ACIT5900

MASTER THESIS

in

Applied Computer and Information Technology (ACIT)

May 2023

Applied Artificial Intelligence

A Robust and Secure Edge-Based AI System Against Adversarial Attacks

Stian André Sørensen

Department of Computer Science

Faculty of Technology, Art and Design

OSLOMET

Abstract

Ensuring the safe deployment and use of artificial intelligence (AI) in safety-critical systems is crucial in the reality of effective adversarial attacks (AAs). AAs involve manipulating the data inputs to AI models to make them behave abnormally and make mistakes. Such attacks may lead AI systems to perform destructive behaviors, leading to unintended outcomes. Therefore, the threat of AAs must be considered in the design process of such systems.

The main objectives of this thesis were to understand the status quo of AAs and defenses in image and video object detection (OD) that needs attention in Kongsberg Defense & Aerospace's (KDA's) context (Goal-01), and use that information to derive a system architecture and requirements for the safe deployment of AI in an unmanned military setting (Goal-02).

A systematic literature review of AAs and defenses in OD was done to uncover the state of the art in the field. The results showed that most existing research focused on AAs in the digital domain with white box knowledge. However, for AAs to become a real threat to unmanned military systems, the research field must focus more on creating physically realizable black box AAs - a challenging task yet to be properly solved and a somewhat premature research field with debatable real-world threats.

A system for the safe deployment of AI in an unmanned military setting was designed based on the systematic literature review results and a requirements engineering process with KDA. Supporting manned vehicles and a back-end server were included to cover the entire AI lifecycle and cope with the limitations of military systems. The use of GPUs for accelerated AI was an essential enabling technology. The threat of most AAs against the system was considered negligible due to the strict security requirements of military systems. Adversarial defenses, like adversarial-training and detection, were recommended to further reduce the threat of AAs, especially the most prominent threat of physically realizable black box AAs.

This thesis delivered a comprehensive review of the state-of-the-art AAs and defenses in OD in the context of unmanned military vehicles – the first of its kind and a valuable resource for the research field and the defense industry. KDA gained valuable information on how AI can safely be deployed in unmanned military systems and how to maintain a cycle of ever-improving AI models throughout the lifetime of the system.

Preface

This master's thesis, titled "A Robust and Secure Edge-Based AI System Against Adversarial Attacks", finalizes my master's education in Applied Computer and Information Technology with a specialization in Applied Artificial Intelligence at OsloMet. My master's thesis was a collaboration with Kongsberg Defence & Aerospace (KDA) - a valuable partner from early project planning until delivery. The planning of the thesis project started toward the end of 2022, with the research and writing of the thesis happening from January to May 2023.

I have learned a lot about myself and my field of research during my work on this thesis. The methods I utilized were entirely new to me, including conducting and writing a systematic literature review and following a requirements engineering process. Before researching and writing my thesis on adversarial attacks, I had little prior knowledge of the topic. Therefore, this thesis has taught me valuable skills and knowledge that will be useful in my future career and studies.

I would like to thank my supervisor Prof. Lothar Fritsch at OsloMet for valuable guidance, feedback, and support throughout the entire thesis. His engagement and expertise in cyber-security were priceless for the research in this thesis. I would also like to thank KDA for enabling this thesis project with their knowledge and experience in the defense industry. A special thanks to Vegar Vejlgaard Morsund - my contact person and supervisor from KDA - for his guidance, feedback, and support throughout the entire thesis. Lastly, I want to thank my friends and family for their incredible support and patience during my studies - especially during the work on this thesis.

Stian André Sørensen
Oslo, May 14, 2023

Non-Disclosure Agreement

A non-disclosure agreement was signed with Kongsberg Defence & Aerospace[1] (KDA) in advance of writing this thesis. Thereby, the thesis does not include classified information and vulnerabilities related to KDA's products and systems. This agreement was kept by focusing on a totally theoretical system.

