Thesis for the degree of Doctor of Philosophy

Outlier Detection as a Safety Measure for Safety Critical Deep Learning

Jens Henriksson

Department of Computer Science and Engineering
CHALMERS UNIVERSITY OF TECHNOLOGY
Göteborg, Sweden 2023

Outlier Detection as a Safety Measure
for Safety Critical Deep Learning
Jens Henriksson
ISBN 978-91-7905-930-9

Doktorsavhandlingar vid Chalmers tekniska högskola
Ny serie nr 5396
ISSN 0346-718X

Department of Computer Science and Engineering
Chalmers University of Technology
SE–412 96 Göteborg, Sweden
Telephone + 46 (0) 31 – 772 1000

Printed by Chalmers Digitaltryck
Göteborg, Sweden 2023

Abstract

Context: Deep learning (DL) has proven to be a valuable component in object detection and semantic segmentation tasks, as the techniques have shown significant performance gains compared to hand-made image processing algorithms. DL refers to an optimization process where a model learns properties and parameters itself through in iterative process running on labeled data. The resulting model contains abstract features that are unintuitive to explain, thus challenging to ensure that the model will work as intended in safety critical applications (SCA).

Aim: The aim of this thesis has been to study how to connect parameters from DL with verification and testing for safety critical applications, and what extensions are necessary to verify deep neural networks. More specifically, this thesis has investigated the use of outlier detection as one testing method to detect when the model is operating on unfamiliar data.

Method: A comprehensive review of DL metrics and outlier detection metrics have been conducted. These metrics have been used to construct several new metrics to evaluate how the model behaves when encountering out-of-distribution (OOD) samples. An evaluation framework has been constructed that performs objective evaluation of OOD detection methods. The framework has been applied on various ranges of image datasets, starting with small scale images and continuing with realistic camera based use-cases from the automotive domain.

Results: This thesis found that one major issue with deployment of DL in SCAs is quantifying and tracing performance measures. The issue exists due to the difficulty in defining requirements and test cases for DL models, and expressing the models performance in safety related metrics. While DL performance is commendable, if the performance cannot be ensured, the technique should not be deployed in SCA. Our experiments show that the effect of OOD samples can be mitigated by extending the model with *safety measures*, i.e., measures that reduce the impact of undesired behavior. This thesis show how to use a *risk-coverage trade-off* metric that connects DL performance with functional safety requirements, such that safety engineers may allocate safety requirements on DL components and evaluate their performance.

Future work: Future works recommend testing the outlier detectors on further real world scenarios and how the detector can be part of a safety argumentation.

Keywords: automotive perception, out-of-distribution, outlier detection.

Acknowledgments

This research has been funded by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg foundation.

First and foremost, I would like to thank my academic supervisor and industrial supervisors. Christian Berger, my academic supervisor, for his continuous constructive and honest advice every time I have been in need. Your technical knowledge has supported me in exploring unique aspects in this thesis. And Stig Ursing, my industrial supervisor, your expertise has been of great value to ensure that what we have researched in this thesis remain relevant towards the industry. It is entertaining to work with you as every time we discuss a topic we solve one issue and define two new ones.

I would also like to thank all my colleagues at Semcon and Chalmers CSE department. Thank you for taking the time to listen to my explanations of my experiments and giving me constructive tips and tricks. Also, a big thank you to the members of the Vinnova research projects SALIENCE4CAV and SMILE. A special thank you to Markus Borg in the SMILE-project, who has been part of several of my studies and made me think about safety and requirements in a useful manner. Together we have made some great publications that are bridging the gap between deep learning and safety engineering.

Finally, I would like to express my gratitude to my family and friends. I want to thank my parents Regina and Lars-Olof for all your support and cheerful motivation. I want to thank my partner, Anna for your love and support and our daughter Thea for your smiles. If you read this when you are older Thea, remember that I used to be smart!

Jens Henriksson

Gothenburg, September 2023

List of Publications

Appended Papers

Paper I: **Jens Henriksson**, Markus Borg and Cristofer Englund. "Automotive Safety and Machine Learning: Initial Results from a Study on How to Adapt ISO 26262 Safety Standard", in *1st International Workshop on Software Engineering for AI in Autonomous Systems (SEFAIS)*, pp. 47-49, Gothenburg, Sweden, May 2018. IEEE/ACM

Contribution: Designed and conducted the interviews, lead and summarized the workshop discussions, structured and wrote the majority of the paper. The paper introduced the gap in standardization of safety mitigation techniques for machine learning.

Paper II: **Jens Henriksson**, Christian Berger, Markus Borg, Lars Tornberg, Cristofer Englund, Sankar Raman Sathyamoorthy and Stig Ursing. "Towards Structured Evaluation of Deep Neural Network Supervisors". In: *IEEE International Conference On Artificial Intelligence Testing (AITest)*, pp 27-34, San Francisco, USA, April 2019. IEEE

Contribution: Designed the comparison framework including datasets and metrics, conducted the baseline performance evaluation, structured and wrote the majority of the paper. The paper highlighted lack of testing measures for deep models but more importantly that existing testing techniques use ad-hoc evaluation.

Paper III: **Jens Henriksson**, Christian Berger, Markus Borg, Lars Tornberg, Sankar Raman Sathyamoorthy and Cristofer Englund. "Performance Analysis of Out-of-Distribution Detection on Trained Neural Networks". In: *Journal of Information and Software Technology*, Volume 130, 106409, February 2021. Elsevier

Contribution: Designed the experimental training and method comparison setup, conducted the comparisons, structured and wrote the majority of the paper. The paper did an in-depth study of out-of-distribution detectors, and saw that their performance vary when applied at different stages of the training procedure of deep models.

Paper IV: **Jens Henriksson**, Christian Berger and Stig Ursing. "Understanding the Impact of Edge Cases from Occluded Pedestrians for ML Systems". In: *2021 47th Euromicro Conference on Software Engineering and Advanced Applications (SEAA)*, pp 316-325, Palermo, Italy, September 2021. IEEE

Contribution: Designed the experimental training and method comparison setup, conducted the comparison study, structured and wrote the majority of the paper. The study found that parallel or expert systems can improve the prediction confidence.

Paper V: Markus Borg, **Jens Henriksson**, Kasper Socha, Olof Lennartsson, Elias Sonnsjö Lönegren, Thanh Bui, Piotr Tomaszewski, Sankar Raman Sathyamoorthy, Sebastian Brink and Mahshid Helali Moghadam. "Ergo, SMIRK is Safe: A Safety Case for a Machine Learning Component in a Pedestrian Automatic Emergency Brake System". In: *Software Quality Journal*, pp 1-69, March 2023. Springer

Contribution: Designed the outlier detection technique, method comparison setup, review of results and wrote major parts of the paper. The paper conducted a full use case implementation of a *methodology for Assurance of Machine Learning for use in Autonomous Systems (AMLAS)* and highlighted the strengths and challenges when deploying and testing machine learning based models.

Paper VI: **Jens Henriksson**, Christian Berger, Stig Ursing and Markus Borg. "Evaluation of Out-of-Distribution Detection Performance on Autonomous Driving Datasets". In: *IEEE International Conference On Artificial Intelligence Testing (AITest)*, Athens, Greece, July 2023. IEEE

Contribution: Designed the experimental training and method comparison setup, conducted the comparisons, structured and wrote the majority of the paper. The study evaluates the risk-coverage ratio on public autonomous driving datasets, and showcases that outlier detection is a useful safety mechanism to showcase safe usage in safety critical applications.

Additional Papers

The list of additional papers states additional publications made during the PhD studies. The additional papers introduced different aspects of *AI testing*, as it highlighted the need for evaluation of both testing of AI techniques and AI techniques for testing.

Paper A: **Jens Henriksson**, Christian Berger, Markus Borg, Lars Tornberg, Sankar Raman Sathyamoorthy and Cristofer Englund. "Performance Analysis of Out-of-Distribution Detection on Various Trained Neural Networks". In: *2019 45th Euromicro Conference on Software Engineering and Advanced Applications (SEAA)*, pp 113-120, Kallithea, Greece, September 2019. IEEE. ***Best Paper Award***

Contribution: Designed the experimental training and method comparison setup, conducted the comparisons, structured and wrote the majority of the paper. The study provided initial results on out-of-distribution detectors and their performance variation during training. The study was significantly extended in Paper III.

Paper B: Dhasarathy Parthasarathy, Karl Bäckström, **Jens Henriksson** and Sólrún Einarsdóttir. "Controlled Time Series Generation for Automotive Software-in-the-Loop Testing using GANs". In: *IEEE International Conference On Artificial Intelligence Testing (AITest)*, pp 39-46, Oxford, UK, August 2020. IEEE

Contribution: Designed the variational autoencoder architecture that was part of the test case exploration. Wrote minor parts of the paper and collaborated cross-company. The study investigated generating test cases through GANs by using linear interpolation as guidance for test case generation.

Paper C: Kasper Socha, Markus Borg and **Jens Henriksson**. "SMIRK: A machine learning-based pedestrian automatic emergency braking system with a complete safety case". In: *Journal of Software Impacts*, Volume 13, 100352, August 2023. Elsevier

Contribution: Reviewed a majority of code for the release of the software, and wrote parts of the publication. The publication demonstrated and released a fully transparent driver-assistance system called SMIRK and added distribution of the dataset constructed in Paper VI. The dataset was built using the automotive simulator ESI Pro-Sivic.

Paper D: **Jens Henriksson**, Stig Ursing, Murat Erdogan, Fredrik Warg, Anders Thorsén, Johan Jaxing, Ola Örsmark, Mathias Örtenberg Toftås. "Out-of-Distribution Detection as Support for Autonomous Driving Safety Lifecycle". In: *Requirements Engineering: Foundation for Software Quality: 29th International Working Conference, REFSQ*, pp 233-242, Barcelona, Spain, June 2023. Springer

Contribution: Designed, reviewed, wrote the majority of the paper, and organized workshop sessions to collaborate cross-company to define a suitable safety lifecycle and showcased how out-of-distribution detection acted as support for this lifecycle. The paper described a vision of how abstraction levels for an automotive safety feature that could be detailed in lower levels that allowed for allocation of safety requirements on ML components.

