

# Vehicle Damage Detection using Semi-Supervised Object Detection

MSc Research Project  
Data Analytics

Maria Raap  
Student ID: X19141700

School of Computing  
National College of Ireland

Supervisor: Majid Latifi

National College of Ireland  
Project Submission Sheet  
School of Computing



<b>Student Name:</b>	Maria Raap
<b>Student ID:</b>	X19141700
<b>Programme:</b>	Data Analytics
<b>Year:</b>	2021
<b>Module:</b>	MSc Research Project
<b>Supervisor:</b>	Majid Latifi
<b>Submission Due Date:</b>	23/09/2021
<b>Project Title:</b>	Vehicle Damage Detection using Semi-Supervised Object Detection
<b>Word Count:</b>	7301
<b>Page Count:</b>	22

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

**ALL** internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

<b>Signature:</b>	
<b>Date:</b>	22nd September 2021

**PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:**

Attach a completed copy of this sheet to each project (including multiple copies).	<input type="checkbox"/>
<b>Attach a Moodle submission receipt of the online project submission</b> , to each project (including multiple copies).	<input type="checkbox"/>
<b>You must ensure that you retain a HARD COPY of the project</b> , both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	<input type="checkbox"/>

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

<b>Office Use Only</b>	
Signature:	
Date:	
Penalty Applied (if applicable):	

# Vehicle Damage Detection using Semi-Supervised Object Detection

Maria Raap  
X19141700

## Abstract

The visual inspection of vehicles for external damages is a major activity in many industries worldwide. Not only does the detection of abnormalities depend on the assessor's expertise, this process is also time-intensive as it must be carried out manually. Recent developments in the area of object detection however provide the opportunity to provide a more automated solution to this problem. While fully supervised learning has already proven successful, this study investigates the potential use of semi-supervised learning-enabled with saliency propagation for vehicle damage detection. In a direct comparison, the semi-supervised learning was not able to achieve the same accuracy rates as fully supervised models, can however accomplish a precision of 56.3% when using labelled data.

## 1 Introduction

### 1.1 Background

The manual inspection of vehicles for external damages is a process known to most and particularly prevalent in the vehicle rental industry, where the accuracy of damage detection is of high importance. The lack of official guidelines and often subjective assessment of damages is making it no surprise that car rental disputes account for the third-largest category of complaints handled by the European Consumer Centre Ireland (European Consumer Centre Ireland; 2020). Not only are such visual inspections time consuming for lessee as well as a rental company, the high level of complaints often due to unexpected supplementary charges for subjectively assessed vehicles damages is leading to negative customer experiences as well as an additional cost for all parties involved for further investigation and formal responses before an acceptable resolution can be achieved. These shortcomings call for an objective solution. With the significant progress made in object detection and instance segmentation, this field of computer vision provides a promising solution. With the possibility to replace manual inspections with visual automated ones, the dependency on the expertise of the assessor is removed. Furthermore, artificial damage recognition reduces the time and effort required for an inspection while also providing a reproducible and objective assessment for all parties involved. While fully supervised models, such as Mask R-CNN have proven to achieve high accuracy rates for the detection of objects (Zhang et al.; 2020; Bouarfa et al.; 2020; Qianqian et al.; 2019; Kim and Cho; 2020), the required effort to annotate a large number of images to achieve these results requires high levels of effort and time. The recent emergence of enhanced

semi-supervised learning models is providing a viable alternative as the even with small percentages of labeled data, good accuracy rates can be achieved (Liu et al.; 2021a; Gong et al.; 2015; Zhou et al.; 2020).

## 1.2 Scope and Limitations

The intention of this study is to identify if a semi-supervised learning model can achieve similar or even better vehicle damage detection accuracy rates as fully supervised models. This work goes beyond the implementations applied in previous studies as it is aiming to use the abundant learning information in the bounding boxes and employ an enlarged training dataset consisting of mainly unlabeled data to reduce the time and cost incurred for the preparation of the data.

Even though the effort for the data preparation is reduced, a large amount of effort is still required to compile the dataset to ensure suitable images are available for training. Furthermore, the annotation of a small portion of the images takes a lot of time. The researcher has only limited software development knowledge and experience and with that, a large portion of available time and effort will go into to setup and configuration of the model. Since the overall project timeline is limited to 6 months, the overall scope is limited to the use of Mask R-CNN as the most suitable base convolutional neural networks (CNN) architectures identified by previous literature (Zhang et al. (2020); Bouarfa et al. (2020); Qianqian et al. (2019); Kim and Cho (2020)) enhanced by a module called "Unbiased Teacher" (Liu et al.; 2021a) that is using saliency propagation to exploit the bounding box information and with that enable semi-supervised learning. This selection of base CNN allows for a direct comparison with previous studies.

## 1.3 Objectives and Research Question

This study attempts to investigate therefore the following research question:

*"Can the accuracy of object masking in semi-supervised instant segmentation models used for damage detection be improved through the use of saliency detection?"*

To accomplish this the study aims to meet the following objectives:

1. Investigate how a state of the art CNN model enhanced with saliency propagation can overcome object detection and segmentation pitfalls of existing models
2. Implement an enhanced state of the art CNN mode for a vehicle damage detection
3. Evaluate the performance of an enhanced state of the art CNN model for semi-supervised learning
4. Compare the performance of a through saliency propagation enhanced state of the art CNN model with standard CNN in the context of vehicle damage detection

The major contribution of this research is a detailed analysis and evaluation of the performance of a state of the art CNN model enhanced with saliency propagation module in the context of vehicle damage detection in a semi-supervised learning setting.

## 1.4 Document Overview

Section 2 of the paper is providing an overview of related literature with a focus on instance segmentation and saliency propagation and its application to damage detection. Section 3 discusses the applied research methodology, providing details on the set hypothesis and further specification on the 3-tier model architecture applied. In section 4 the different elements of the CNN model are discussed in more detail providing the basis for the model implementation and training details described in section 5. This section is also providing the results of the different models run. The evaluation of the results of the fully and semi-supervised models is discussed in section 6. The paper concludes the research in section 7 while also discussing future work.

## 2 Related Work

For this study, a comprehensive review of the relevant literature of fully as well as semi-supervised object detection and instance segmentation has been conducted with the main focus on damage detection. Section 2.1 provides an overview of findings of recent work on object detection and instance segmentation through CNN. Followed by the advancements and current limitations in the area of semi-supervised learning in object detection models in section ???. A special focus is placed on the evolution of saliency detection methods and how these are used to enhance existing object detection models. Section 2.3 provides details on previous studies that focused on damage detection on different surfaces through the use of state-of-the-art CNN. This chapter concludes with section 2.4 with an overall summary of findings of the review of related work.

### 2.1 Object Detection and Instance Segmentation

CNN models used for object detection and instance segmentation have seen a significant increase in applications in industries reaching from medicine to detect anomalies on CT scans to damage detection on vehicles (Hoeser and Kuenzer; 2020). While the overall structure of the CNN architecture has not changed, the accomplished prediction and accuracy levels have improved drastically. Particularly the optimization of feature extraction modules contributed to this development and leveraged the application of CNN models from a general classification to a multi-object detection and classification question with the convolutional backbone at its centre (Chen et al.; 2018; Minaee et al.; 2020). This development enabled also a more accurate and detailed detection of even small objects, a crucial part of vehicle damage detection. Kumar et al. (2020) identifies in that particularly for damage detection the architecture must be able to identify multi-scale features and able exploit image context to produce a satisfactory degree of classification and detection.