[1] Kongsberg Defence & Aerospace (KDA): https://www.kongsberg.com/kda/

Contents

Abstract i

Preface iii

Non-Disclosure Agreement v

1 Introduction 1
- 1.1 Project Description . . . 2
  - 1.1.1 Systematic Literature Review of Adversarial Attacks and Defenses in Object Detection . . . 2
  - 1.1.2 Edge-Based AI System for Unmanned Military Vehicles . . . 4
- 1.2 Goals . . . 5
- 1.3 Contributions . . . 5
- 1.4 Thesis Report Structure . . . 5

2 Background 7
- 2.1 Deep Learning-Based Object Detection . . . 7
- 2.2 Adversarial Attacks and Defenses in Deep Learning . . . 8
  - 2.2.1 Terminology . . . 8
  - 2.2.2 Adversarial Attacks . . . 10
  - 2.2.3 Adversarial Defenses . . . 12
- 2.3 Related Work . . . 12
  - 2.3.1 Surveys and Reviews on Adversarial Attacks and Defenses in Computer Vision . . . 13
  - 2.3.2 Adversarial Attacks in Computer Vision in the Defense Industry . 14

3 Methods 15
- 3.1 Discovering the State of the Art in Adversarial Attacks and Defenses in Object Detection . . . 15
  - 3.1.1 Justifying the Systematic Literature Review . . . 16
  - 3.1.2 Database Search Strategy . . . 16
  - 3.1.3 Snowballing . . . 17

3.1.4 Data Extraction 18
3.2 Securing the Edge-Based AI System for Unmanned Vehicles 21
3.2.1 Understanding the Stakeholder Needs and Requirements 21
3.2.2 Derviving the System Architecture and Requirements 22
3.2.3 Deriving the Security Requirements 22

4 Results **23**
4.1 State-of-the-Art Adversarial Attacks and Defenses in Object Detection . 23
4.1.1 Results From the Literature Search 24
4.1.2 Categorization and Analysis of the Reviewed Literature 29
4.2 Ensuring the Safe Deployment of AI Against Adversarial Attacks in Unmanned Systems 38
4.2.1 Threat Model 38
4.2.2 Stakeholder Needs 39
4.2.3 Stakeholder Requirements 40
4.2.4 System Architecture – More than Just Unmanned Vehicles . . . 41
4.2.5 Logical System Architecture 43
4.2.6 Physical System Architecture 44
4.2.7 AI Pipeline / Functional System Architecture 46
4.2.8 Acknowledging and Mitigating the Threats of Adversarial Attacks 48
4.2.9 System and Security Requirements 54

5 Discussion **59**
5.1 The Research Field of Adversarial Attacks and Defenses in Object Detection 59
5.1.1 Discussion of the Systematic Literature Review 60
5.1.2 Physically Realizable Adversarial Attacks 61
5.1.3 Conflicting Terminology 61
5.2 Discussion of the Derived System Architecture and Requirements . . . 62
5.2.1 How the Stakeholder Needs and Requirements were Met 63
5.2.2 Combining Knowledge from Different Specializations 65
5.2.3 Choice of Object Detectors 65
5.2.4 The Current Threat of Adversarial Attacks 66
5.3 Ethical Considerations 67
5.4 Future Research 68

6 Conclusion **71**

A Tools to Aid Future Research on Adversarial Attacks and Defenses **101**

B Other Data Types and Multi-Sensor Object Detection **103**

List of Figures

1.1 Project outline 2
1.2 Example use case of the edge-based AI system 4

2.1 Visualization of adversarial examples 11

4.1 Adversarial attack methods 30
4.2 Adversarial attack types 31
4.3 Attack methods' domain distribution 33
4.4 The covered adversarial attack methods' domain distribution 34
4.5 Covered object detector families 35
4.6 Adversarial defense methods 36
4.7 The defense methods' targeted attack methods 38
4.8 Logical system architecture 43
4.9 Physical system architecture 45
4.10 AI pipeline / functional system architecture 46
4.11 Threat analysis of digital adversarial attacks 49
4.12 Threat analysis of physical adversarial attacks 53
4.13 System requirements implementation in the AI pipeline 58

List of Tables

3.1 Inclusion and exclusion criteria for data extraction 19

4.1 Adversarial attacks in object detection 27
4.2 Adversarial defenses in object detection 28

A.1 Tools for adversarial attacks and defenses in object detection 101

Chapter 1

Introduction

Artificial intelligence (AI) applications have become more widespread in recent years and are becoming part of people's everyday life. People interact with AI daily through the use of personal assistants on smart devices, chatbots, recommender systems on media platforms, email spam filtering, personalized marketing, and even in safety-critical systems like self-driving cars. AI has also become a hot topic in everyday conversation due to the recent releases of groundbreaking models like DALL·E 2 and ChatGPT, resulting in both spiked public interest and worry about the field.