Contents

Abstract **i**

List of Publications **v**

1 Introduction **1**
1.1 Background . . . 2
1.2 Aim and Research Questions . . . 5
1.3 Scope and Delimitations . . . 6
1.4 Outline . . . 7

2 Frame of Reference **9**
2.1 Metrics . . . 9
2.2 Safety and Standardization . . . 11
2.3 Outlier Detection . . . 13
2.4 Testing with Artificial Intelligence . . . 15

3 Research Approach **17**
3.1 Combining Metrics with Risk Reduction . . . 20

4 Summary of Appended Papers **25**
4.1 Paper I . . . 25
4.2 Paper II . . . 25
4.3 Paper III . . . 26
4.4 Paper IV . . . 26
4.5 Paper V . . . 27
4.6 Paper VI . . . 27
4.7 Paper Contributions to Research Questions . . . 28

5 Discussion **29**

6 Conclusions and Future Work **31**

Bibliography **33**

List of Figures

1.1 A visualization of the varying image dimensions that this and related works are evaluating on. Values below refer to the amount of parameters there are per sample in the sets. 5

2.1 Visualization of the goal of SOTIF: To reduce the risk of encountering unsafe unknown scenarios, by detecting and categorizing scenarios (HARA) and applying risk mitigation strategies to handle them [4]. . 11
2.2 A comparison between anomaly detection (AD), Open Set Recognition (OSR), and Out-of-Distribution (OOD) detection. To the left, a hypothetical set of samples, and automotive examples of a training (a regular image), anomalous (with lens scratch), or novelty image (new driving scenario). Images are from the Audi Autonomous Driving Dataset [36]. 13
2.3 A training sample to the left, with semantic segmentation labels to the right from the Cityscapes training set [22]. Each color in the label corresponds to one specific class, e.g., dark blue represents cars in the dataset. 14

3.1 A historical view of how research studies have led into the succeeding studies. Papers I through VI are the appended papers in this thesis. . 18
3.2 The AUROC score visualized with regards to the achieved testing accuracy for two state-of-the-art DNN architectures. The experiments were tested with and without training augmentation (marked with A in the figure legend) and adjusted learning rate (marked L in the figure legend) for two variants of OOD detection, namely Baseline (BL) [41] and OpenMax (OM) [44] 19
3.3 In Paper II, outlier detection was performed between CIFAR-10 (inlier set) and Tiny ImageNet (outlier set). The figure shows the histogram of the anomaly scores and the corresponding ROC-curve for one OOD detector. 22
3.4 A visualization of the risk-coverage trade-off curve for the training set of Paper VI. The curve is achieved by varying the accepted anomaly distance threshold. 23

Chapter 1

Introduction

During the last decade, the field of computer vision has become a pivotal part of automation as a result of drastic improvements in performance on the back of Deep Learning (DL). The ImageNet Large Scale Visual Recognition Challenge (ILSVRC) [1] ran between 2010-2017, where the top-5 error rate was reduced from 28% (the best result in 2010) to 2.3% in 2017, compared to the human error that is estimated to be 5.1% [1]. Since 2012, the top contenders incorporated deep learning through a Deep Neural Network (DNN), a layered matrix multiplication structure that incorporates non-linearity through activation functions that are applied after each layer.

With the promising performance of DL, several domains have successfully incorporated DNNs in their products, e.g., improved web searching, recommendation systems in advertisement and streaming content, and voice assistants like in consumer smartphones are just a few examples of domains that have been revolutionized [2]. For Safety Critical Applications (SCA) however, the incorporation of DL is not as straight-forward as in non-critical applications. The automotive domain has a long history of incorporating new automation features that has gone through rigorous safety verification before being released to the public. However, the existing frameworks and standards were not designed with Machine Learning (ML) algorithms in mind, thus they fall short in ensuring that the intent of the ML model is consistent [3]. ISO standards such as ISO 21448 SOTIF [4] and the upcoming ISO/PAS 8800 [5] aim to address these issues, but they do not elaborate on how to evaluate safety measures for DL models.

This thesis has detailed and evaluated how outlier detection methods, specifically from the subfield of out-of-distribution (OOD) detection, can be used as a safety measure that works on DNNs in SCAs. The safety measure is demonstrated by connecting ML performance metrics with safety related metrics on varying complexity of image data, ranging from small scale images of digits to large scale automotive datasets. As there is a need to bridge the gap between safety assurance and DNN

metrics, OOD detection provides one qualitative metric that can be utilized in safety argumentation throughout the safety lifecycle of autonomous driving.

The remainder of this section elaborates on the background of this thesis, specifically outlier detection and the advancements of automotive safety and what safety standards are available for DNNs. The background lays the foundation to the research questions and the corresponding limitations of the conducted studies.

1.1 Background

The automotive sector is a prime target for increased automation, as it can increase public accessibility and improve road safety [6]. Advanced Driver Assistance Systems (ADAS) have improved comfort and automation and are already incorporated into vehicles, since they save lives by preventing or reducing the impact of accidents. Estimates suggest that integration of ADAS can potentially reduce up to 40% of accidents [6]. While ADAS provide partial automation, e.g., through adaptive cruise control and lane-keeping assistance, the driver is always responsible for monitoring the driving tasks and taking control when the function exits its defined operational design domain. [7]. However, for Autonomous Driving Systems (ADS) with high level of automation, the fallback responsibility should instead be put on the system, as described in the SAE J3016 standard [8]. However, transferring the fallback to the system requires more effort in testing, verification and validation to ensure safe deployment before vehicle manufacturers can accept the responsibility for self-driving vehicles.

One of the essential tasks to enable both ADAS and ADS is constructing situational awareness of the vehicle, also referred to as perception. Perception is achieved through fusion of sensor information that perceives the surroundings of the vehicle, typically achieved through the use of cameras, radars and LiDARs [9]. For ADAS, the primary sensor setup includes sensor fusion between radar for object detection and classification through traditional camera algorithms that are not derived through ML.

For ADS, high precision of perception is crucial as a majority of decisions are based on how the environment is perceived. Hence, incorporating the advances from data-driven algorithms into the autonomous driving industry can improve perception and thereby improve safety. Data-driven algorithms refer to methods that are not explicitly programmed to perform a task, but instead learn from data samples and adjust their algorithm based on the training data received. A DNN is a data-driven method that constructs a chain of layered computations, where each layer consists of a set of nodes (also known as neurons) that receives a weighted sum of inputs from the previous layer and combines the sum with a non-linear activation function. In practice, the amount of layers and nodes used in a DNN is selected empirically, where practitioners will experiment with various configurations. The amount of parameters

in the DNN grows rapidly with depth and width of layers and thus, it will quickly become infeasible to scrutinize.

Given the importance of high-precision perception for autonomous driving, development guidelines and evaluation standards are necessary. ISO 26262 [10] is an international standard for the automotive industry that was originally released in 2011 and further updated in 2018. The standard provides guidance for electric and electronic systems in road vehicles, and describes the full safety lifecycle from concept and development phase to operational and maintenance phase. The main goal of ISO 26262 is to provide a systematic approach to functional safety that provides risk reduction techniques through tools or ways-of-working guidelines that historically have reduced failures in the systems that could lead to harmful accidents. The potential hazards are gathered through hazard identification on an item, where an item is defined as a system or component that is under study. For each hazard identified, the severity (i.e., potential harm), the exposure (i.e., how often the hazard can happen), and controllability of the hazard are assessed. Depending on these three factors, an Automotive Safety Integrity Level (ASIL) is assigned to each hazard, ranging from A to D, representing the amount of risk reduction techniques needed, where ASIL D requires the highest level of safety requirements.

The risk reduction techniques from ISO 26262 have evidently been beneficial for development of traditional algorithms in the automotive industry. However, the safety standard was not designed to consider DNN algorithms [3]. *ISO 26262 Part 6: Product development at the software level* introduces 75 risk reduction techniques, from which 34 apply at the unit level and the remaining on the architecture level. For a hazard with ASIL D, roughly two out of three techniques are either not applicable for ML or would require adaptations from the safety standard to be applicable [3].

One risk reduction technique that would require adaptation to be applicable for software unit testing is fault injection, which refers to injecting arbitrary faults, e.g., through corrupting values or variables to test robustness of the system without compromising the system. While this method is technically applicable to DNNs, a randomly modified or corrupted weight of a convolutional DNN will not indicate robustness as a model can contain several millions of parameters and the amount of permutations needed to evaluate robustness is practically infeasible. The method can be modified however, e.g., to describe a Bayesian approximation as suggested by Gal et al. [11]. Dropout layers are used during training as a regularization technique to reduce overfitting of DNNs and normally turned off during testing. However, Bayesian dropout refers to enabling the dropout layers during testing to construct an uncertainty estimation by processing the sample multiple times through the network. This technique could be a suitable DNN replacement for fault injections, as it would study how the model behaves when certain parts of the model are cancelled out.

The item definition and hazard analysis and risk assessment from ISO 26262 are applicable to DNNs, but the risk reduction techniques that are to be used need to be designed for data-driven algorithms. This is partially a reason why several

frameworks and standards have been released, e.g., *ISO 21448: Safety of the Intended Functionality* (SOTIF) [4] and *Assurance of Machine Learning for use in Autonomous Systems* (AMLAS) framework [12] that aim at bridging the gap by assuring the performance of DL through different ways of working.

The potential risk reduction techniques for ML have been an engaging topic recently, as there is an inherent interest in identifying limitations of ML models, specifically when applied to SCA. Mohseni et al. [13] describe three safety strategy categories with open challenges: inherent safe design, safety margins, and safe fails. For inherent safe design, structuring and incorporating design and implementation transparency is discussed as safety strategies, and different forms of domain generalization [14] with robustness evaluation for perturbation and corrupted data as safety margins [15]. For safe fails, uncertainty estimation and sample evaluation for outlier detection is suggested.

This thesis considers outlier detection as the generic field to detect samples outside the scope of the system. The field contains several subfields, where Out-of-Distribution (OOD) detection is one suggested safe fail for automotive [13]. The subfield in OOD detection is closely related to Anomaly Detection (AD) and Open Set Recognition (OSR), compared in Section 2.3. In short, AD refers to the detection of anomalous samples that are in-distribution (i.e., normal images) but erroneous in some form, e.g., through sensor anomalies such as scratches on camera lenses providing a tear in the image or dirt that pollutes the image. In most cases, ML models are trained as closed-set classification models, i.e., trained where each input sample belongs to one of the pre-defined classes of the model. If a sample is provided to the model that is outside the pre-defined classes, the model will still provide the most probable prediction. OSR is a form of novelty detection that addresses this issue by acknowledging the problem and includes an *unknown unknown* class, similar to how SOTIF defines their unknown scenarios, that can be used during testing for samples that does not fall into the pre-defined classes. Last, OOD detection aims at separating test samples that from different distributions compared to the training distribution. For evaluation of OOD performance, the aim is to have an out-distribution that does not overlap with the in-distribution. However, this is rarely the case, specifically for real-world applications [16].

The fields AD, OSR, and OOD detection are all vital for ML to be deployed in SCA [17], as performance may degrade on data outside the scope of the ML model. Unfortunately, the vast majority of related work is applied on small scale images. The issue with these images is that the results are not directly transferable to realistic use cases, where the camera resolution can be several orders of magnitude larger. ImageNet [1] is considered the most well-rounded dataset for classification and is commonly used as the in-distribution dataset for OOD evaluation. In related OOD evaluation works, the evaluation is done on classification networks, i.e., not designed for semantic segmentation networks or object detection and classification networks in realistic scenarios, something that is addressed in this thesis. Furthermore, there is an interpretation gap of how to translate computer vision metrics into safety-related

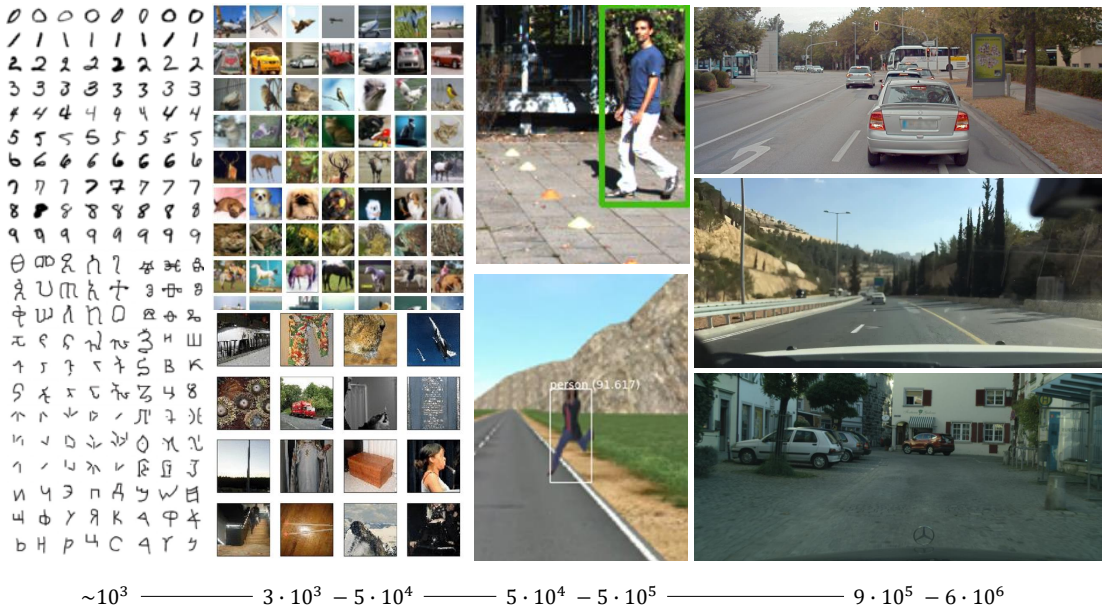

Figure 1.1: A visualization of the varying image dimensions that this and related works are evaluating on. Values below refer to the amount of parameters there are per sample in the sets.

performance guarantees, as well as a gap of how to evaluate realistic performance guarantees for ML models.