In He et al. (2017) addresses this issue and propose a new state-of-the-art CNN model with enhanced object detection capabilities, called Mask R-CNN. This new model is built on the previously proposed Faster R-CNN (Ren et al.; 2016) but extends the architecture with a mask head as well as a Feature Pyramid Network (FPN) in the convolutional backbone that enables additional segmentation to the formerly used classification and bounding box regression heads. While also other attempt to optimize CNN architectures, Mask R-CNN is considered the most suitable due to its ability to detect asymmetrical, small and even multiple object accurately (Hoeser and Kuenzer; 2020). For example, Cai

and Vasconcelos (2018) are aiming to improve previous models by increasing bounding box refinement through cascading IoU, this approach struggles with complex detections of objects and is therefore not suitable for damage detection due to its asymmetrical character.

Even though Mask R-CNN is able to enrich feature extraction for enhanced object recognition, the backbone structure relevant for the object detection itself is not as optimised and requires training from scratch for every instance (Liu et al.; 2020; Hoeser and Kuenzer; 2020). In Kumar et al. (2020) selects Mask R-CNN as the base model and compares its detection performance in combination with the different backbones in a case study around varying degrees of vehicle damages. Different results are displayed, where ResNet outperformed others significantly. This result concurs with similar research on damage detection (Kim et al.; 2021; Zhang et al.; 2020; Bouarfa et al.; 2020).

While state-of-the-art CNN models such as Mask R-CNN display exceptionally accurate object detection, researchers commonly identify the need for fully annotated datasets as obstacles (Kim and Cho (2020); Kim et al. (2021); Zhang et al. (2020); Bouarfa et al. (2020); Hoeser and Kuenzer (2020)). Since extensive manual effort is required to prepare a large enough dataset to facilitate sufficient learning and with that accomplish satisfactory precision and accuracy, more and more research investigates the use of semi-supervised learning to overcome this limitation (Liu et al.; 2020; Kim et al.; 2021; Liu et al.; 2021a; Kuo et al.; 2019; Qianqian et al.; 2019).

## 2.2 Semi-Supervised Learning

A hybrid of unsupervised and supervised learning can be known as semi-supervised learning. Here only a part of the training data is labelled and annotated. This usually smaller part facilitates supervised learning. Such semi-supervised methods usually utilize low-density separation used by transductive support vector machines (TSVM) or graph-based methods such as saliency propagation. TSVM is an extension to supervised SVM, here the principles of transductive interference are added to facilitate the learning on unlabelled data (Kim et al.; 2021; Kuo et al.; 2019; Zhou et al.; 2020). The principle of transduction is also used in graph-based methods, however, the provided data is evaluated and then mapped to the Euclidean space to form connections between neighbouring entities, also called relationships. These relationships provide details on the similarity between superpixels or nodes (Kim et al.; 2021; Zhou et al.; 2020) through the distance between them. The complexity in models used for object detection however makes it difficult for semi-supervised techniques to transfer the learning from the classification to object detection stage (Zhang et al.; 2020; Liu et al.; 2021a).

### 2.2.1 Saliency Propagation

Saliency detection is making use of multiple instance learning (MIL) and utilized this initial step to activate the saliency. To do this, box annotations are firstly assigned to identifying and classifying objects in an image including the number of detected objects together with the relevant location. The identified box can also be described as a set of pixels where each pixel corresponds to a binary label  $y$  indicating if a pixel belongs to the mask of the detected object or not, measured by the IoU (Zhou et al.; 2020; Kim et al.; 2021). If the IoU high enough the box is considered positive and brought forward to the next stop. Should this not be the case the box is disregarded and not used for

further training. Through this approach the learning  $F$  is kept lightweight as the focus is on box-level labels:

$$L = \sum_c \left( \sum_{p_i \in P^c} -\sigma(F(\omega(p_i), \theta)) + \sum_{n_j \in N^c} +\sigma(F(\omega(n_j), \theta)) \right) \quad (1)$$

”where  $L$  is the loss,  $\theta$  is the learnable parameters in  $F$ ,  $\omega$  is a 2D aggregation operator, and  $\omega$  is a region feature pooling operator” (Zhou et al.; 2020). Based on the collected information high-salient regions in the image are predicted building the shape prior information. This information is essential for the mask head used later in the CNN. To further enhance the learning through propagation and exploit identified relationships between pixels, a latent space message passing process is implemented (Zhou et al.; 2020). Convolutional blocks are used to encode an instance saliency map  $M \in R^{H \times W}$  to latent features  $M \in R^{C \times H \times W}$  as shown in figure 1. Subsequent processes are made more robust through the iterative encoding of extracted features and propagation weight prediction between the group of pixels. Through normalization and shuffling, these predicted propagation weights are used to build location-specific kernels. Once the iterative process is complete a convolutional layer is used to decode the updated latent features and build a map combining the information from box and mask. This is providing the model with an intermediate object serving an even stronger shape prior information to the subsequent mask head. The learned information are then integrated into the instance segmentation modules mask head to streamline the subsequent learning of mask predictions.

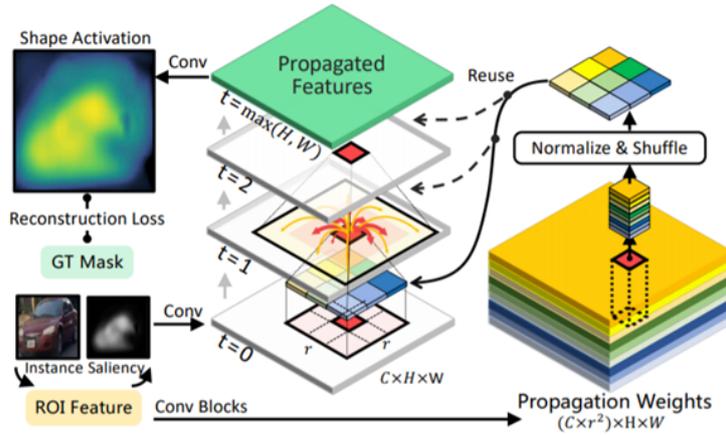


Figure 1: Saliency Propagation model (Zhou et al.; 2020)

## 2.3 Automated Damage Detection

The evolution of object detection models generated increased interest in many industries, such as insurance, maintenance or construction (Kumar et al.; 2020). The interest in the insurance and car rental industry for automated solutions is however enormous (Kumar et al.; 2020; Zhang et al.; 2020). Limited literature is available that is investigating the application of state-of-the-art CNN for damage detection. All of the previous literature focused on damage detection uses fully supervised learning, no study has been identified where semi-supervised learning has been applied. Table 1 displays an overview of the identified relevant literature.