The research community has also stated its worry as deep learning algorithms have in recent years shown to be prone to adversarial attacks [1, 2]. Adversarial attacks are a type of attack against the functionalities of deep learning algorithms that leads to models making wrong decisions, leading to undesired behavior. Such attacks can be applied to various input data, including images, video, audio, text, and 3D data like point clouds derived from LiDAR scans. Adversarial attacks apply small changes to the input data, called adversarial perturbations, to create adversarial examples that cause mistakes in the target model's predictions. These adversarial perturbations are not always perceptible to humans [3] and can, in such cases, even bypass human monitoring.

Plenty of possible real-world uses of adversarial attacks exist, including tricking face detectors into misclassifying persons as somebody else, misclassifying medical images, bypassing malware detection and spam filtering, and making vehicles and pedestrians invisible to self-driving cars. In the worst case, such attacks can have fatal consequences, especially when applied to safety-critical systems like autonomous vehicles. Therefore it is crucial to develop robust defenses against such attacks to retain both trust and reliability of AI-driven systems. Adversarial attacks and defenses have consequently become a growing area of research within the field of AI.

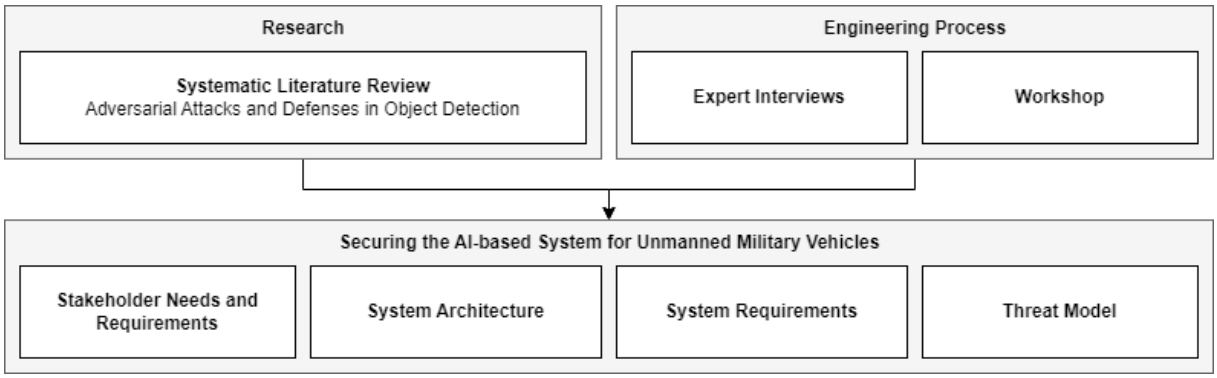

Figure 1.1: Project outline.

1.1 Project Description

Ensuring the safe deployment and use of AI in safety-critical systems is crucial in the reality of effective adversarial attacks. It is vital that unmanned systems can trust the decisions made by its enabling AI to prevent the system from making mistakes leading to undesired outcomes - especially in military settings. Previously an operator could quickly uncover attacks or decoys by looking through their binoculars. However, operators no longer have the capacity to live-monitor the input of the many sensors installed in one or several unmanned or partially-automated systems. Therefore, the development of a robust and secure AI system is highly relevant in an increasingly unmanned reality.

This master's thesis is a collaboration with Kongsberg Defence & Aerospace (KDA) and is, for the most part, theoretical in nature. KDA has much experience running unmanned system demos and was a valuable resource for this thesis. The thesis is divided into two main parts: a systematic literature review on adversarial attacks and defenses against object detectors, and recommendations for a robust and secure edge-based AI system against adversarial attacks in the military setting. Figure 1.1 is a visualization of the project outline, including the methodologies and results of the thesis.

1.1.1 Systematic Literature Review of Adversarial Attacks and Defenses in Object Detection

For the first part of the thesis, a systematic literature review on adversarial attacks and defenses in object detection for images and videos was conducted. The result of the literature review was the presentation of the state-of-the-art adversarial attacks and defenses relevantly categorized to the use case of unmanned military systems.

The use of AI in unmanned military vehicles can be compared to that of self-driving

cars, with the added element of a significantly increased probability of being a target for attacks. Due to the many possible uses of AI in autonomous systems, including environmental perception, decision-making, route planning, predictive maintenance, and safety enhancements, a systematic literature review on the entire topic would have been far too time-consuming.