1.2 Aim and Research Questions

The difficulty in incorporating computer vision DNNs into SCA have indicated that more research is needed to bridge the gap between ML model development, and safety verification and validation. Thus, describing what constitutes a safety measure for DNNs and what determines its usefulness for arguing integrity of a system that incorporates DL is of interest in this thesis. Furthermore, it has not been shown how the safety measure performs when applied to realistic automotive images. To this end, two concrete research goals are formulated for this thesis:

G1: To Understand and motivate testing with safety measures for DNNs, with ADS as target application.

G2: To understand the translation from a DNN metric, e.g., accuracy, to a concrete safety requirement allocated on a DNN.

In this thesis, safety measure evaluation is investigated from the DL perspective, i.e., what metrics can DL provide, and how can they be used in the lifecycle of

autonomous driving. Furthermore, the high-dimensional data aspect is studied with varying image dimensions, e.g., Fig. 1.1, where automotive images with up to $2 \cdot 10^6$ pixels have been studied. Specifically, these research questions have been addressed in the research:

RQ 1: What would constitute suitable metrics that can be extracted from DNNs and utilized as safety measures in the automotive verification?

RQ 2: How can OOD detection be formulated as one potential safety measure?

RQ 3: At what stages can OOD detection be beneficial to the safety lifecycle of autonomous driving features?

By studying these three research questions, the aim is to find a way to translate DNN terms, i.e., model accuracy or precision that are commonly used during training, in such a way that the metrics may be used in requirements breakdown and evaluation that the DL models perform in a safe and controlled way. An illustrative example of average precision in a requirements form is "*In a sequence of images from a video feed any object to be detected should not be missed more than 1 in 5 frames.*" described by Gauerhof et al. [18].

1.3 Scope and Delimitations

The scope of this thesis is to follow OOD detection performance on varying inputs, DNN models, and evaluation sets with the goal of applying state-of-the-art OOD detection methods on complex automotive scenarios. Initially, this thesis will focus on well-known datasets from academia, e.g., MNIST [19] and CIFAR [20], and afterward continue towards full-sized images as in KITTI [21] and Cityscapes [22]. Note that the tasks in the aforementioned datasets vary, hence the application of the OOD detection method needs appropriate adjustments, an aspect that is addressed in Section 3.1.

For automotive, and in general any system with continuous input streams, the history provides significant information for detection and classification systems. However, due to lack of access to publicly available datasets with this feature, the time-series aspect had to be disregarded during the studies, and instead, all OOD metrics evaluated are considered as per-image basis.

Also, to widen the exploration of OOD methodology, the training aspects of the DNNs are partially neglected in some studies. The upsides to neglecting the training are resulting models are pre-trained with proven performance, as well as allowing the focus to be put on evaluating the OOD methods and metrics more thoroughly, and less time spent on training and hyperparameter search of models.

1.4 Outline

The remainder of this thesis is structured as follows:

Section 2: *Frame of Reference* - Introduces metrics, standardization and frameworks, specifically SOTIF [4] and AMLAS [12]. Related works in OOD methodology are presented, and described in context of the frameworks. Finally, alternative AI testing techniques are presented that could work in parallel with this thesis.

Section 3: *Research Approach* - Introduces the connection of appended papers and how they contribute to the research questions of this thesis. Furthermore, the metrics derived in this thesis are presented, specifically the *risk-coverage trade-off* metric that has been prominent in all studies, and how it can be used for risk reduction.

Section 4: *Summary of Appended Papers* - Presents the key findings for each of the six appended research publications that are included in this thesis and describes how they contribute to the research questions formulated in Section 1.2.

Section 5: *Discussion* - Combines the contributions from the six appended publications and highlight their combined contribution. The contributions are further elaboration upon and used to answer the research questions are formulated in this thesis. Furthermore, the drawbacks and threats to validity are discussed.

Section 6: *Conclusions and Future Work* - Summarizes and concludes the key messages from this thesis, and offers a final remark of this research.

Chapter 2

Frame of Reference

2.1 Metrics

To be able to link OOD detection with safety measures and be framed within the context of AI testing, one first has to acknowledge the confusion matrix that is essential to extracting true and false positive ratios. The *Encyclopedia of Machine Learning and Data Mining* defines the confusion matrix as

> "*A **confusion matrix** summarizes the classification performance of a classifier with respect to some test data. It is a two-dimensional matrix, indexed in one dimension by the true class of an object and in the other by the class that the classifier assigns. [...] A special case of the confusion matrix is often utilized with two classes, one designated the positive class and the other the negative class. In this context, the four cells of the matrix are designated as true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN)*" [23]

The matrix is comparing a predicted condition to the actual condition, visualizing the performance overlap. The four cells described in the confusion matrix are the basis for several evaluation metrics used in OOD detection and DNN training in this thesis, as well as related works.

The four cells reoccur, but in different forms both in SOTIF and OSR [24]. The fields of OSR and OOD detection do not explicitly state the four cells, but incorporate the "unknown" class for OSR and consider in and out-distributions in OOD. SOTIF has a related approach, where the scenarios are categorized based on two properties: whwther a scenario is known, and if a scenario is safe. Combining the two properties, all scenarios can be categorized into four areas: known safe, known unsafe, unknown safe, and unknown unsafe scenarios. If one consider the positive samples in the

confusion matrix as a safe scenario from SOTIF, a connection between the confusion matrix and the SOTIF areas is created. Safety measure are then applied to ensure that the ADS is only operating on positive samples, i.e., samples that are considered safe, and rejects operation on scenarios deemed unsafe.

The confusion matrix is used for several metrics and visualization tools [25]. The Receiver Operating Characteristics (ROC) curve is a common tool used to visualize the trade-off between correct behavior, i.e., true positive rate, and false alarms, i.e., false positive rate, which is adopted in this thesis. For object detection and semantic segmentation, Intersection over Union (IoU) is the most common metric [26]. IoU, also known as the Jaccard index for tasks outside of computer vision, is computed as the intersection between prediction and ground truth, i.e., TP, divided over the union (TP + FP + FN), thus providing a score between 0 (no overlap) and 1 (full overlap) between the prediction and ground truth. In object detection tasks, the object is considered detected if the score is above a certain threshold, e.g. above 0.5.

In traditional OOD detection, the task is to separate samples to either be included or excluded for processing. One drawback for this approach is the fact that it does not consider whether the model output will be correct or not. OOD detection can be applied in this case to also consider risk of misclassifications, and not only if the sample is in-distribution or not. This approach is applied in selective prediction, a subfield within ML that considers extending the ML model with a reject option [27]. The main point of the field is to reduce the false positives by allowing the model to acknowledge its limitation, something that can reduce the classification risk by reducing test coverage. This way the precision of the output will increase, but the model will predict fewer samples, something that is referred to as the *risk–coverage trade-off* and has been shown with the reject option for several classification and regression tasks [28]. This thesis extends the risk-coverage trade-off with the OOD detection safety measure and puts it into context for the automotive safety lifecycle.

OOD detection and the confusion matrix are typically used for classification challenges. However, pure classification tasks are seldom the use case for ADS. Instead, ADS relies on semantic segmentation or object detection and classification models. OOD detection can be applied in the same way to these models by considering the pixels in semantic segmentation tasks and each bounding box in an object detection and classification model as an independent classification task. This way, the OOD detection method can be applied on each independent task, thereby excluding regions of pixels or a set of bounding boxes for each image processed through the DNNs, something that has been tested and presented in Papers IV-VI. Note that in traditional OOD detection, the positive condition is if a sample should be excluded or not. Thus, a true positive is an outlier that has correctly been rejected. For semantic segmentation and object detection and classification networks, the performance is instead reported in risk-coverage trade-off that is accessed through varying the accepted distance threshold.

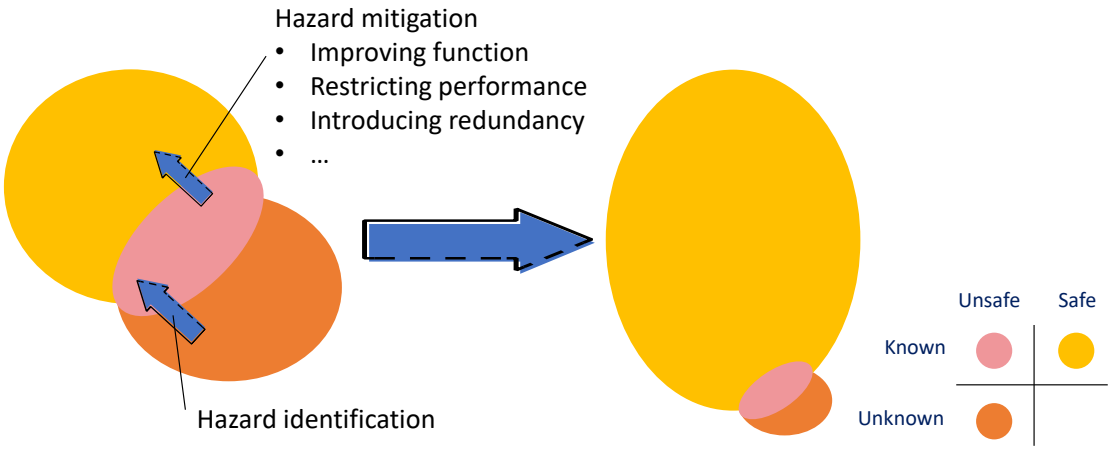

Figure 2.1: Visualization of the goal of SOTIF: To reduce the risk of encountering unsafe unknown scenarios, by detecting and categorizing scenarios (HARA) and applying risk mitigation strategies to handle them [4].

2.2 Safety and Standardization

During the time of this thesis, AI safety became a well discussed topic. Initially, the peculiar issues that occurred in ML were reviewed and discussed, e.g., issues in poorly defined goals, unknown negative side effects, model robustness and the ability to conduct safe exploration of new input variations [29]. In addition, the need to understand how DNNs generalize has been discussed, as large DNNs tend to overfit and can fit to random labels [14], thus raising the question of how to properly test for generalization.