Table 1: Overview of recent research on automated damage detection

Source	Specification	Advantage	Limitation/ Shortcoming
Kumar et al. (2020)	<ul style="list-style-type: none"> <li>- Focus on Car Damage detection with 451 images</li> <li>- Use of Mask R-CNN with different classifiers</li> <li>- SVM</li> <li>- Softmax</li> <li>- Autoencoder</li> </ul>	<ul style="list-style-type: none"> <li>- detailed image characterization of dataset</li> <li>- high precision of 88.5% (Softmax)</li> </ul>	<ul style="list-style-type: none"> <li>- low precision of 54.3% (Autoencoder)</li> <li>- no information on backbone used</li> <li>- small dataset</li> <li>- limited precision on small and multiple damages</li> <li>- no details on evaluation method</li> </ul>
Zhang et al. (2020)	<ul style="list-style-type: none"> <li>- Focus on Car Damage detection with 2000 images</li> <li>- Use of Mask RCNN with optimized ResNet50</li> </ul>	<ul style="list-style-type: none"> <li>- high precision of 99.5 on normal damages</li> <li>- high precision of 93.3 on minor damages</li> <li>- high precision of 98.6 on multi damages</li> <li>- different exposure and light conditions of images</li> <li>- detailed analysis and evaluation of results</li> </ul>	<ul style="list-style-type: none"> <li>- only use ResNet50</li> <li>- partially inaccurate mask instance segmentation</li> </ul>
Kim and Cho (2020)	<ul style="list-style-type: none"> <li>- Focus on Concrete Damage detection 765 images</li> <li>- Use of Mask RCNN with optimized loss function</li> </ul>	<ul style="list-style-type: none"> <li>- high precision of 90.41%</li> <li>- detailed image characterization of dataset</li> </ul>	<ul style="list-style-type: none"> <li>- unable to detect small concrete cracks</li> </ul>
Bouarfa et al. (2020)	<ul style="list-style-type: none"> <li>- Focus on Aircraft Damage detection with 50 images</li> <li>- use of Mask R-CNN with updated weights</li> </ul>		<ul style="list-style-type: none"> <li>-low precision of 57.3%</li> <li>- small dataset</li> <li>- unable to detect minor damages</li> </ul>

Most of the datasets used in the review are very small, Bouarfa et al. (2020); Zhang et al. (2020) state that even the gathering of images for the study has been difficult as very little high-quality images are publicly available. Kim and Cho (2020); Kumar et al. (2020) mention that the labelling of larger datasets requires too much time. Here semi-supervised learning provides a suitable solution. The size of the dataset used on the study seems to be a big contributor to accomplished precision rates. (Bouarfa et al.; 2020) uses the smallest dataset with only 50 images and only achieves a precision rate of 57,3%, while (Zhang et al.; 2020) compiled a dataset with 2000 labelled images and accomplished a 99.5% accuracy.

In an in-depth study Kim and Cho (2020) analyse the use of Mask R-CNN for the detection of cracks on concrete structures and accomplish an impressive 90% precision rate. The researchers not only evaluate the performance of the standard model but optimize the loss function so such high accuracy rates for the detection can be achieved.

When however faced with small and not clearly distinguishable damages the model is not able to produce such high precision rates. This is a commonality in most studies reviewed (Kumar et al.; 2020; Zhang et al.; 2020; Bouarfa et al.; 2020).

The most successful approach was taken by Zhang et al. (2020) where the base Mask R-CNN model was used with an optimized ResNet50 backbone. No comparison with other backbones like ResNet101 was carried out. The high precision rates even for minor and multi damages might be due to the large training dataset of 2000 labelled images rather than the optimized backbone. In the evaluation of the study, the researchers advise that the mask segmentation is partially inaccurate and more research is required to determine the cause.

## 2.4 Literature Summary

While a few studies attempted to utilize Mask R-CNN for damage detection in different settings like concrete (Kim and Cho; 2020), aircraft (Bouarfa et al.; 2020) or vehicles (Kumar et al.; 2020; Zhang et al.; 2020; Qianqian et al.; 2019; Harshani and Vidanage; 2017), only a few were able to achieve satisfactory accuracy rates. All identified research focused on fully supervised learning and no published papers were available that utilize semi or weakly supervised learning. The reviewed papers all highlight the importance of detailed labelling of damage details such as length and depth to enable any model to identify these. The results of these papers also demonstrate that the larger and more detailed the provided training dataset is, the more accurate the predicted boxes and masks of the damages are. At the same time, the authors acknowledge that this is not always feasible due to a lack of available high-quality images and limited resources to prepare the dataset appropriately. Altogether it must be acknowledged that it is more difficult for a model to detect damages due to its asymmetric characteristics.

The reviewed literature is showing a clear trend that the architecture of state-of-the-art CNN models utilizing fully supervised learning has matured over the past decade. When used on benchmark datasets, these models achieve very high accuracy. When these models are however used on smaller bespoke datasets for a particular purpose like damage detection the accuracy rates drop significantly (Kumar et al.; 2020; Bouarfa et al.; 2020). The literature agrees that this is mostly due to the limited learning potential when only a few training images are provided to the model. A requirement, therefore, emerges to enlarge the dataset without necessarily increasing the effort to prepare it through labelling and annotations. This can either be done through artificial augmentation methods like K-fold cross-validation or the use of semi-supervised learning. Since damages are unlike other objects and can take any shape or form as well as position, the use of augmentation methods does not necessarily provide the CNN model with more learning material and therefore semi-supervised learning appears to be the best alternative to provide models with additional learning material.

## 3 Methodology

In this study, the damage detection accuracy in a semi-supervised setting using a Mask R-CNN model extended with an 'unbiased teacher' module is analysed. It builds on previously conducted research that only use the base Mask R-CNN model in a fully supervised setting. This chapter starts with the definition of the research hypothesis in section 3.1. Section 3.2 provides an accurate description of the research procedure

followed including steps followed and how data was gathered and prepared. In section 3.3 materials and equipment used is described. This chapter concludes with section 3.4 that explains how the measurements and calculations were performed and which statistical techniques were used.

### 3.1 Hypothesis

The research question outlined in section 1.3 suggest that the accuracy of object masking can be improved through the use of larger datasets utilizing semi-supervised learning, particularly for vehicle damage detection. The supposition that the accuracy of base Mask R-CNN can be improved through the use of saliency propagation has been already validated Gong et al. (2015); Kuo et al. (2019); Zhou et al. (2020). Furthermore, Liu et al. (2021a) demonstrates that similar high model accuracy rates can be achieved using only part of the labelled data while providing larger portions of unlabelled data for semi-supervised learning. The hypothesis for this study is, that through the use of saliency propagation enabled pseudo-labelling on unlabelled data high accuracy rate can be achieved in the application of vehicle damage detection. A clear superiority in precision and recall in comparison to previous studies must be displayed to accept this thesis. The set hypothesis can be accepted or rejected depending on the evaluation and ultimately allows this study to answer the defined research question. If H0 is accepted, the research question can be answered with ‘yes’, vehicle damage detection can be enhanced in a semi-supervised setting through the use of an enhanced pseudo-labelling module in the base Mask R-CNN model. Should H0 be rejected, the research question is answered with ‘no’, semi-supervised learning with enhanced pseudo-labelling is the example of vehicle damage detection does not necessarily produce more accurate results than a base Mask R-CNN model.