Therefore, the research question was focused on a single AI component of autonomous systems, the deep learning-based object detection task within computer vision. The security, trust, and viability of object detection algorithms are crucial for deployment in the described use case, leading to the research question of this systematic literature review:

> *RQ: Does the current state of adversarial attacks pose a threat to image and video object detectors in unmanned military systems? - And how can they be defended?*

Background

The following background information is provided to further elaborate on the main concepts and foundations of the research question in KDA's context, including the object detection task, the selected data types, and adversarial attacks.

Object detection is a popular field within computer vision, which aims to localize and classify objects in data like images, video, and point clouds. KDA considered object detection one of the most valuable and essential tasks in unmanned systems. Also, the current state of technology offers solutions to the problem, making it possible to implement or experiment with object detection.

Due to several factors, images and videos were the media in focus in this thesis. One was that most research on adversarial attacks and defenses in the object detection task had focused on using images as the target medium. KDA also considered images the most relevant medium in a military system, as unmanned systems are likely to make decisions based on images in the future, making images a target for attacks. Object detection on videos was also included, as videos are simply sequences of images.

Moreover, this systematic literature review was necessary due to object detectors' vulnerability to adversarial attacks - modifications to images or videos that cause object detectors to behave abnormally and make mistakes. Such attacks may lead AI systems to perform destructive behaviors, leading to unintended outcomes. Therefore, the threat of adversarial attacks must be considered in the design process of systems using object detection.

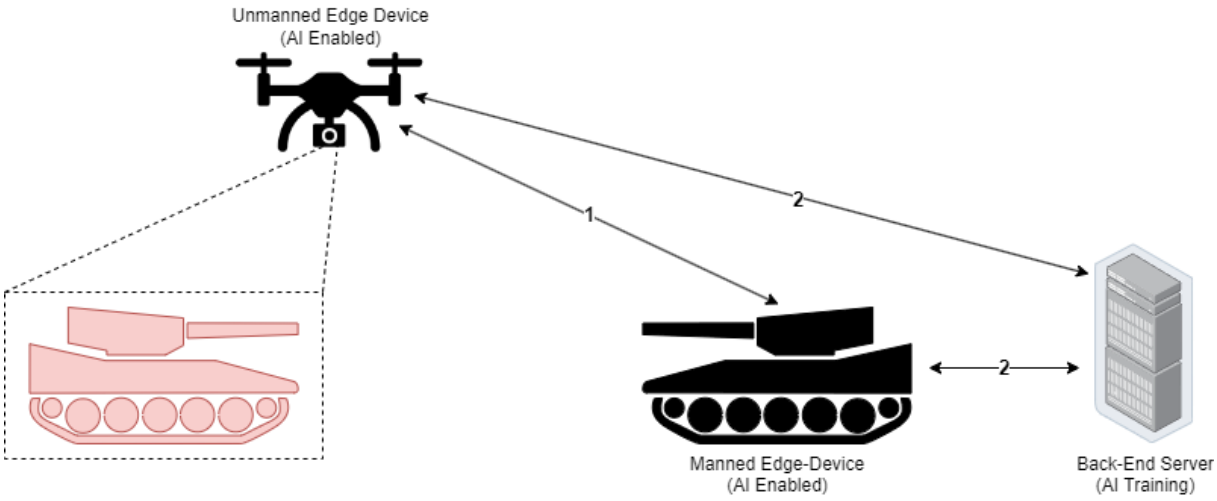

Figure 1.2: Example use case of the system. The numbered connection transmits the following data: **1:** Shared information (AI predictions and sensor data), **2:** Trained AI models (server $\rightarrow$ edge), recorded mission data (training data) (edge $\rightarrow$ server).

1.1.2 Edge-Based AI System for Unmanned Military Vehicles

The second part of the thesis project consisted of designing a system architecture and a set of system requirements for the safe deployment of AI in unmanned military systems. The system architecture describes not only the unmanned vehicles themselves but also the surrounding enabling systems, including manned vehicles and a back-end server. The system is entirely theoretical and was mainly based on the results from the systematic literature review on adversarial attacks and defenses in object detection. To supplement the research, workshops with KDA and expert interviews were conducted to get a better understanding of the specific use case. The restrictions that followed the military use case, like limitations in communication technology and computing power, and the strict requirements for confidentiality, integrity, and availability, needed to be considered in the design process.