ISO 21448 SOTIF (Safety of the Intended Functionality) was originally released as a publicly available specification in 2019, and further released in 2021 with the goal to provide measures to eliminate hazards or reduce risk [4]. SOTIF aims at identifying system insufficiencies or hazardous scenarios through these measures (referred to as methods in the standard), visualized in Fig. 2.1.

In SOTIF, the scenarios are distributed over four areas: safe known, unsafe known, unsafe unknown, and unknown safe scenarios. The process of SOTIF reduces the amount of unsafe scenarios through 103 risk reduction methods, where 47 aim at analyzing and identifying functional insufficiencies and foreseeable misuse, and the remaining 56 aim at verifying sensing, planning, actuation, and integration, as well as evaluating residual risk of unknown scenarios that can occur in real-life situations. SOTIF provides guidelines on *what* methods to apply, whereas this thesis describes *how* to use a safety measure.

Assurance of Machine Learning in Autonomous Systems (AMLAS) [12] was released in early 2021 with the goal of establishing justified confidence in ML models, specifically for the use in applications that are safety critical. The publication contains guidance to integrate safety assurances into ML components and connect them to an evidence

base that justifies that safety is achieved in the components. AMLAS provides an iterative process containing one scoping phase and five assurance phases that together bridge system safety requirements to a safety case for the ML component that can be followed throughout requirements' allocation, data management, model training, model verification, and model deployment.

The AMLAS guidance specifies 33 objectives and artifacts that utilize Goal Structured Notation (GSN) for the objectives, something they have previously shown to provide confident arguments for safety argumentation [30]. AMLAS provides several examples for the artifacts provided, as well as notes for elaborating common practices from literature. In similarity to SOTIF, this thesis follows outlier detection as one safety measure that can act as a method to provide the artifact output needed for their ML requirements assurance.

Burton et al. [31] discusses the challenges in deriving an acceptance criterion for ML components. Their publication construct and motivates safety contracts for ML components and exemplifies it with contracts for a pedestrian detection DNN. Furthermore, they connect their safety contract to the confusion matrix and argue how safety analysis may set different goals on FP and FN.

One recommended task from Mohseni et al. [13] is to perform formal verification of neural networks. Formal verification of DNNs has been shown by Ehlers [32]. However, to correctly verify a system, a specification of the task at hand is needed. This is an infeasible requirement for object detection and semantic segmentation DNNs, as there is no specification for what constitutes a specific object class [33], e.g., it is infeasible to specify what the class *pedestrian* constitutes in pixel-space due to the vast amount of permutations. Instead, the performance of a DNN is evaluated with accuracy and precision metrics based on a given set of test data. This thesis utilizes these metrics to evaluate *risk* of misclassifications within a DNN that can be studied throughout experiments. In addition to formal verification challenges, Seshia et al. [34] identify four additional major challenges to achieve formally-verified DNNs, focusing on the difficulties in environmental modelling, DNN learning techniques, scalable design for verification and validation of models and data.

Another issue found through literature reviews are the lack of experiments on automotive datasets. Tambon et al. [35] reviewed how to certify ML models in SCAs and found that a majority of methods presented are evaluated on small-scale datasets e.g., MNIST and CIFAR, where results are not guaranteed to be transferrable to realistic automotive challenges. However, their taxonomy aligns with Mohseni et al. [13] and Seshia et al. [34] in asserting that explainability, uncertainty, and robustness are key factors for certification.

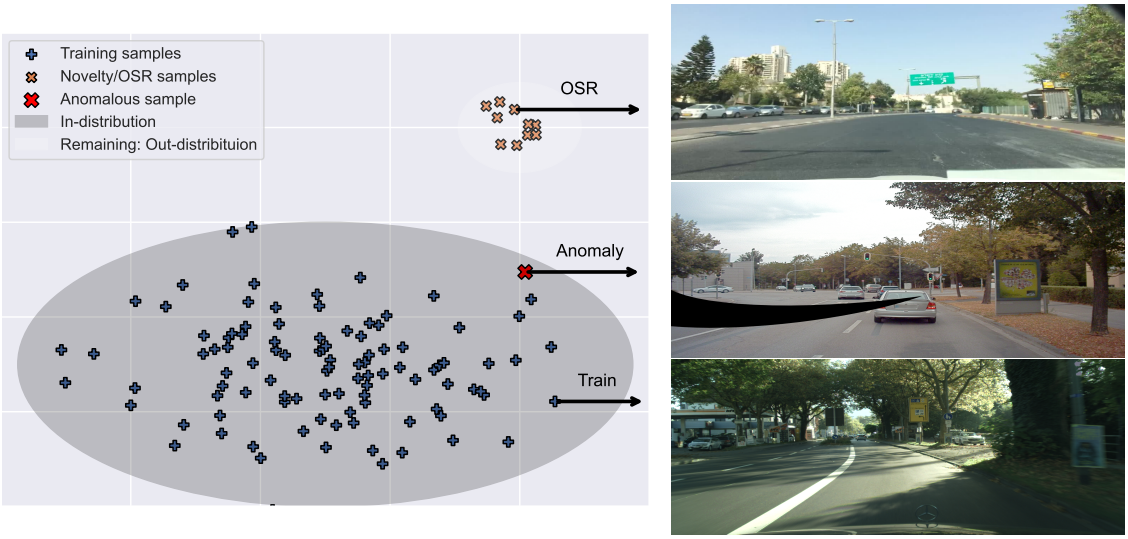

Figure 2.2: A comparison between anomaly detection (AD), Open Set Recognition (OSR), and Out-of-Distribution (OOD) detection. To the left, a hypothetical set of samples, and automotive examples of a training (a regular image), anomalous (with lens scratch), or novelty image (new driving scenario). Images are from the Audi Autonomous Driving Dataset [36].

2.3 Outlier Detection

This section summarizes the definition of outlier detection for this thesis, by comparing AD, OSR, and OOD detection, visualized in Fig. 2.2. In addition, surrounding works in evaluation of robustness of DNNs through the scope of outlier detection are discussed. It is worth noting, that the majority of research uses the terms interchangeably, outlier detection and similar synonyms (i.e., anomaly, novelty, outlier, OOD, and OSR), have been researched for more than a century, covering a wide set of approaches [37]. Geng et al. [38] conducted a survey on OSR, and summarized datasets, evaluation metrics, and algorithm comparisons in the field this thesis aligns to. Specifically, they describe the four base categories (referred to as four cells in Section 2.1) such that OSR terminology overlaps with SOTIF and the confusion matrix. OOD detection simplifies this to consider all erroneous samples to be OOD [17].

Initial experiments used principal component analysis and one-class support vector machines to implement anomaly detection, but in the past years more sophisticated methods have been applied [39]. Most OD methods belong to either distance based, reconstruction based, probabilistic or classification based variants, where this thesis has focused on distance and reconstruction based methods. Reconstruction based methods refer to compressing the input information through some bottleneck, followed by decompression and studying the reconstruction error. One common method to do this is with variational autoencoders [40], since they can be constructed with a suitable bottleneck dimension, as well as varying dimensions of the encoder/decoder part such that advanced features can be extracted.

Figure 2.3: A training sample to the left, with semantic segmentation labels to the right from the Cityscapes training set [22]. Each color in the label corresponds to one specific class, e.g., dark blue represents cars in the dataset.

For distance based anomaly scores, some related work can be seen as milestones in the field. Hendrycks et al. [41] introduced a baseline approach to OOD detection for DNNs. The approach consisted of studying the softmax output of a DNN, and concluded that the magnitude of the most probable class could be used as an inclusion/exclusion criterion. Additional techniques to detect OOD samples have been used, e.g., training anomaly detectors by exposing them to outlier samples during training, such that it can generalize and detect previously unseen outliers [42].

Liang et al. [43] continued the work, and proposed reversed temperature scaling. They noticed that a stronger separability could be achieved by modifying the input with a small perturbation based on what was the most probable prediction for the model. The added perturbation affected the inlier samples significantly, while only marginally affecting the outlier samples, thus it could be used as a separability measure. From OSR, the work by [44] et al. constitutes a milestone as they modified the softmax layer into an *Openmax* layer which estimates the probability of an input to belong to a unknown class. One of their key findings was that a meta-recognition concept could be applied to the penultimate layer (i.e., the layer prior to the final activation function), and thereby get a statistical measure to reject unknown samples. Finally, Lee et al. [45] utilized a statistical approach to the OOD detection field by computing an outlier score based on the Mahalanobis distance for an input sample with regard to the most probable class distribution. Furthermore, their experiments saw even better results when combined with temperature scaling in [43]. Lee et al. experimented on several datasets, however not on automotive datasets. This thesis continued with the Mahalanobis distance but adjusted to work on semantic segmentation challenges, including large scale automotive images.

Applying outlier detection to automotive data is a rather new task, and even less explored for semantic segmentation. Semantic segmentation refers to each pixel being classified in an image, demonstrated in Fig. 2.3. Applying AD to a semantically segmented image results in each pixel being evaluated independently. A recent publication reviewed AD for ADS between the years 2015 and 2022 and showed that some methods have been applied and tested on camera, radar, and LiDAR data [46]. Their review shows 21 different approaches on a broad set of datasets, where only a handful attempts were applied on Cityscapes [22], Berkeley Deep Drive [47], and

Audi Autonomous Driving Dataset (A2D2) [36]. In fact, A2D2 [36] is rather new and only a handful of DNN training approaches have been tested, and even less OOD methods [48]. Prominent approaches from [46] include Bayesian SegNet [49], one of the first application of uncertainty to a semantic segmentation dataset. Their approach constructed a saliency map by applying Monte Carlo dropout sampling and assessing the variance as uncertainty. Furthermore, a dataset referred to as *Fishyscapes* [50] that could be utilized for future work is introduced. The dataset combines Lost and Found [51] with Cityscapes [22] for the purpose of outlier detection on Cityscapes trained models. This work complements those studies, by significantly extending OOD methods to be applicable on automotive datasets.

2.4 Testing with Artificial Intelligence

Deploying AI in SCA requires adjustments of testing, but AI can also be utilized to improve testing procedures. Several works are utilizing the advances of DL to construct artificial test cases that can extend the test coverage for autonomous driving. Ribeiro et al. [52] notices that *trust* is an important aspect to adopt DNNs into SCAs, and thereby introduces a method to introduce interpretable representations from any model. The representations allow a prediction to highlight which parts of an image motivate the prediction. Wagner and Koopman [53] also highlight the need for building trust in autonomous driving as well as ML.

Generative Adversarial Networks (GANs) have also been introduced as a way to construct additional training samples. Carlini and Wagner [54] utilize adversarial samples to estimate the robustness of DNNs, and describe two approaches to evaluate DNNs: Proving a lower bound of the performance or demonstrating an upper limit based on adversarial attacks. Since the field of GAN mainly focuses on synthesizing and detecting adversarials, it has been excluded in this thesis.

Pei et al. [55] introduce DeepXplore, a tool aiming at detecting erroneous behavior for DNNs by leveraging neuron coverage of the models. This is used to systematically measure parts of the DNN that are prone to misbehave. By following neuron coverage, thousands of edge cases can be discovered in state-of-the-art DNNs. Ma et al. [56] apply a similar approach and construct sets of test cases through generation of adversarial samples that maintain neuron coverage. Last, Tian et al. [57] also utilize neuron coverage for guided test generation. However, instead of adversarial samples, they construct realistic test cases with traditional computer vision effects, such as stretching, skewing, and tearing, but also adding weather effects such as rain and fog.