**H0:** CNN models enriched with enhanced pseudo-labelling methods achieve a statistically higher (with 95% confidence) precision and recall in a semi-supervised setting than simple CNN models when utilised to detect the damage on vehicles.

### 3.2 Research Procedure

For this study, an extensive review of published and peer-reviewed literature has been conducted before a selection of the base CNN and enhancement module was made. The literature has been accessed through the National College of Ireland’s (NCI) library portal to ensure only peer-reviewed and qualitative literature is considered.

The dataset used for the study is a compilation of publicly available datasets (Shah; 2019; Amir; 2021). All images were reviewed for duplicates and to ensure only high-quality images with visible damages are included in the set. The final dataset was then split and ten per cent of randomly selected images were annotated with COCO annotator (Brooks; 2019), resulting in a required JSON file for the following model training. Once the decision has been made to use Mask R-CNN as base CNN model (Abdulla; 2017) and enhance this further with a semi-supervised module ‘Unbiased Teacher’ (Liu et al.; 2021b) the relevant codebases which are publicly available and are authorized<sup>12</sup> to be used for

---

<sup>1</sup><https://github.com/facebookresearch/unbiased-teacher/blob/main/LICENSE>

<sup>2</sup>[https://github.com/matterport/Mask\\_RCNN/blob/master/LICENSE](https://github.com/matterport/Mask_RCNN/blob/master/LICENSE)

Model	Backbone
Fully Supervised Mask R-CNN	ResNet50 ResNet101
Semi-Supervised with 1% labelled data	ResNet50 ResNet101
Semi-Supervised with 5% labelled data	ResNet50 ResNet101
Semi-Supervised with 10% labelled data	ResNet50 ResNet101

Table 2: Model Overview

further research were retrieved from Github and the code amended and extended were required to cater for the bespoke requirements of this study.

For the Mask R-CNN model, the epochs were reduced since the dataset used in this study is significantly smaller than the benchmark dataset COCO used to evaluate the performance of the base model. Additionally configuration files were added to be able to run the model on backbone ResNet50 and ResNet101 separately as the base model only includes ResNet50. For the enhanced model, the base structure of Liu et al. (2021b) was taken and the Faster R-CNN structure was replaced with the more comprehensive Mask R-CNN one. Also here additional configuration files were added to facilitate the learning with 1,5 and 10 per cent of labelled data as well as the different backbones. For both models, tensorboard was added to enable a comparable evaluation. The previously prepared dataset was then centrally stored and the link was added to the configuration file for each model. For the base Mask R-CNN and the semi-supervised model, separate anaconda3 environments were configured as both had different requirements on package versions.

After the environment setup, the different models were triggered through a command in the terminal calling the different configuration files. See also table 2 for all the models to run for this study. The evaluation has been carried out as the final step in tensorboard

### 3.3 Materials and Equipment

Materials and Equipment used in this study:

- Azure Virtual Machine with NVIDIA compatible GPU
- Ubuntu 18.04 operating system
- Anaconda3
- Python with packages like Keras, Tensorflow, Cuda
- Mask R-CNN base model (adapted from Abdulla (2017))
- Semi-Supervised Module 'Unbiased Teacher' (adapted from Liu et al. (2021b))
- Coco Annotator (Brooks; 2019)
- prepared and annotated dataset based on Shah (2019); Amir (2021)

### 3.4 Model and Result Evaluation

To measure the performance of the set up model correctly, the performance metric consisting of precision, identification of actual damage classification, and recall, identification of damages localisation and segmentation, are selected. The calculation of precision and recall are as follows:

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

$$Recall = \frac{TP}{TP + FN} \quad (3)$$

Where:

- **TP** stands for true positive and is equal to the number of correctly detected damages
- **FP** stands for false positive and is equal to the number of incorrectly identified damages
- **FN** stands for false negative and is equal to the number of damages not detected by the model

Furthermore, to enable a direct comparison with previous studies, mean average precision (mAP) is used.

$$mAP = \frac{\sum_{q=1}^Q AveP(q)}{Q} \quad (4)$$

Where:

- **q** is a given query
- **Q** is the number of queries
- **AveP(q)** is the average precision for a query

## 4 Design Specification

The following sections provide more details of the different modules used in the adopted architecture, a combination of Mask R-CNN modules followed by the saliency propagation modules responsible for both object detection as well as segmentation. The overall framework architecture is illustrated in Fig. 2.