Figure 1.2 shows a visualization of an example use case of such a system. Both the unmanned air vehicle (UAV) and the main vehicle (tank) are installed with their own AI to assist in operations. In this theoretical scenario the UAV scouts ahead of the manned tank to look for potential threats using its AI-based environmental perception, including object detection. When detecting the potential threat, in this case, an opposing tank, it reports back to the main vehicle. The systems can exchange data, e.g. video feeds or predictions, due to the secure connections between them. Based on the intel and data from the UAV, the manned tank can evaluate how to evade or engage the detected potential threat. At the end of the operation, the main vehicle uploads all the relevant newly recorded data to the back-end server for training new or improved machine learning models for future missions.

1.2 Goals

The main objectives of this thesis can be summarized in two goals, one primarily aimed at the research community (Goal-01) and the other tailored to KDA's interests and the industry (Goal-02). The contributions of Goal-01 are valuable for both the research community and KDA, while Goal-02 is primarily valuable for KDA and the defense industry.

Goal-01 Understand the status quo of attacks and defenses in deep learning-based object detection that needs attention in KDA's context.

Goal-02 Derive a system architecture and requirements for the safe deployment of AI in an unmanned military setting.

1.3 Contributions

The research contribution of this thesis is the drafting of a threat model for attacks and defenses of deep learning-based object detectors in the use case of unmanned military systems. A comprehensive literature search found no other systematic literature reviews on adversarial attacks and defenses in object detection aimed at defense industry use cases, making this systematic literature review the first of its kind.

This thesis has provided KDA with important insights into the current state-of-the-art in adversarial attacks and defenses in object detection. KDA received guidance on safely deploying AI in unmanned military settings, focusing on AI implementation. Furthermore, the systematic literature review can aid in evaluating the maturity and robustness of object detection, which will determine whether it should be implemented or disregarded in future or ongoing projects. Additionally, this thesis project can inspire further research into more accurate and reliable object detection.

1.4 Thesis Report Structure

The rest of the thesis report is structured as follows: Chapter 2 - Background describes the main topics of this thesis, including deep learning-based object detection and adversarial- attacks and defenses. It also presents related work, including similar reviews on adversarial attacks and defenses in object detection and adversarial attacks in computer vision in the defense industry. Chapter 3 - Method describes the systematic literature review methodology and the requirements engineering process to reach Goal-01 and Goal-02, respectively. Chapter 4 - Results present the results from the literature review and the derived system architecture and requirements for the

safe deployment of AI in unmanned military vehicles. Chapter 5 - Discussion presents suggestions for future work and discuss the results of the thesis. At last, Chapter 6 concludes the thesis with some summarizing words.

Chapter 2

Background

This chapter summarizes the main concepts and the foundations of the project, including deep learning-based object detection and adversarial- attacks and defenses. As this thesis is a collaboration with KDA, and by that not only aimed at other researchers, it is assumed that not every reader of this thesis is an expert in AI. Therefore, this chapter gives an introduction and overview to readers who might not be as familiar with the respective field. However, some preliminary knowledge of deep learning is expected of the reader to extract the full value of the information.

2.1 Deep Learning-Based Object Detection

Applications within computer vision have become some of the most popular uses of deep learning following the success of convolutional neural networks (CNNs). These applications include image classification, object detection, visual tracking, and semantic segmentation [4, 5]. Image classification is the most straightforward task, aiming to classify images into predefined classes. A common use of image classification is medical image classification [6]. Even though image classification is the most straightforward computer vision task, it forms a base for other, more complex problems. Object detection extends the image classification task by localizing and classifying multiple objects within images. The localization feature of object detection makes it more useful in real-world applications like self-driving cars [7] and UAV bridge inspection [8]. Semantic segmentation takes another step forward by classifying every pixel in an image into predefined classes, e.g., road, pedestrian, car, and background, in a traffic scene. Visual tracking is another challenging task aimed at tracking objects across video frames, i.e., people or vehicles. The remaining sections will not describe image classification, semantic segmentation, and object tracking in more detail, as this thesis focuses on the object detection task.

As previously described, object detection aims to localize and classify objects in

images. While image classification classifies the entire image as one class, object detectors can localize and classify multiple objects in images, making it a much more complex problem. Therefore, the task is often split into multiple tasks: localization or region proposal and classification. One part of the object detector focuses on the localization of objects by proposing regions where objects might be, e.g., by using a Region Proposal Network (RPN) [9]. The RPN is a fully convolutional neural network that uses anchor boxes to propose boundaries for where objects might be. Regions are proposed with an objectness score that measures the RPN's confidence in an object being in the proposed region. The proposed regions are then sent to a classifier network that decides which class the object belongs to, including "background" to disregard object proposals. Object detectors with such task-divided architectures are referred to as two-stage- or proposal-based object detectors. A widely used family of two-stage object detectors is Faster-RCNN [9]. Two-stage object detectors have proven highly accurate but at the cost of performance. They are, therefore, often regarded as too slow for real-time use cases as they prioritize accuracy over performance.