Chapter 3

Research Approach

This thesis has investigated and systematically evaluated how verification of high-dimensional DNNs can be conducted in practice. Early in the process, it became evident that the performance gains by adopting ML is only useful as long as the results can be systematically assessed for quality desired quality properties. That is, it is better to operate with worse performance if the limitations can be known, compared to operating with excellent performance, but the system can fail unexpectably. The work conducted in this thesis has continuously been aligned with plausible use-cases such that it is relevant towards the automotive industry, to ensure the methodology is providing valuable insights. Alignment has been achieved through collaborations with both industry and academia through workshops and research projects.

The thesis has operated in a waterfall flow, where results from the previous study lead into the succeeding one. This flow is visualized in Fig. 3.1, and are categorized in *industrial relevance*, *method screening and evaluation*, and *automotive use case*. The remainder of this section walks through the studies, followed by Section 3.1 that combines the results from the attached papers to present the final evaluation metrics from DNNs that are recommended for outlier detection, and then how to utilize the outlier detection as a safety measure that is applicable for ADS.

The initial research aimed at defining what the major issues are that disrupts verification of DNNs for SCAs. More specific, the target audience was autonomous driving, thus a suitable initial point was the ISO 26262 standard that is prominent in the automotive industry. By studying the standard and interviewing two experts (Paper I), followed by workshop discussions, it became evident that the standard could not handle the complexity of DNNs nor was it designed to do so. This sparked the need to review what metrics exist that are used when evaluating DNNs, as well as what OOD methods exist, which motivated an additional literature review that were conducted through two layers of backwards and one layer of forward reference-snowballing approach with Hendrycks et al. [41] as initial seed, as the paper initiated the field. This approach yielded two contributions, 1) a M.Sc. thesis that

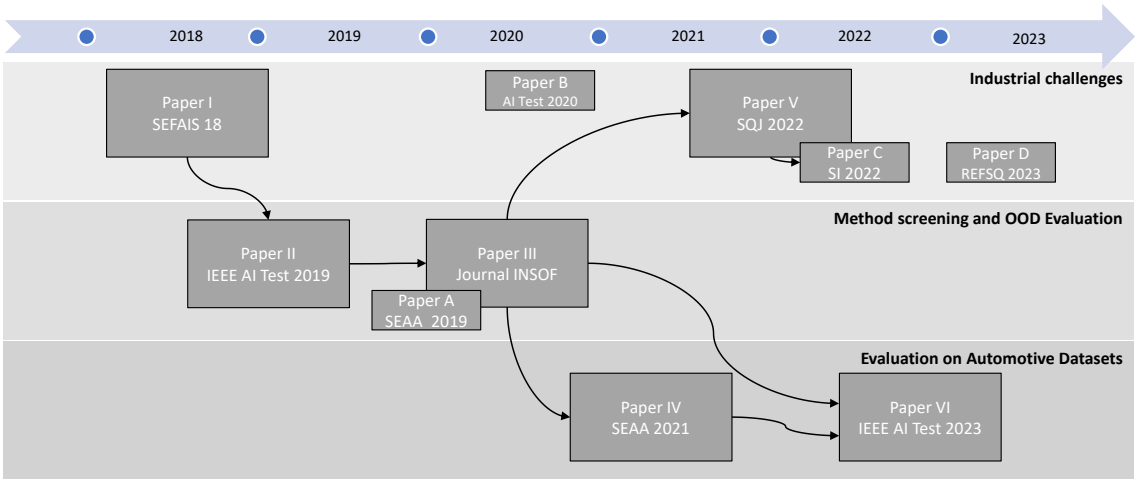

Figure 3.1: A historical view of how research studies have led into the succeeding studies. Papers I through VI are the appended papers in this thesis.

reviewed outlier detection methods. The thesis reviewed 18 methods, from which five were implemented and tested by Landgren and Tranheden [16], and 2) the variations in metrics and lack of comparable evaluation between outlier detection methods in Paper II. Furthermore, it was noted that a majority of studies used different models and datasets, thus the performance in the studies are not comparable.

To combat the issue of unfair comparisons, a comparison study (*Paper A*) was formulated as a way to provide a fair evaluation of the most prominent outlier detection methods that had been found at this stage of research. In addition, it was noted during the experiments in Paper II that overfitting had an effect on outlier detection. Therefore, in Paper A two OOD methods were compared on two DNN architectures that were re-trained from scratch and studied for OOD detection performance during the training cycle to see if the outlier detection performance, i.e., the AUROC score, was correlated with the accuracy of the DNN. The results can be seen in Fig. 3.2, and the concept was well received by the academic community, resulting in an extended version of the concept in Paper III. The extended version compared three OOD methods on four trained models as well as three outlier sets. Both the initial paper, and the extended version showed that there exist correlation between model accuracy and OOD detection capability, such that when the model accuracy performance is increased, it also improves the performance of OOD detection. The extended version also investigated overfitting of models, and saw that the performance of OOD detection performance quickly deteriorated when the model started to overfit, to the point that the model fails to separate in-distribution data to random noise samples. For more details of this behavior, see Fig. 2 in Paper III.

Paper III also introduced a significant change of coverage breakpoint metrics compared to Paper A. Coverage breakpoints refers to the percentage of sample coverage that is achieved with a pre-determined target performance level is assigned. In Paper A, the

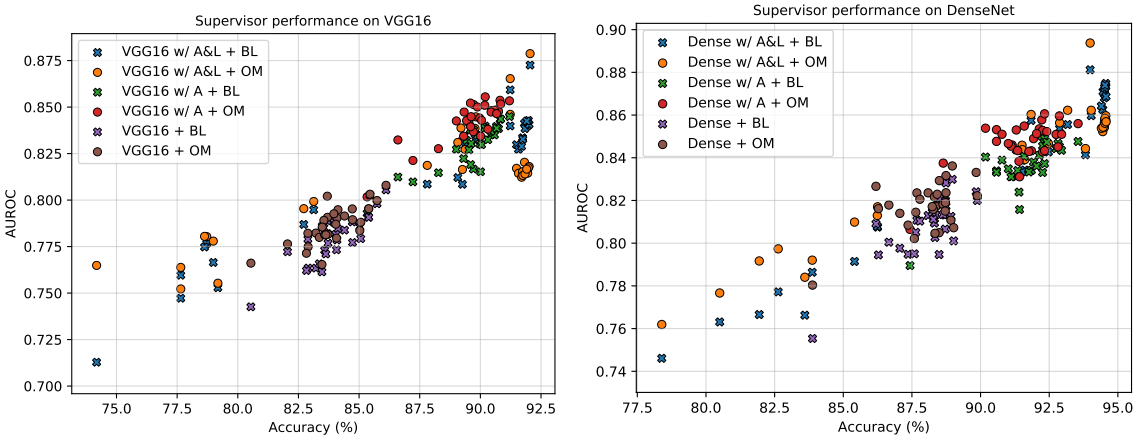

Figure 3.2: The AUROC score visualized with regards to the achieved testing accuracy for two state-of-the-art DNN architectures. The experiments were tested with and without training augmentation (marked with A in the figure legend) and adjusted learning rate (marked L in the figure legend) for two variants of OOD detection, namely Baseline (BL) [41] and OpenMax (OM) [44] .

target performance was to reduce coverage until the testing set accuracy was achieved. A major drawback of this strategy is that better performing models will have to reject more samples to recover the lost accuracy of outlier samples. To combat this, Paper III instead introduced coverage breakpoints at a fixed performance level, therefore requiring more restrictions on worse performing models. See Fig. 4 and 5 in Paper III to compare both versions of coverage breakpoints.

Up until this point, the datasets used in the studies constituted of small scale images. The next step was to consider more realistic use-cases that could infer performance on real-world applications by expanding the methodology to be applied on more complex datasets. Furthermore, the previous studies utilized classification networks, i.e., DNNs that predicted a class for each input image. However, realistic ADS perception will incorporate a combination of semantic segmentation DNNs and object detection DNNs. Regarding the latter, it is realistic to question how to adapt the OOD safety measure to handle varying input images. Both Paper IV and Paper V studied bounding box DNNs with two different approaches, and two different datasets. Paper IV extended the OOD methods to work on the well-known object detector YOLO-v3 [58], and extracting the class probability vector for each accepted bounding box, thus allowing the evaluation criteria from Paper III to still be applicable.

The goal in Paper V was to utilize an OOD safety measure to verify that the safety requirements were fulfilled. The paper did a full review of the AMLAS publication, and constituted that in a simulator environment the full scope of AMLAS could be covered. To this end, a scenario of a moving pedestrian that is crossing a road in front of a moving vehicle was constructed and used as a basis to generate a multitude of test examples where several parameters could be adjusted. Through extensive testing, Paper V showcased an open sourced safety case for an ML-based component

for a pedestrian automatic emergency braking system.

Semantic segmentation is the task of assigning a class prediction to each pixel in an image. In fact, considering each pixel as an individual instance, the output format will then resemble the format of classification challenges. This revelation sparked interest in the final study, as semantic segmentation has been studied broadly through Cityscapes Benchmark Suite [22]. Paper VI utilized Cityscapes pre-trained DNNs and evaluated a class-conditional Mahalanobis distance with the attempt to see if false positives could be reduced over four automotive semantic segmentation models. Results showed that risk-coverage trade-off existed in all datasets, however more importantly, labeling format and class definition played a large part in the success of DNNs.

The studies are summarized in Table. 3.1, and categorized as *R* for reviewing if the goal was to investigate surrounding research of the topic, or *Ex* for experimental if the aim was to showcase usage of a methodology, and *EM* for evaluation metrics if the aim was to discuss or describe metrics.

Table 3.1: Summary of studies conducted in this thesis. Category refers to focus area of the publication where R: Review, EM: Evaluation metrics, and Ex: Experimental

Paper	**Category**			**Objective of the study**
	R	*EM*	*Ex*	
Paper I	✓			Interview study
Paper II		✓		Literature study / Review of SEFAIS Workshop
Paper III		✓	✓	Experimental metric evaluation on public datasets
Paper IV			✓	Experiments on public datasets
Paper V	✓	✓	✓	Full review and application of AMLAS [12]
Paper VI		✓	✓	Full evaluation on several automotive datasets

3.1 Combining Metrics with Risk Reduction

This section describes outlier detection scores from the confusion matrix and puts them into the context of safety assessment. From the confusion matrix, two states are considered for the OOD detector: 1) a positive condition, i.e, that the OOD detector is triggered, and the sample should not be processed, or 2) a negative condition, i.e., the sample should be considered inside the scope of the model. From these two states the four cells in the confusion matrix are computed such that TP refers to a correctly dismissed outlier sample, FP to a wrongfully dismissed inlier sample, FN to a wrongfully accepted outlier sample, and TN as a correctly accepted inlier sample. With this in mind, the true positive rate (TPR) and false positive rate (FPR) of the model can be computed as

$$\text{TPR} = \frac{TP}{TP + FN} \tag{3.1}$$

$$\text{FPR} = \frac{FP}{FP + TN} \tag{3.2}$$

When computing TPR and FPR without any restrictions, 100% of inlier and outlier samples will be processed. Consider the idea to extend the model with a discriminator that either accepts or rejects the output from the model based on an estimation of how likely the sample belongs to the in-distribution of the model. We refer to this concept as a safety measure (the concept has also been referred to as a safety cage and supervisor in Papers II-V) when combining a distance measure $\mathcal{S}$, referred to as anomaly score, with a discriminator that has an acceptable threshold ϵ as

$$P = \mathcal{M}(\mathbf{x}) \tag{3.3}$$

$$d = \mathcal{S}(P; [\mathcal{D}_{train}, \mathbf{x}, \dots]) \tag{3.4}$$

$$D = \begin{cases} reject, & \text{if } d > \epsilon \\ accept, & \text{otherwise} \end{cases} \tag{3.5}$$

where $\mathcal{M}(\mathbf{x})$ is the DNN output for input $\mathbf{x}$. Note that $\mathcal{S}$ is denoted with optional inputs, as there are no limitations on designing outlier measures, e.g, statistical or distance based approaches as used in Paper III and VI, or reconstruction based approaches such as the VAE used in Paper II and V. A distance value close to zero refers to a sample deemed similar to some training sample.