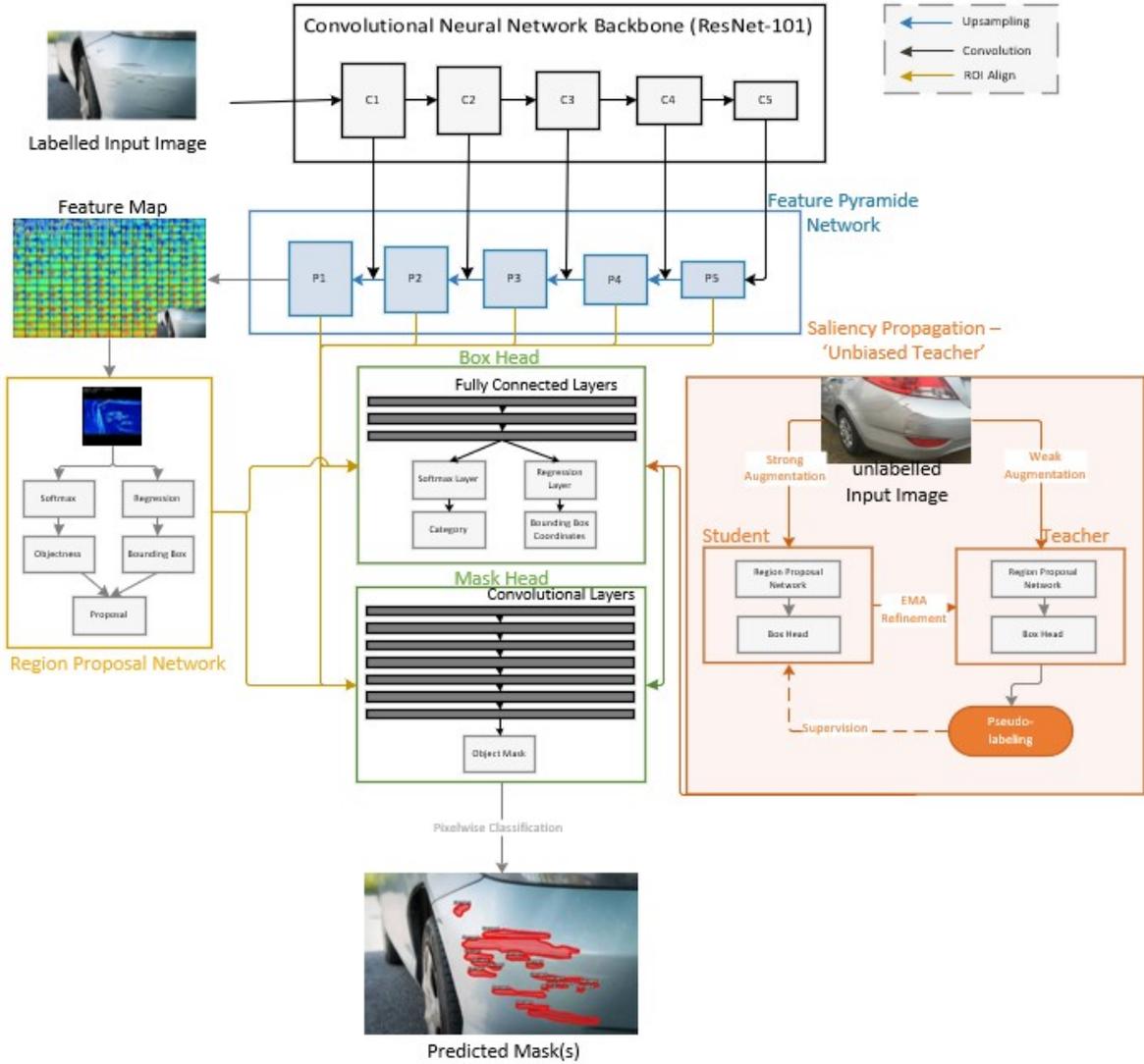


Figure 2: Overall Framework. Adapted from (He et al.; 2017) and (Liu et al.; 2021a)

## 4.1 Mask R-CNN modules

### 4.1.1 Feature Extraction

The input image is initially passed through several convolutional layers (C1 - C5) of different sizes with increasing depth and decreasing resolution of the backbone ResNet-50/-101 to extract major features. Weights and bias are imported from the pre-trained model, however, optimized while training by using loss functions (He et al.; 2017). The generated feature maps are fed then into a feature pyramid network (FPN) to restore the spatial information potentially lost in the previous step, building a feature map.

### 4.1.2 Region Proposal Network (RPN)

This map contains a large number of frames with the region of interest (ROI) which is subsequently passed through the RPN where a classification and regression classifier detect the existence and location of potential objects and marks these as proposed bounding boxes using a small fully connected network (FCN). Here a  $n \times n$  sliding window scans the

relevant map, compiles the result and feeds these into the subsequent classification and regression layer with 2k and 4k elements, respectively, where k is the number of anchors of a window. In this way, different sized candidate boxes are created that represent the object class as well as the horizontal and vertical coordinates of the proposed bounding boxes together with the width and height of the anchor (Ren et al.; 2016; He et al.; 2017).

### 4.1.3 Box Head

In this step, the previously proposed bounding boxes are refined further through ROIAlign (He et al.; 2017). Features in the bounding boxes are extracted with higher accuracy using the bilinear interpolation. In the subsequent regression layer, these features are corrected and more accurate bounding boxes are proposed. At the same time, the classification layer classifies the object using the softmax function (Girshick; 2015) while employing a smooth L1 loss for optimization.

### 4.1.4 Mask Head

With the provided information the mask head is estimating the shape of the detected object regardless of the class. This is achieved through the distinction of fore- and background object objects by pixel-by-pixel binary classification and the use of a binary cross-entropy/ log loss function (He et al.; 2017). The originally proposed setup by He et al. is used, including four 3x3 convolutional layers that maintain the resolution, a 2x2 deconvolutional layer and an output layer of 28x28 for the predicted mask.

## 4.2 Saliency Propagation modules

### 4.2.1 Learning with Pseudo-Labeling

As illustrated in Fig 2 unlabeled images were assessed based on the previous learning of the fully supervised module and provided pseudo-labels from the 'Teacher' module. The model is attempting to generate object labels with the pseudo-labelling method based on the previous learning of the fully supervised module and provided pseudo-labels from the 'Teacher' module. If more than one object has been identified, class-wide non-maximum suppression (NMS) is applied to remove repetitive predictions and remove noisy pseudo-labels and reduce bias. For all images, the predicted bounding boxes are assessed based on the confidence threshold  $\sigma$ , if it has been reached the images is considered strongly augmented and directly fed to the 'Student'. Should  $\sigma$  not be reached, the image is fed to the 'Teacher' for evaluation and generation of pseudo-labels. These generated pseudo-labels are made available in form of learnable weights to the 'Student' via back-propagation (Liu et al.; 2021a).

### 4.2.2 Refinement via Exponential Moving Average

To enhance the generated pseudo-labels by the 'Teacher', exponential moving average (EMA) is applied. Here the information learned by the 'Student' is fed back to the 'Teacher' via propagated network weights. A step is inserted to further exploit the relationship between pixels and enhance the incomplete object region responses. This is a more effective approach to explore the latent feature space between pixels instead of only relying on pixel-wise classification. Considering deep pixels as nodes and the model

is propagating the space as a message between the spatially adjacent nodes and predict propagation weights (Gong et al.; 2015; Liu et al.; 2021a).

### 4.3 Integration of propagated saliency

The learned shape prior and saliency propagation features are concatenated with the input region features and fed into the mask head. As proven by Liu et al. the additional information provides 'strong prior information of objects' possible shapes' and with that significantly improve the predicted masks. In this way, the mask head can focus on fine-detailed information.

## 5 Implementation

This section discusses the implementation of the different models used the detection of damages on real-life vehicle images. Section 5.1 discusses the initial data understanding and preparation. Section 5.2 lists the environment requirements before the set up models can be executed while section 5.3 describes how the prepared data is loaded into these models for the algorithm training. Afterwards section 5.4 which describes how the different models are trained in detail. This chapter concludes with section 5.5 detailing the evaluation stage of the project.

### 5.1 Data Preparation

**Image sanitation:** While the selected datasets were specifically designed for vehicle damage detection they included several images that are unsuitable for this project and were therefore removed to avoid negative influence of the overall performance of the learning algorithm.

Images removed based on:

- missing parts as this is not classified as damage in this project
- total wreckage of cars
- very low quality
- images from images sharing sites like shutterstock, stockphoto...
- no damage visible/not damaged cars
- car partially blocked by other objects

**Image Annotation:** As this project is training the algorithm to detect damages using Mask R-CNN with semi-supervised learning, a part of the images require mask labels, so-called annotations, for the algorithm to retrieve sufficient learning to apply generalisation on previously unseen data. Annotations are precisely identified regions of a damage in an image. This is based on previous attempt of semi-supervised learning that only used 1% of annotated images to achieve outstanding results (Zhou et al.; 2020). To perform this task, COCO Annotator is used for 10% of selected images and polygon masks are drawn along the damages for each image. The annotations, including the coordinates of the bounding boxes and masks are exported as JSON file which is fed later into the algorithm for training.