Single-stage detectors, also called regression-based detectors, on the other hand, combine localization and classification into a single task. Popular single-stage detectors include the YOLO [10] and SSD [11] families of detectors. As single-stage detectors prioritize performance, it does come at the cost of accuracy. Therefore, the trade-off between accuracy and performance must be considered when deciding what kind of object detector is fit for the specific use case.

2.2 Adversarial Attacks and Defenses in Deep Learning

This section introduces adversarial attacks and defenses in deep learning, along with an explanation of the relevant terminology. Readers with more experience in the field may consider skipping to Chapter 4 where the results of the literature review on adversarial attacks and defenses in object detection are presented.

2.2.1 Terminology

This section describes the commonly used terms in the research field that are relevant to and actively used in this thesis. The terminology used in this thesis is based on terminology from the previous review- and survey papers in the field [12–21]. Some of these terms still need to be fully developed and may have other definitions elsewhere, as adversarial attacks are a relatively young area of research.

Adversarial Examples are images with noise-like changes that aim to be imperceptible to the human eye, similar to the ones in Figure 2.1. Adversarial examples are the most common attack method in the context of adversarial attacks in computer vision on images and videos.

Advesrarial Patches are patch-shaped adversarial examples that aim to be more useful in the physical world as they can be painted or printed as a sticker. They are usually square-shaped but might also come in different shapes.

Adversarial Camoflages differ from adversarial patches as they cover entire objects in an adversarial camouflage or texture to fool detectors rather than placing patches on objects. Like adversarial patches, adversarial camouflages usually aim to be physically realizable.

White Box Attacks refers to attacks with full access to the target model, including parameters, architecture, training- method, and data. Attacks in this category often exploit the gradient or loss of the target model while crafting the worst possible attack to trigger misclassifications with high confidence. White box attacks usually degrade the performance of the target model more effectively than other attack methods. However, they are more challenging to perform in real-life scenarios due to needing full access to the target model.

Black Box Attacks differ from white box attacks as black box attacks assume no knowledge of the target model. These attacks focus on crafting universal adversarial attacks that can degrade any model. Effective black box attacks are naturally more complex to craft due to the limitation of not having access to the target model. However, they are considered more applicable to real-world use cases, as attackers can rarely assume full access to the target model, especially in the military setting.

Transferability of adversarial attacks can have several meanings based on the specific context. However, transferability may have two meanings in this thesis: cross-model transferability or cross-task transferability. Cross-model transferability often refers to a white box attack's ability to attack other models in a black box setting. E.g., crafting an adversarial patch using a YOLO model and attacking an SSD model using that same patch. On the other hand, cross-task transferability refers to an attack method's ability to attack deep learning models specializing in a different problem. E.g., using image classification attacks on object detectors.

Targeted Attacks are adversarial attacks that target a specific class for misclassification. This outcome can be achieved by maximizing the loss of the correct class while minimizing the loss of the target class. E.g., an attack that aims for the object detector to misclassify a car as a bicycle. In this case, the attack targets the bicycle class.

Untargeted Attacks does not target a specific class when attacking object detectors. The main goal of untargeted attacks is to trigger misclassification. Thereby not caring about the specific output class as long as it is incorrect. This outcome can be achieved by maximizing the loss of the correct class - a more straightforward task than targeted attacks.

Digital Attacks are performed in the digital domain, i.e., applied to digital images or videos. Crafting adversarial attacks in the digital domain offers more flexibility as it does not require feasibility in the real world.

Physical Attacks are specifically crafted to be applicable in the real world and are, by that, not limited to the digital domain. Crafting physical attacks is more challenging than digital attacks because of the constraints imposed by the physical world.

Proactive Defenses are defense methods that aim to mitigate an attack before it happens. Typical proactive defense methods against adversarial attacks include adversarial training, robustifying model architectures, and input transformations. These defense methods are further described in Section 4.1.2.

Reactive Defenses are defense methods that attempt to mitigate attacks when they happen. The primary reactive defense method against adversarial attacks is adversarial detection. Adversarial detection is further described in Section 4.1.2.