An OOD detector from Paper II is exemplified in Fig. 3.3. By constructing varying thresholds $\epsilon_1, \epsilon_2, \dots, \epsilon_n$ with accepted anomaly distances, the TPR and FPR will vary, thereby creating the ROC-curve. The area under the curve (AUROC) constitutes the average discriminative performance of the OOD measure, where 1 constitutes perfect separation, and 0.5 refers to random guess. The AUROC measure is the most common metric for comparing OOD methods and is applicable for classification and segmentation tasks.

The OOD detector determines solely whether a sample is within the in-distribution and does not consider if the sample would be handled correctly or not. Simply looking at the discriminative performance is not always suitable for real-life applications. As several outlier samples will be correctly handled by the model, and a portion of inliers will be handled incorrectly, a better approach is to incorporate the prediction from the DNN in combination with the discriminator to see if the performance of the system increases by rejecting a sample. We first introduce risk in Paper II, and define it as the complement of performance metric for the task as hand. For classification tasks, the performance metric used is model accuracy on the test set.

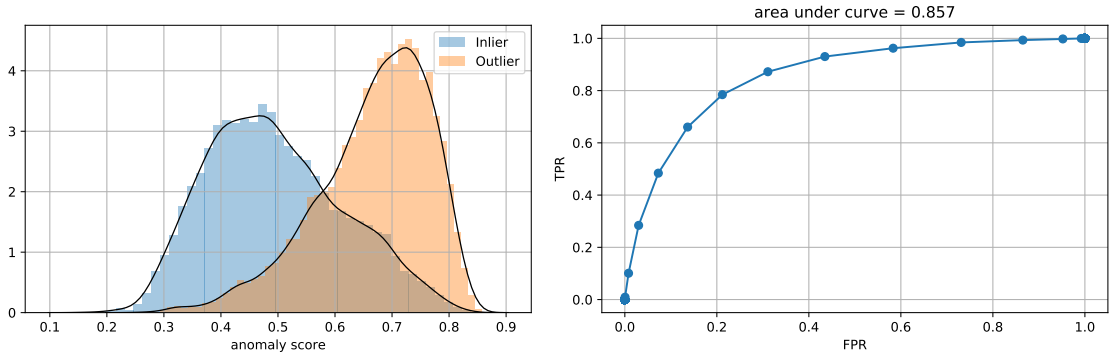

Figure 3.3: In Paper II, outlier detection was performed between CIFAR-10 (inlier set) and Tiny ImageNet (outlier set). The figure shows the histogram of the anomaly scores and the corresponding ROC-curve for one OOD detector.

For segmentation tasks and object detection and classification tasks, the IoU [26] measure is used as performance metric giving the risk as

$$\text{risk} = 1 - IoU = 1 - \frac{P \cap L}{P \cup L} \tag{3.6}$$

where P denotes predictions from the model as described in Eq. 3.3 and L refers to the annotated label of the sample. Extending Eq. 3.6, to also utilize the discriminator defined in Eq. 3.5, the risk can be expressed as a function of the accepted anomaly score ϵ as exemplified in Fig. 3.3. Furthermore, the coverage can be expressed as the percentage of samples that has been processed, i.e., for a semantic segmentation task the coverage refers to the amount of pixels that are accepted to be processed of the DNN. Risk and coverage can then be formulated as

$$\text{risk}(\epsilon) = \frac{P(\epsilon) \cap L(\epsilon)}{P(\epsilon) \cup L(\epsilon)} \tag{3.7}$$

$$\text{coverage}(\epsilon) = \frac{\sum P(\epsilon)}{\sum L(\epsilon)} \tag{3.8}$$

Similar to the ROC-curve, the risk-coverage curve can be visualized, c.f. Fig. 3.4 that visualizes the risk-coverage curve for the performance evaluation on the training set of Paper VI. The yielded function shows a trade-off between risk in the system and functional coverage, something this thesis has referred to as risk-coverage trade-off.

Based on Eq. 3.7 and 3.8, requirements can now be put on model risk and coverage. Based on the safety standards e.g., SOTIF as described in Section 2.2, potential risks of the system can be assessed with a HARA. From the HARA, safety goals can be defined, from which concrete functional safety requirements can be derived. If the

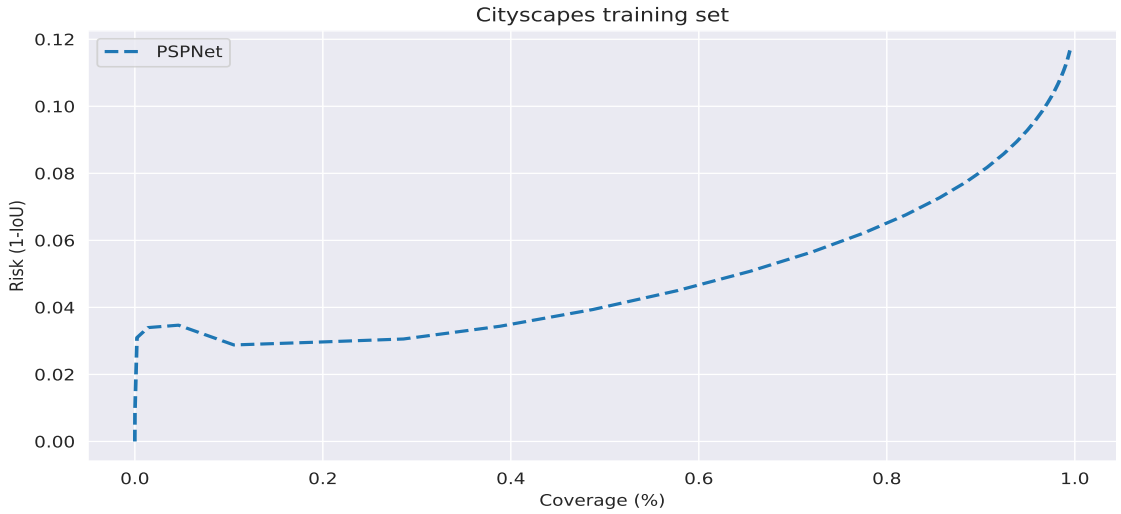

Figure 3.4: A visualization of the risk-coverage trade-off curve for the training set of Paper VI. The curve is achieved by varying the accepted anomaly distance threshold.

requirements are performance based, they can be allocated towards the DNN. To illustrate, performance requirements focus on quantitative targets of the DNN, e.g.,

Req 1: *The component shall have an IoU overlap of 95% for objects that are within 80m*

Req 2: *The component shall predict with 50% coverage for samples that are within the operational design domain.*

Req 3: *In 99% of sequences of 10 consecutive frames from a 10Hz video feed the IoU shall not deteriorate below 80% in more than 20% of frames.*

For the exemplary requirements, Req 1. is `failed`, Req 2. is `passed`, and Req 3. is `not applicable`, as the dataset is not annotated with high enough frame rate. Applying the safety measure from Paper VI as exemplified in Fig. 3.4, the option to restrict samples is possible. Restricting the system with safety measures until risk reaches 0.05, causes the coverage to be reduced to 0.62, thereby rendering Req 1. as `passed` and keeping Req 2. as `passed`. However, consider the case where Req 1. instead stated 98% IoU. For this case, no matter the chosen ϵ threshold, this requirement cannot be passed according to the risk-coverage curve. For this case the options are requirements elicitation, extended training for performance improvements, or operational design domain reduction as discussed mainly in Paper D but also touched on in Paper V.

Chapter 4

Summary of Appended Papers

4.1 Paper I

> **Jens Henriksson**, Markus Borg and Cristofer Englund. "Automotive Safety and Machine Learning: Initial Results from a Study on How to Adapt ISO 26262 Safety Standard", in *1st International Workshop on Software Engineering for AI in Autonomous Systems (SEFAIS)*, pp. 47-49, Gothenburg, Sweden, May 2018. IEEE/ACM

Paper I analyzed ISO 26262 and discusses *Part 6: Product development at the software level*, and which part of that are difficult to be applied for ML components. Through two in-depth interviews with experts from the software development domain, recommended adaptations could be defined and categorized into three concrete areas: 1) the training phase, where requirements need to capture the essence of design and training of the model, and not only the model itself, 2) models need to be evaluated how sensitive they are towards input disturbances, and 3) test case design needs to be tailored more thoroughly for ML.

4.2 Paper II

> **Jens Henriksson**, Christian Berger, Markus Borg, Lars Tornberg, Cristofer Englund, Sankar Raman Sathyamoorthy and Stig Ursing. "Towards Structured Evaluation of Deep Neural Network Supervisors". In: *IEEE International Conference On Artificial Intelligence Testing (AITest)*, pp 27-34, San Francisco, USA, April 2019. IEEE

Paper II aimed at creating a unified framework to structure how OOD detection methods could achieve a fair comparison. This was done by reviewing existing

metrics that had been found in related work and what datasets the evaluation were conducted on. The paper focused on seven metrics that focused on TPR, FPR, FNR, and coverage metrics. Coverage is referring to a percentage describing how much of the original dataset size is used to achieve the stated performance. The paper continued to demonstrate the metrics on two use-cases, 1) a baseline anomaly score test between CIFAR-10 and Tiny ImageNet, to see if separability could be achieved on small scale images with a classification DNN, and 2) a VAE as an OOD detection method that was applied on a generated automotive dataset of highway scenes with and without weather disturbances. The recreation error from the VAE could be used as discriminator and distinguish between regular and foggy driving conditions.

4.3 Paper III

> **Jens Henriksson**, Christian Berger, Markus Borg, Lars Tornberg, Sankar Raman Sathyamoorthy and Cristofer Englund. "Performance Analysis of Out-of-Distribution Detection on Trained Neural Networks". In: *Journal of Information and Software Technology*, Volume 130, 106409, February 2021. Elsevier

The third paper conducted an extensive evaluation of three OOD detection methods as safety measures: A baseline equation [41], ODIN [43], and OpenMax [44], applied to four DNNs. The application of the safety measures were applied at every 10th epoch during training such that the performance could be studied as the models' performance varied over epochs. The paper found that a linear upper bound existed emerged during training, namely that the better the model generalized, the better the safety measures could separate between in and outlier samples. However, the behavior deteriorated as soon as the model started to overfit. The study showed this behavior on three outlier sets.

4.4 Paper IV

> **Jens Henriksson**, Christian Berger and Stig Ursing. "Understanding the Impact of Edge Cases from Occluded Pedestrians for ML Systems". In: *2021 47th Euromicro Conference on Software Engineering and Advanced Applications (SEAA)*, pp 316-325, Palermo, Italy, September 2021. IEEE

The fourth paper expanded the safety measure concept by moving towards large-scale images from KITTI. Specifically, the study investigated how occlusion of pedestrians could affect the performance by allowing the system to reject false positives. In addition, the paper showcased improved performance through subnetworks, that were applied if the predictions from the original pedestrian detector were rejected by the safety measure.