Figure 3: Image no Annotation



Figure 4: Image with Annotation

## 5.2 Environment Set-up

Before training the designed model, the setup of the correctly configured environment is crucial. For an enhanced performance, an Azure Standard NC6 with 6 vCpus, 56 GiB memory powered with NVIDIA Tesla K80 GPU is set up with Ubuntu 18.04 as operating system. Elements used in the environment:

- Anaconda3
- Python 3.6
- Python Packages (Tensorflow, Keras, PyTorch, OpenCV, Numpy, Scipy, Scikit Learn, Scikit Image, Pandas, Matplotlib, Pillow, Caffe, Java JDK, PyCocoTools (MS COCO dev kit), CUDA 10.2)
- Pre-trained Mask R-CNN including pre-configured weights (Abdulla; 2017)
- Pre-defined saliency propagation module (Liu et al.; 2021a)

## 5.3 Loading Dataset

The subsequent training of the algorithm is split into two approaches. First, the fully supervised Mask R-CNN network is trained to enable a comparison with previous papers. Secondly, the semi-supervised Mask R-CNN enriched with saliency propagation to test the set hypothesis. For the fully-supervised instance only the approximately 268 annotated images together with the generated JSON file are fed into the algorithm. In addition also the remaining 2791 unannotated images are provided to the algorithm for enhanced learning for the semi-supervised learning instance.

## 5.4 Algorithm Training

### 5.4.1 Fully Supervised Mask R-CNN

The implementation of the fully supervised training used the FPN + ResNet101 and ResNet50 backbone layer for the detection of low as well as high level features. Subsequently the generated feature map is passed to the RPN, scanning over more than 300k anchors to check for car damages, resulting in class and bounding box regression and ultimately ROI which are passed to the box and mask head. For this implementation two classes are generated, damage and background. the background class is discarded as not relevant

Table 3: Hyper-parameter setting for Mask R-CNN

Configuration	Description
ResNet Architecture	ResNet50 ResNet101
Learning Rate (LR)	0.001
LR momentum	0.9
Number of Epoch	20
Steps per epoch	250
Validation steps	5
Num classes	2

Table 4: Hyper-parameter setting for enhanced Mask R-CNN

Configuration	Description
ResNet Architecture	ResNet50 ResNet101
Learning Rate (LR)	0.001
LR momentum	0.9
Number of Epoch	20
Steps per Epoch	250
Validation steps	5
Num classes	2

for further evaluation. In the box head, the proposed box are further refined through the crop and resize function to handle the varying ROI box size and convert them into fixed size. The mask head is taking the anchors with detected damage, positive anchors, and generates high quality masks over the identified object. Feature weights are initialized by the COCO-pretrained mode also used by He et al., no further data augmentation is carried out.

The following network hyper-parameter were set up for the model as shown in table 3. Due to time, memory and processing power constraints the training was performed on 283 training and 52 validation images, including in total 762 damage annotations. As an indicator for how well the algorithm models, a multi-task loss for each ROI is defined as following:

$$L = L_{class} + L_{box} + L_{mask} \quad (5)$$

where  $L_{class}$  and  $L_{box}$  describe the class and bounding box loss which is adopted from the Fast RCNN model (Girshick; 2015).  $L_{mask}$  is a pixel-wise cross-entropy loss for RPN mask localization (He et al.; 2017). AP50:95 (denoted as mAP) is used as an evaluation metric.

#### 5.4.2 Semi-Supervised Mask R-CNN enriched with saliency propagation

The same implementation of the fully supervised training used for Mask R-CNN as in section 5.4.1 was used. To facilitate the semi-supervised learning, an additional module for the saliency propagation is added. Here the annotated training images are fed into the teacher module that is learning from these and leveraging these learnings on the un-annotated images through pseudo-label. Random horizontal flip for weak augmentation and randomly add color jittering, grayscale, Gaussian blur, and cutout patches for strong augmentations are used for data augmentation based on the original work by Wu et al..

The following network hyper-parameter were set up for the model as shown in table 4. The training was performed on 2791 un-annotated and 283 annotated images, while validation used 163 images. The same evaluation metric was employed and based solely on the teacher module.

## 5.5 Model Results

Table 5: Experiment Outcomes

Configuration	Description	mAP
<i>Fully Supervised</i>	ResNet50	84.37
	ResNet101	64.34
<i>Semi Supervised</i>		
- 1% labelled data	ResNet50	56.06
	ResNet101	48.74
- 5% labelled data	ResNet50	65.15
	ResNet101	62.68
- 10% labelled data	ResNet50	81.28
	ResNet101	74.84

## 6 Evaluation

The evaluation chapter discusses the findings displayed in chapter 5 and is detailed in table 5. Section 6.1 discusses the results for the fully supervised model in comparison with previous studies. Section 6.2 evaluates the outcomes from the semi-supervised model result in comparison with the fully supervised model.

### 6.1 Fully Supervised Learning Evaluation

The fully supervised model is performing well on the relatively small dataset provided. With ResNet50 a mAP of 84.37% was accomplished, while ResNet101 only achieved a mAP of 64.34 %, see also figure 5. The results are similar to the ones achieved in previous studies displayed in table 1. The model performed significantly better than Bouarfa et al. (2020) model. This is however not surprising as the dataset used is also five times the size of the one used in this study. Compared to Kumar et al. (2020); Zhang et al. (2020); Kim and Cho (2020) the mAP of this experiment is not as high, however, this might be explained by the smaller dataset used. Furthermore, no optimisations either in the loss function or backbone layer were applied.

The model is performing very well on images where the damage is clearly visible. Similar to previous studies our model displays inaccuracies in the mask segmentation for multi damages. The model is able to detect most damages in the presented images, however, the model is grouping some of the smaller damages in one mask, see also comparison in figure 6. This issue might be due to the small training dataset, as the majority of training images have only single damages. To overcome this, additional images with multi-damages might need to be fed to the model fit it to be able to detect multiple instances more accurately. Another explanation might be the use of pre-trained weights from the COCO dataset. As the object for this model is relatively large but uniform, the model might require further adjustment on the weights to cater for the asymmetric character of vehicle damages. A further experiment where the model is trained from scratch solely on vehicle damage data with a larger training dataset is recommended to overcome this issue. Also when the image quality is not as high, the model struggles to identify any damage and in some cases applies an incorrect mask. A possible explanation

for this is the lacking pixel-level differentiation which is used by the mask head to accurately determine the outlines of a particular identified damage. Overall of 14 test images, the model was able to only identify 5 pictures with damages. Of these 5 the identified damages were only in one instance accurate, where there was only one damage visible.