2.2.2 Adversarial Attacks

Deep neural networks are prone to adversarial attacks, which is a type of attack that exploits the weaknesses in the functionality of deep learning algorithms. The initial adversarial attack against deep neural networks was hardly perceptible perturbations added to images, called adversarial examples [1]. Szegedy et al. [1] showed that non-randomly crafted perturbations could trigger misclassification in image recognition algorithms. Not only did the adversarial examples trick the targeted model, but they were also transferable to models with different architectures and models trained on different datasets. Figure 2.1 shows some of these original adversarial examples. The

Figure 2.1: Images on the left are the original images with their correct classification, while images on the right are the adversarial examples classified as an ostrich. The noise in the middle is the adversarial perturbations with a magnitude of 10 for visibility. The image is from a research paper by Szegedy et al. [1]

figure shows that these adversarial perturbations are not perceptible to the human eye but still fool the image recognition algorithms.

Following the introduction of adversarial attacks, Goodfellow, Shlens, and Szegedy [2] proposed the fast gradient sign method (FGSM) family of attacks in 2015. FGSM is a fast, cheap, and effective method of generating adversarial examples that exploit the gradient of the model during training. It is, therefore, dependent on having full access to the target model, i.e., a white box attack. However, they found that the adversarial examples were generalizable or transferrable to other models. The successful generalizability of adversarial examples was explained as the product of different models converging to similar parameters when trained on the same task. This early proof of transferability was the start of frequent experimentations with white box methods within a black box setting in recent studies [22–24].

Along with most research on adversarial attacks in computer vision, the FGSM attack was designed for the image classification task. Only in recent years have the research field started covering more complex problems in computer vision, like object detection. DPatch by Liu et al. [25] is one of the early adopters of adversarial attacks on object detection. Different from the untargeted FGSM attack, DPatch can be used as a targeted attack to target a specific class for misclassification. In other words, DPatch can purposefully make a detector classify a bird as a plane or a person as a chair. The DPatch attack aims to generate an adversarial patch that attacks both the box regression and image classifier of object detectors to make the targeted objects invisible to the detector. Advesrarial patches generated by the DPatch method vary in size but are very noticeable to the human eye.

On the other hand, the RPAttack by Huang et al. [26] builds upon the concept of FGSM to create imperceptible patches to objects in images to make objects invisible to detectors while keeping the patches imperceptible to humans. Adversarial patch attacks have also branched into adversarial camouflages that cover entire objects [27]. The move toward adversarial- patches and camouflages is primarily due to their applicability in the physical world, while adversarial perturbations are mostly limited to the digital domain.

2.2.3 Adversarial Defenses

Object detectors need functional and practical defense methods due to their vulnerability to adversarial attacks. Effective defenses against such attacks are crucial for the safe deployment of object detectors, or deep neural networks in general, in safety-critical systems like unmanned vehicles. Along with the introduction of adversarial examples in 2014, Szegedy et al. [1] simultaneously introduced the concept of including adversarial examples in the training data of classifiers to create a more robust model. Later, in 2015, Goodfellow, Shlens, and Szegedy [2] made adversarial training more practical by using FGSM for faster generation of adversarial examples.

Much later, in 2019, adversarial defenses were studied in the context of object detection [28, 29]. Zhang and Wang [28] focused on adversarial training for object detection, which is a multi-task problem in comparison to the single-task problem of classification. On the other hand, Xiao et al. [29] focused on another type of adversarial defense, namely adversarial detection. While adversarial training is a proactive defense method by robustifying the model before an attack occurs, adversarial detection is a reactive defense method. Instead of robustifying the model with adversarial training and risk lowering the performance of the detector on raw input, adversarial detection aims to use an AI model to detect the adversarial attack when it happens and then act in some predefined way to mitigate the threat. Other proactive adversarial defenses are also proposed in research, including attempts at robustifying the model architecture [30] and applying input transformations to lower the effect of attacks [31].

2.3 Related Work

This section covers some of the related work to this thesis. The first part covers other surveys and review articles related to adversarial attacks and defenses in object detection. In contrast, the second part covers related work on adversarial attacks in computer vision in the defense industry.

2.3.1 Surveys and Reviews on Adversarial Attacks and Defenses in Computer Vision

Several surveys and review articles have previously been conducted in the field of adversarial attacks and defenses in deep learning-based computer vision applications. In 2018, Akhtar and Mian [12] released a survey on the threats of adversarial examples on deep learning in computer vision. The main portion of the surveyed articles covers adversarial attacks on image classifiers. However, attacks on autoencoders, generative models, recurrent neural networks, deep reinforcement learning, semantic segmentation, and object detection are also covered, but in way less detail.