4.5 Paper V

Markus Borg, **Jens Henriksson**, Kasper Socha, Olof Lennartsson, Elias Sonnsjö Lönegren, Thanh Bui, Piotr Tomaszewski, Sankar Raman Sathyamoorthy, Sebastian Brink and Mahshid Helali Moghadam. "Ergo, SMIRK is Safe: A Safety Case for a Machine Learning Component in a Pedestrian Automatic Emergency Brake System". In: *Software Quality Journal*, pp 1-69, March 2023. Springer

Paper V took the recently released AMLAS framework [12], and reviewed it in full. AMLAS provides a high-level guidance of incorporating ML into SCA, but requires case-specific details to be useful. The paper designed and formulated a safety case named SMIRK, a pedestrian automated emergency braking system, and generated extensive test cases for the potential hazard where a pedestrian walks into the vehicle path. The safety case was demonstrated in an automotive-grade simulator and provided a complete safety case for a limited operational design domain. Furthermore, SMIRK introduced safety requirements for the perception system to detect pedestrians such that the safety requirement can be evaluated during training and testing of the DNN model.

4.6 Paper VI

Jens Henriksson, Christian Berger, Stig Ursing and Markus Borg. "Evaluation of Out-of-Distribution Detection Performance on Autonomous Driving Datasets". In: *IEEE International Conference On Artificial Intelligence Testing (AITest)*, Athens, Greece, July 2023. IEEE

The sixth paper transferred the safety measure towards semantic segmentation DNNs by constructing a Mahalanobis distance OOD measure that considered the class-conditional probability distribution accessed from a trained DNN. The distribution were used such that each pixel in the classified image received an anomaly score that could be used by a discriminator and reject false positive samples. The methodology was evaluated on three individually trained DNNs as well as four automotive semantic segmentation datasets. The experiments indicated that safety measures are applicable and come with a risk-coverage trade-off, where the level of accepted risk determines how restrictive the safety measures have to be. Furthermore, the consistency between annotations are more important than where or how training samples have been gathered, since discrepancies between annotations caused the most harm for the performance of the DNNs.

4.7 Paper Contributions to Research Questions

The appended paper of this thesis contribute to the research goals and contribute to bridging the gap between safety assurance and DNN development. To summarize the joint contributions of all appended papers, their relevance towards the research questions are assessed. The contributions are described in subjective perspective through *no contribution*, *minor contribution* and *major contribution*, and presented in Table 4.1.

Table 4.1: Assessment of contributions from the appended papers towards the formulated research questions of this thesis.

Paper	RQ1	RQ2	RQ3
I	**Minor contribution** Introduced the missing link of transferrable metrics to safety assurance	No contribution	**Minor contribution** Shared information of lack of safety measures for DNNs
II	**Major contribution ✓** Introduced and recommended metrics that are found in related work and suitable for outlier detection evaluation	**Minor contribution** Introduced the coverage and risk breakpoints	**Minor contribution** Initial rejection experiments in an automotive simulator
III	**Major contribution ✓** Extensive tests of metrics and their usage during training of DNNs	**Minor contribution** Formulated the ground work for safety measures	No contribution
IV	**Minor contribution** Highlighted the existing issues	**Minor contribution** Applied the safety measure on automotive samples and showcased the benefit	No contribution
V	**Minor contribution** Applied metrics on a realistic automotive scenario	**Major contribution ✓** Followed a safety case that incorporated safety measures in the argumentation	**Major contribution ✓** Summarized how to use safety measures as during runtime or during training
VI	**Minor contribution** Highlighted the existing issues	**Major contribution ✓** Formulated a VAE as a safety measure and utilized it to show compliance with allocated ML safety requirements	**Major contribution ✓** Summarized how the safety measure can be applied during verification of the system

Chapter 5

Discussion

The goals of this thesis were to understand how to translate metrics between ML and safety, and shed some light on the effect that safety measures can have on data-driven algorithms. In this thesis, OOD detection methods have been formulated as one safety measure that has been studied on datasets with a range of complexity. This section summarizes answers to each research questions and highlights potential drawbacks of the conducted studies.

RQ 1: What would constitute suitable metrics that can be extracted from DNNs and utilized as safety measures in the automotive verification?

In this thesis, several DNN metrics are presented in Paper II, III, V, and VI. In short, these metrics can be extracted from the confusion matrix and the evaluation set that is used during the training. From the automotive verification perspective, the thesis has shown how requirements can be expressed as a product of minimum true positive rate, maximum allowed false positives, and maximum allowed false negatives. Furthermore, these are combined to express a risk-coverage trade-off, a metric that puts the DNN performance into context of performance requirements on the model, to highlight the necessary encapsulation needed to achieve requirements.

RQ 2: How can OOD detection be formulated as one potential safety measure?

The OOD detector can be formulated as a discriminator, where predictions above an accepted threshold level are considered outside the scope for the function at hand. This can be utilized as a pass/fail requirement, to ensure all samples are within the accepted distance, something that was demonstrated in Paper V. The discriminator can also be seen as a spectrum during evaluation, i.e., by varying the accepted distance threshold, the function coverage varies accordingly. This trade-off allows safety requirements to be allocated on the model by defining an accepted level minimum coverage, which was suggested in Paper II and III, and further demonstrated in Paper VI.

RQ 3: At what stages can OOD detection be beneficial to the safety lifecycle of autonomous driving features?

This thesis has primarily considered OOD detection as a testing tool of DNNs, that can be applied during training. To this end, Paper II, Paper V and Paper VI demonstrates the usage of OOD detection methods on datasets with wide variety of complexity. In the context of those studies, requirements' can be allocated towards the perception DNN and verified with a set of safety measures, where we showcased that OOD detection can act as one measure. Furthermore, it is also noted that the technique can be applied during on-road testing, e.g., as described in Paper D, through shadow mode, or as an assurance that the model performs within boundaries during runtime. For the shadow mode, the OOD safety measure can identify samples that are far off from the training domain, such that they constitute new test cases that are not covered in the training.

As noted through all studies conducted in this thesis, the datasets are the key ingredients to succeed with ML. Throughout this journey, the first OOD methods were applied to small scale images where a concrete distributional shift could be measured. This allowed the methods to perform extraordinary well, however moving towards real-world applications with more complex images, the shift distributional shift disappeared and estimated anomaly scores became increasingly similar between in- and outlier samples [16]. Separability will be increased with 1) better trained models, and 2) better safety measures, however based on the results from this thesis, there will always be a trade-off between coverage and rejection of outlier cases.

Throughout this thesis, experiments have suggested training on one dataset and evaluating on another. However, it has rarely been the case that both sets contain the same labeling principles, or the same class definition. This yields a case of comparing *apples* to *oranges*, i.e., an unfavorable comparison between the sets, since the model is trained on one but not the other. Paper VI showcases this issue, as the performance is severely degraded on datasets with labeling differences. This issue will be present in the automotive industry if different labeling-service providers are used unless they have a proper evaluation structure of dataset labeling.

It has to be noted that the OOD safety measure evaluation is not exhaustive. Studies show the effectiveness of the method, however, to truly demonstrate usefulness the method needs to be applied on a larger scale, suggestively on a larger scale of dataset. Furthermore, the safety requirements specified in the studies are for research purpose only. Each individual SCA will have independent assessment of what constitute good safety requirements for their specific SCA. As long as the assessment formulates requirements based on the confusion matrix, then OOD detection is applicable as one safety measure.

Chapter 6

Conclusions and Future Work

This thesis has investigated OOD detection as a safety measure that is applicable to SCA. The goal was to bridge the gap between ML developers and safety engineers, by describing what joint metrics can be utilized in both ML and as safety requirements allocated on ML algorithms. During the course of this thesis, several metrics were identified and described. Most prominent was the usage of ROC-metrics, that are extracted from the confusion matrix. Furthermore, this thesis combined ROC metrics with coverage metrics accessed through evaluation on the training set, such that requirements can be formulated based on DNN performance and require that the DNN maintain a specific coverage to be reliable. We summarize this as safety requirements on the *risk-coverage trade-off* metric and believe that this metric will be useful as evaluation of safety measures in general. The metric incorporates the exclusion of uncertain samples and investigates how the performance of the model varies when allowing this exclusion. In most scenarios this has a risk reducing benefit, see e.g., Fig. 3.4 where misclassifications are reduced by almost half (0.12 to 0.06) by reducing coverage roughly 25%. In certain experiments the risk increased however, something Paper III and Paper VI concluded occurred due to unfamiliar scenarios and overfitting.

All OOD methods that have been evaluated in this thesis have shown reduced false positive rates on experiments on small-scale datasets containing handwritten digits and tiny images. For the later parts of this thesis the same methods have been applied to realistic use-cases with object detection and semantic segmentation DNNs that are useful for autonomous driving. One key insight has been that these methods are not always transferrable to the realistic use-cases and often require significant tuning. Furthermore, if the OOD method requires parameter tuning, this step has to be renewed for each new version of the DNNs.

It is expected that this thesis sparks interest in more demonstrations of the positive effect of safety measures, such that the different techniques motivated by SOTIF [4], Mohseni et al. [13], and similar are demonstrated on automotive datasets before being

recommended. This thesis has followed OOD detection, one of the recommended techniques, through different DNN tasks with various difficulty to demonstrate the effectiveness and trade-offs that occur. This form of demonstrate is important to legitimize the recommendation of safety measures, and should be seen as a first step in a broader demonstration of safety measures and their benefit for safety critical deep learning applications.

The experiments conducted in this thesis are still too few to provide any generalized consensus. The concept of safety measures has to be evaluated thoroughly, directly at the automotive manufacturer. It is suggested that for future work, the evaluation to be conducted on a large vehicle fleet, such that the concept can guide suitable operational design domain limitations or highlight functional insufficiencies. The thesis has utilized several prominent open source datasets that are suitable for performance evaluation tasks. However, there is a lack in the research community of high detailed annotations of time-series autonomous datasets. Until recently, the time-series aspect has been neglected in public datasets due to lack of availability, but with the release of A2D2 [36] in 2020 and KITTI-360 [21] in 2022, there is an opportunity to continue the safety measure evaluation with time-series in mind. With time-series advantages, object detection can incorporate tracking over time to improve robustness and quality of the predictions and semantic segmentation tasks can feed back the previous prediction to achieve a better future estimate. Combining both with LiDAR or radar scans, all sensors can support and provide an occupancy grid, to ensure safe drivable paths for the ADS.

Bibliography

[1] Olga Russakovsky et al. “ImageNet Large Scale Visual Recognition Challenge”. In: *International Journal of Computer Vision* (2015), pp. 211–252 (cit. on pp. 1, 4).

[2] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. “Deep learning”. In: *nature* 521.7553 (2015), pp. 436–444 (cit. on p. 1).

[3] Rick Salay, Rodrigo Queiroz, and Krzysztof Czarnecki. “An Analysis of ISO 26262: Using Machine Learning Safely in Automotive Software”. In: *Arxiv preprint [1709.02435.]* (2017) (cit. on pp. 1, 3).

[4] International Organization for Standardization. *ISO 21448:2022 Road Vehicles - Safety of the intended functionality*. ISO Standard No. ISO 21448:2022. 2022 (cit. on pp. 1, 4, 7, 11, 31).