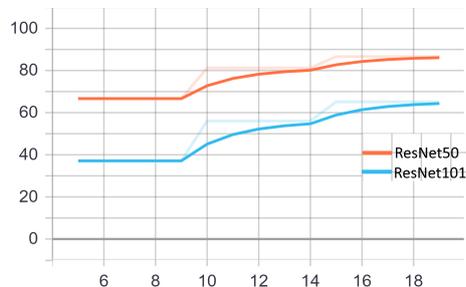


Figure 5: mean Average Precision

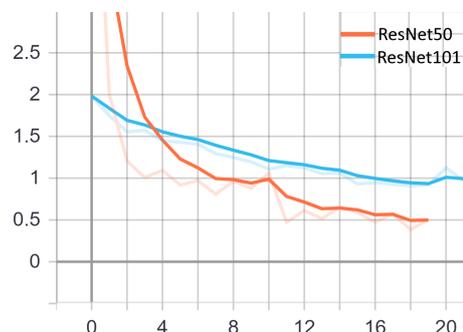


Figure 6: Overall Function Loss

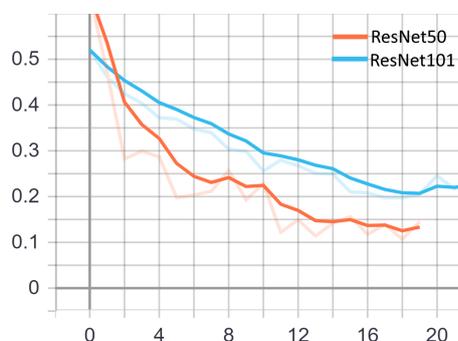


Figure 7: Bounding box loss

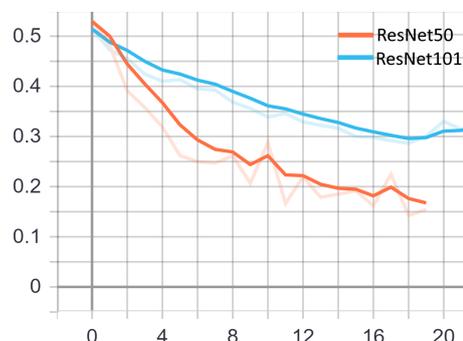


Figure 8: Mask loss

Two models were run, one with the ResNet50 and a second with the ResNet101 backbone. Each model took approximately 35 hours to run with each epoch taking around 1.45 hours. The model with ResNet101 took a little longer to run which can be explained by the additional layers. This is a relatively long run time for such a small training set, however, no indication is given in previous studies how long it took to train their respective models and for that no comparison is possible. As can be seen in figures 6, ResNet50's loss function is declining faster than ResNet101 one. It should however be considered to run training with more epochs to check if ResNet101 will outperform ResNet50 over longer training periods.

## 6.2 Semi-Supervised Learning Evaluation

The semi-supervised model is not performing as well as the fully supervised model even though a larger dataset with an additional 2791 unlabelled images was provided during the training stage. See detailed results in table 5. Overall the efficacy of the Mask R-CNN model enhanced with the 'Unbiased Teacher' module performs best with the highest percentage of labelled data. With 10% labelled training images the highest mAP of 81.28% was achieved with ResNet50 coming close to the results from the fully supervised training. Similar to the basic Mask R-CNN all the models trained also still struggle with multi damage instances and cannot identify damages where the image quality is sub-optimal, see also examples in table 7. This is to some extent expected as also the base

Table 6: Image Annotations Mask R-CNN with ResNet50

Description	Original Picture	Model Annotated Picture
Low-quality Image		
One Damage		
Multi Damage		

model is struggling with these and generated pseudo-label would not be able to bridge the lack of learning.

Two models were run, one with the ResNet50 and a second with the ResNet101 backbone. Each model took approximately 50 hours to run with each epoch taking around 2.5 hours. It can be observed, that models trained on 10% labelled data achieves better results than the ones trained on lower percentages of trained data. This can be explained through the increased learning potential of the base model which can provide the 'Trainer' element of the model that is generating the pseudo-labels for the 'Student'. Without this substantial learning the accurate recognition of damages is not possible in the unlabelled data and can therefore also not be learned from the student. However, potentially this lacking learning experience with lower percentage models might overall impact the 'Teacher-Student' learning relationship. As usually, the 'Student' would provide feedback to the 'Teacher' this feedback becomes increasingly inaccurate the more images have incorrect or unreliable pseudo-labels. Further research is required to investigate this claim.

As can be seen in figures 11, 10 and 12, the loss which measures the fit of the bounding box and mask of the predicted annotation reduces at a similar rate for all models. Through the advantage of increased learning of the model with 10% labelled data, the loss is marginally greater. This is ultimately contributing to the overall better performance.

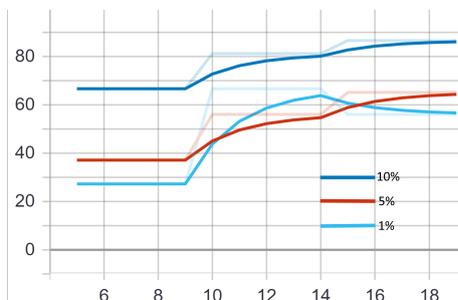


Figure 9: mean Average Precision

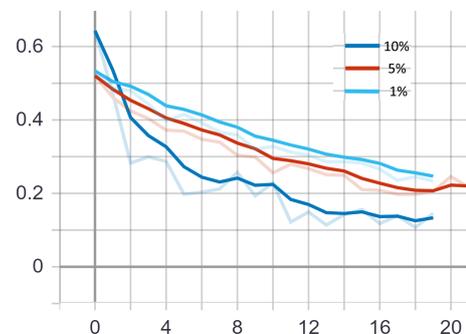


Figure 10: Overall Function Loss

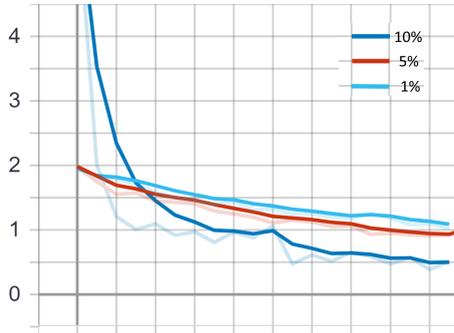


Figure 11: Bounding box loss

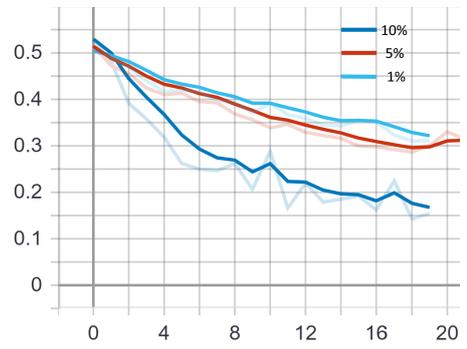


Figure 12: Mask loss

Table 7: Image Annotations of enhanced Mask R-CNN with ResNet50

Description	Original Picture	Model Annotated Picture
Low quality Image		
One Damage		
One Damage		
One Damage		
Multi Damage		

## 7 Conclusion and Future Work

Based on the carefully carried out evaluation, the set hypothesis in section 3.1 cannot be accepted as the base Mask R-CNN model outperforms the model utilizing the semi-supervised learning module. With that also the research question must be rejected, as semi-supervised learning with enhanced pseudo-labelling in the example of vehicle damage detection does not produce more accurate results than a base Mask R-CNN model.

The initial dataset was compiled from several kaggle datasets. An extensive cleaning extensive was carried out to ensure that only images of acceptable quality and with visible vehicle damages are included in the training dataset. The cleaning of the dataset can be seen as the first crucial step to the high precision achieved by the models. Furthermore, the detailed annotation of the selected 268 images is another contributor to the high accuracy rate. Extra time and effort were spent to ensure a precise annotation of the individual damages identified in an image to enable the model to learn different shapes and locations of damages even in the same picture. To enable such a detailed annotation of each individual damage, polygons were selected as tool of choice as these allowed to capture all characteristics of each damage. The detail of each damage is then presented to the model in form of a JSON file with the detailed coordinates of each damage, enabling the model to a more enhanced learning. The selected dataset had a size of 268 annotated images for fully supervised learning and 2791 unlabelled images for unsupervised learning. It is considered a reasonable size for the training on vehicle damages based on a comparison with related research, see table 1. Due to this, it is assumed that there is no over-or underfitting, this can be however only be ruled out through further research.

There is only limited detail provided in the related literature on how much effort was put into the selection of the dataset and the annotation itself. Only Zhang et al. (2020); Kim and Cho (2020) provide an overview of the image selection and annotation process as well as the characterisation of different images. As these two studies also have the highest precision rates in automated damage detection, it is assumed that careful images selection and annotation plays a vital part in the performance of a model.

Another contributing factor to the good performance of the models is also the decision to provide the models with pre-trained weights from the COCO dataset training of the base Mask R-CNN (He et al.; 2017). This transfer learning enables the models to leverage of the previous training on 80 different categories on 330k images (Microsoft; 2014). It provides the model with an advantage in the feature representation and detection and reduces training time since weights only need to be fine-tuned to cater for specific characteristics of vehicle damages (He et al.; 2017; Zhang et al.; 2020; Bouarfa et al.; 2020). There is also the possibility that the validation dataset has been too small and contained too many similar images resulting in a too optimistic and high variance estimation of model performance. Due to limited time available, this has not been investigated further in the study, should be however considered for further research.

While there are numerous applications that achieve promising high precision rates on large benchmark datasets with semi-supervised learning such as Kuo et al. (2019); Zhou et al. (2020); Liu et al. (2021a), more research is required for the application on vehicle damage datasets. These studies show that with a relatively small set of training data only mediocre results are accomplished. Potentially the results can be improved by increasing the overall dataset size, however, would contradict the overall purpose to reduce the effort in the preparation of the data. Future studies should investigate different model optimisation, such as the use of different classifiers and optimised loss functions.

## 8 Acknowledgement

I would like to thank my supervisor Majid Latifi for his great support and encouragement throughout this whole project. Furthermore, I would like to thank Keith MacHale and Matt Tracey from Transit9 for the enlightening discussions as well as my partner Sandeep Khatri for fully supporting and motivating me during the project. Furthermore my great appreciation goes to Majid Latifi for presenting the research in computing module in an interesting and engaging way while always being available to answer queries and questions.

## References

- Abdulla, W. (2017). Mask r-cnn for object detection and instance segmentation on keras and tensorflow, [https://github.com/matterport/Mask\\_RCNN](https://github.com/matterport/Mask_RCNN).
- Amir, M. H. (2021). Damaged cars data.  
**URL:** <https://www.kaggle.com/muhammmadhamzaamir/damaged-cars-data>
- Bouarfa, S., Doğru, A., Arizar, R., Aydoğan, R. and Serafico, J. (2020). Towards automated aircraft maintenance inspection. a use case of detecting aircraft dents using mask r-cnn, *AIAA Scitech 2020 Forum*, p. 0389.
- Brooks, J. (2019). COCO Annotator, <https://github.com/jsbroks/coco-annotator/>.
- Cai, Z. and Vasconcelos, N. (2018). Cascade r-cnn: Delving into high quality object detection, *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6154–6162.
- Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F. and Adam, H. (2018). Encoder-decoder with atrous separable convolution for semantic image segmentation, *Proceedings of the European conference on computer vision (ECCV)*, pp. 801–818.
- European Consumer Centre Ireland (2020). Annual report 2019.
- Girshick, R. (2015). Fast r-cnn. arxiv 2015, *arXiv preprint arXiv:1504.08083*.
- Gong, C., Dacheng Tao, Wei Liu, Maybank, S. J., Meng Fang, Fu, K. and Yang, J. (2015). Saliency propagation from simple to difficult, *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2531–2539.
- Harshani, W. R. and Vidanage, K. (2017). Image processing based severity and cost prediction of damages in the vehicle body: A computational intelligence approach, *2017 National Information Technology Conference (NITC)*, IEEE, pp. 18–21.
- He, K., Gkioxari, G., Dollár, P. and Girshick, R. (2017). Mask r-cnn, *Proceedings of the IEEE international conference on computer vision*, pp. 2961–2969.
- Hoeser, T. and Kuenzer, C. (2020). Object detection and image segmentation with deep learning on earth observation data: A review-part i: Evolution and recent trends, *Remote Sensing* **12**(10): 1667.

- Kim, B. and Cho, S. (2020). Automated multiple concrete damage detection using instance segmentation deep learning model, *Applied Sciences* **10**(22): 8008.
- Kim, M., Woo, S., Kim, D. and Kweon, I. S. (2021). The devil is in the boundary: Exploiting boundary representation for basis-based instance segmentation, *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 929–938.
- Kumar, S. S., Devaki, K. et al. (2020). Assessing car damage using mask r-cnn, *arXiv preprint arXiv:2004.14173* .
- Kuo, W., Angelova, A., Malik, J. and Lin, T.-Y. (2019). Shapemask: Learning to segment novel objects by refining shape priors, *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9207–9216.
- Liu, Y.-C., Ma, C.-Y., He, Z., Kuo, C.-W., Chen, K., Zhang, P., Wu, B., Kira, Z. and Vajda, P. (2021a). Unbiased teacher for semi-supervised object detection, *Proceedings of the International Conference on Learning Representations (ICLR)*.
- Liu, Y.-C., Ma, C.-Y., He, Z., Kuo, C.-W., Chen, K., Zhang, P., Wu, B., Kira, Z. and Vajda, P. (2021b). Unbiased teacher for semi-supervised object detection, <https://github.com/facebookresearch/unbiased-teacher>.
- Liu, Y., Wang, Y., Wang, S., Liang, T., Zhao, Q., Tang, Z. and Ling, H. (2020). Cbnet: A novel composite backbone network architecture for object detection, *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34, pp. 11653–11660.
- Microsoft (2014). Common objects in context.  
**URL:** <https://cocodataset.org>
- Minaee, S., Boykov, Y., Porikli, F., Plaza, A., Kehtarnavaz, N. and Terzopoulos, D. (2020). Image segmentation using deep learning: A survey, *arXiv preprint arXiv:2001.05566* .
- Qianqian, Z., Sen, L. and Weiming, G. (2019). Research on vehicle appearance component recognition based on mask r-cnn, *Journal of Physics: Conference Series*, Vol. 1335, IOP Publishing, p. 012026.
- Ren, S., He, K., Girshick, R. and Sun, J. (2016). Faster r-cnn: towards real-time object detection with region proposal networks, *IEEE transactions on pattern analysis and machine intelligence* **39**(6): 1137–1149.
- Shah, A. (2019). Car damage detection.  
**URL:** <https://www.kaggle.com/anujms/car-damage-detection>
- Wu, Y., Kirillov, A., Massa, F., Lo, W.-Y. and Girshick, R. (2019). Detectron2, <https://github.com/facebookresearch/detectron2>.
- Zhang, Q., Chang, X. and Bian, S. B. (2020). Vehicle-damage-detection segmentation algorithm based on improved mask rcnn, *IEEE Access* **8**: 6997–7004.
- Zhou, Y., Wang, X., Jiao, J., Darrell, T. and Yu, F. (2020). Learning saliency propagation for semi-supervised instance segmentation, *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10304–10313.