[5] International Organization for Standardization. *ISO/AWI PAS 8800 Road Vehicles — Safety and artificial intelligence*. ISO Standard No. ISO/AWI PAS 8800:2023. 2023 (cit. on p. 1).

[6] Aaron Benson et al. *Potential Reduction in Crashes, Injuries and Deaths from Large-Scale Deployment of Advanced Driver Assistance Systems*. Research Brief. Washington, D.C.: AAA Foundation for Traffic Safety, 2018 (cit. on p. 2).

[7] European Road Safety Observatory. *Advanced driver assistance systems*. European Commission, 2018 (cit. on p. 2).

[8] SAE International Surface Vehicle Recommended Practice. *Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles*. SAE Standard J3016. 2021 (cit. on p. 2).

[9] De Jong Yeong et al. “Sensor and sensor fusion technology in autonomous vehicles: A review”. In: *Sensors* 21.6 (2021), p. 2140 (cit. on p. 2).

[10] International Organization for Standardization. *ISO 26262-1:2018 Road vehicles — Functional safety*. ISO Standard No. 26262:2018. 2018 (cit. on p. 3).

[11] Yarin Gal and Zoubin Ghahramani. “Dropout as a bayesian approximation: Representing model uncertainty in deep learning”. In: *international conference on machine learning*. PMLR. 2016, pp. 1050–1059 (cit. on p. 3).

[12] Richard Hawkins et al. “Guidance on the Assurance of Machine Learning in Autonomous Systems (AMLAS)”. In: *Arxiv preprint [2102.01564]* (2021) (cit. on pp. 4, 7, 11, 20, 27).

[13] Sina Mohseni et al. “Taxonomy of Machine Learning Safety: A Survey and Primer”. In: *ACM Computing Surveys* 55.8 (2023), pp. 1–38 (cit. on pp. 4, 12, 31).

[14] Chiyuan Zhang et al. “Understanding deep learning (still) requires rethinking generalization”. In: *Communications of the ACM* 64.3 (2021), pp. 107–115 (cit. on pp. 4, 11).

[15] Dan Hendrycks and Thomas Dietterich. “Benchmarking Neural Network Robustness to Common Corruptions and Perturbations”. In: *International Conference on Learning Representations*. 2019 (cit. on p. 4).

[16] Mattias Landgren and Ludwig Tranheden. *Input Verification for Deep Neural Networks*. Master Thesis 255752, Chalmers University of Technology, Gothenburg, 2018 (cit. on pp. 4, 18, 30).

[17] Jingkang Yang et al. “OpenOOD: Benchmarking Generalized Out-of-Distribution Detection”. In: *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track* (2022) (cit. on pp. 4, 13).

[18] Lydia Gauerhof et al. “Assuring the Safety of Machine Learning for Pedestrian Detection at Crossings”. In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. Vol. 12234 LNCS. 2020, pp. 197–212 (cit. on p. 6).

[19] Yann LeCun et al. “Gradient-based learning applied to document recognition”. In: *Proceedings of the IEEE* 86.11 (1998), pp. 2278–2324 (cit. on p. 6).

[20] Alex Krizhevsky. “Learning Multiple Layers of Features from Tiny Images”. In: (2009) (cit. on p. 6).

[21] Yiyi Liao, Jun Xie, and Andreas Geiger. “KITTI-360: A Novel Dataset and Benchmarks for Urban Scene Understanding in 2D and 3D”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2022) (cit. on pp. 6, 32).

[22] Marius Cordts et al. “The Cityscapes Dataset for Semantic Urban Scene Understanding”. In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. Vol. 2016-Decem. 2016, pp. 3213–3223 (cit. on pp. 6, 14, 15, 20).

[23] Kai Ming Ting. “Confusion Matrix”. In: *Encyclopedia of Machine Learning and Data Mining*. Ed. by Claude Sammut and Geoffrey I. Webb. Boston, MA, 2017, pp. 260–260 (cit. on p. 9).

[24] T E Boult et al. “Learning and the Unknown: Surveying Steps toward Open World Recognition”. In: *Proceedings of the AAAI Conference on Artificial Intelligence* 33.01 (2019), pp. 9801–9807 (cit. on p. 9).

[25] Tom Fawcett. “An introduction to ROC analysis”. In: *Pattern Recognition Letters* 27.8 (2006), pp. 861–874 (cit. on p. 10).

[26] Hamid Rezatofighi et al. “Generalized intersection over union: A metric and a loss for bounding box regression”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Vol. 2019-June. 2019, pp. 658–666 (cit. on pp. 10, 22).

[27] Yonatan Geifman and Ran El-Yaniv. “Selective classification for deep neural networks”. In: *Advances in Neural Information Processing Systems*. 2017 (cit. on p. 10).

[28] Yonatan Geifman and Ran El-Yaniv. “SelectiveNet: A deep neural network with an integrated reject option”. In: *36th International Conference on Machine Learning, ICML 2019*. Vol. 36. 2019, pp. 3768–3776 (cit. on p. 10).

[29] Dario Amodei et al. “Concrete Problems in AI Safety”. In: *Arxiv preprint [1606.06565]* (2016) (cit. on p. 11).

[30] Richard Hawkins et al. “A new approach to creating clear safety arguments”. In: *Advances in Systems Safety - Proceedings of the 19th Safety-Critical Systems Symposium, SSS 2011*. 2011, pp. 3–23 (cit. on p. 12).

[31] Simon Burton et al. “Safety Assurance of Machine Learning for Perception Functions”. In: *Deep Neural Networks and Data for Automated Driving: Robustness, Uncertainty Quantification, and Insights Towards Safety*. Ed. by Tim Fingscheidt, Hanno Gottschalk, and Sebastian Houben. Cham, 2022, pp. 335–358 (cit. on p. 12).

[32] Rüdiger Ehlers. “Formal verification of piece-wise linear feed-forward neural networks”. In: *International Symposium on Automated Technology for Verification and Analysis*. Vol. 10482 LNCS. 2017, pp. 269–286 (cit. on p. 12).

[33] Xiaowei Huang et al. “Safety verification of deep neural networks”. In: *International Conference on Computer Aided Verification*. Vol. 10426 LNCS. 2017, pp. 3–29 (cit. on p. 12).

[34] Sanjit A Seshia, Dorsa Sadigh, and S Shankar Sastry. “Towards Verified Artificial Intelligence”. 2020 (cit. on p. 12).

[35] Florian Tambon et al. “How to certify machine learning based safety-critical systems? A systematic literature review”. In: *Automated Software Engineering* 29.2 (2022) (cit. on p. 12).

[36] Jakob Geyer et al. “A2D2: Audi Autonomous Driving Dataset”. In: *Arxiv preprint [2004.06320]* (2020) (cit. on pp. 13, 15, 32).

[37] Varun Chandola, Arindam Banerjee, and Vipin Kumar. “Anomaly Detection : A Survey”. In: *ACM Computing Surveys* (2009) (cit. on p. 13).

[38] Chuanxing Geng, Sheng Jun Huang, and Songcan Chen. *Recent Advances in Open Set Recognition: A Survey*. 2021 (cit. on p. 13).

[39] Lukas Ruff et al. “A Unifying Review of Deep and Shallow Anomaly Detection”. In: *Proceedings of the IEEE* 109.5 (2021), pp. 756–795 (cit. on p. 13).

[40] Jinwon An and Sungzoon Cho. "Variational Autoencoder based Anomaly Detection using Reconstruction Probability". In: *Special lecture on IE 2.1* (2015), pp. 1–18 (cit. on p. 13).

[41] Dan Hendrycks and Kevin Gimpel. "A baseline for detecting misclassified and out-of-distribution examples in neural networks". In: *5th International Conference on Learning Representations*. 2017 (cit. on pp. 14, 17, 19, 26).

[42] Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich. "Deep anomaly detection with outlier exposure". In: *7th International Conference on Learning Representations, ICLR 2019*. 2019 (cit. on p. 14).

[43] Shiyu Liang, Yixuan Li, and R Srikant. "Enhancing the reliability of out-of-distribution image detection in neural networks". In: *6th International Conference on Learning Representations*. 2018 (cit. on pp. 14, 26).

[44] Abhijit Bendale and Terrance E Boult. "Towards open set deep networks". In: *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. 2016 (cit. on pp. 14, 19, 26).

[45] Kimin Lee et al. "A simple unified framework for detecting out-of-distribution samples and adversarial attacks". In: *Advances in Neural Information Processing Systems*. 2018, pp. 7167–7177 (cit. on p. 14).

[46] Daniel Bogdoll, Maximilian Nitsche, and J. Marius Zollner. "Anomaly Detection in Autonomous Driving: A Survey". In: *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*. Vol. 2022-June. 2022, pp. 4487–4498 (cit. on pp. 14, 15).

[47] Fisher Yu et al. "BDD100K: A Diverse Driving Dataset for Heterogeneous Multitask Learning". In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. 2020, pp. 2633–2642 (cit. on p. 14).

[48] Philipp Oberdiek, Matthias Rottmann, and Gernot A Fink. "Detection and retrieval of out-of-distribution objects in semantic segmentation". In: *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*. Vol. 2020-June. 2020, pp. 1331–1340 (cit. on p. 15).

[49] Alex Kendall, Vijay Badrinarayanan, and Roberto Cipolla. "Bayesian segnet: Model uncertainty in deep convolutional encoder-decoder architectures for scene understanding". In: *British Machine Vision Conference 2017, BMVC 2017*. 2017 (cit. on p. 15).

[50] Hermann Blum et al. "Fishyscapes: A benchmark for safe semantic segmentation in autonomous driving". In: *Proceedings - 2019 International Conference on Computer Vision Workshop, ICCVW 2019*. 2019, pp. 2403–2412 (cit. on p. 15).

[51] Peter Pinggera et al. "Lost and found: Detecting small road hazards for self-driving vehicles". In: *IEEE International Conference on Intelligent Robots and Systems* 2016-November (Nov. 2016), pp. 1099–1106 (cit. on p. 15).

[52] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. “"Why should I trust you?" Explaining the predictions of any classifier”. In: *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2016, pp. 1135–1144 (cit. on p. 15).

[53] Michael Wagner and Philip Koopman. “A Philosophy for Developing Trust in Self-driving Cars”. In: *Lecture Notes in Mobility*. 2015, pp. 163–171 (cit. on p. 15).

[54] Nicholas Carlini and David Wagner. “Towards Evaluating the Robustness of Neural Networks”. In: *Proceedings - IEEE Symposium on Security and Privacy* (2017), pp. 39–57 (cit. on p. 15).

[55] Kexin Pei et al. “DeepXplore: Automated Whitebox Testing of Deep Learning Systems”. In: *SOSP 2017 - Proceedings of the 26th ACM Symposium on Operating Systems Principles*. 2017, pp. 1–18 (cit. on p. 15).

[56] Lei Ma et al. “DeepGauge: Multi-granularity testing criteria for deep learning systems”. In: *ASE 2018 - Proceedings of the 33rd ACM/IEEE International Conference on Automated Software Engineering*. 2018, pp. 120–131 (cit. on p. 15).

[57] Yuchi Tian et al. “DeepTest: Automated testing of deep-neural-network-driven autonomous cars”. In: *Proceedings - International Conference on Software Engineering*. Vol. 2018-May. 2018, p. 12 (cit. on p. 15).

[58] Joseph Redmon and Ali Farhadi. “YOLOv3: An incremental improvement”. In: *Arxiv preprint [1804.02767]*. 2018 (cit. on p. 19).