A dedicated survey on the object detection task was needed, and just that was released in 2020 by Xu, Zhu, and Wang [14]. It describes six adversarial attack methods against object detectors and does experiments on how to defend against them. While Xu, Zhu, and Wang [14] focused on the object detection task, the review article by Xi [15] from the same year focused on adversarial machine learning for cybersecurity and computer vision. It differentiates itself from the other articles by having more of a security perspective by describing different adversaries' points of attack, mirrored in their categorization method.

Later, in 2021, Akhtar et al. [16] released a second survey covering the advances in adversarial attacks and defenses in computer vision since their previously released paper in 2018. This survey, as its predecessor, primarily focused on classification with additional smaller sections covering other computer vision tasks. A book chapter by Oh, Xompero, and Cavallaro [17] also covers visual adversarial examples in computer vision applications, including object detection. The same year, Ren, Huang, and Yan [13] released a review of real-world physical adversarial attacks and defenses in image classification and object detection. The survey of adversarial attacks and defenses in deep learning-based autonomous driving systems by Deng et al. [18] is also highly relevant to this thesis due to its focus on autonomous systems.

The following year, in 2022, Amirkhani, Karimi, and Banitalebi-Dehkordi [19] released another highly relevant survey on adversarial attacks and defenses for object detection and their applications in autonomous vehicles. A comprehensive and technical survey on adversarial attacks and defenses in deep learning for image recognition by Wang et al. [20] was also released the same year.

The latest survey on adversarial examples in object detection was released in 2023 by Mi et al. [21].

2.3.2 Adversarial Attacks in Computer Vision in the Defense Industry

In some cases, adversarial attacks in object detection have also been studied in defense applications. These papers are highly relevant to this thesis as they are aimed at the same industry.

In 2019, the Norwegian Defence Research Establishment (FFI) published a paper on adversarial camouflage for naval vessels [32]. Aurdal et al. [32] discussed whether traditional camouflages designed to fool human observers are sufficient in a future with intelligent systems and thereby studied how to create camouflages to fool such systems. A case study was conducted on a neural network-based ship classifier distinguishing between civilian and military crafts. To create the camouflages, a generative adversarial network was trained to generate adversarial patches to be placed on the naval vessels, which successfully fooled the classifier in digital experiments.

The same year, Alfimtsev et al. [33] released a similar paper on adversarial camouflage generation for military uniforms. This study aimed to generate adversarial camouflages in the style of traditional camouflages to fool human observers and intelligent systems. Different from the paper by Aurdal et al. [32] that focused on image classification, they focused on the even more relevant object detection task of people detection. While their digital experiments were highly successful, the physical experiments in the real world were less impressive.

Later, in 2021, Wang et al. [34] proposed Dual Attribute Adversarial Camouflage (DAAC) to fool both human and artificially intelligent systems in the military setting. A limitation of their experiments is that they were only tested in the digital domain, and attacks in the real world were left as future research.

The same year, Kim et al. [35] released a similar paper on adversarial patch generation with the style of military camouflages. Use case experiments with military tank detection proved the success of their adversarial patches in the digital domain. Again, physical, real-world experiments were left as future work.

Chapter 3

Methods

This chapter describes the methodology used to achieve the two main goals addressed in Section 1.2 of the introduction of this thesis. Goal-01 on understanding the status quo of adversarial attacks and defenses in deep learning-based object detection in KDA's context was addressed with a systematic literature review and classification of papers. Goal-02 on deriving a system architecture and requirements for the safe deployment of AI in an unmanned military setting was achieved through workshops, technical analysis, expert opinion, and expert interviews with employees in KDA. The following sections describe the respective methodologies in detail.

3.1 Discovering the State of the Art in Adversarial Attacks and Defenses in Object Detection

This section summarizes the protocol for the systematic literature search of this thesis. The protocol includes the search strategy used to identify relevant papers and describes the criteria for selecting articles for inclusion in the study. Information regarding the searched database and the keywords used to craft the search string is also given. The process of snowballing in systematic literature studies is also described as it was used for a more comprehensive search. Information on how the papers were classified according to the research question of this literature review is also given.

Several papers describing the process of conducting a systematic literature search and writing a literature review were studied in advance of this systematic literature review [36–38]. In addition, Peters et al. [39] provided guidance on how to develop scoping review protocols, and Wohlin [40] described the process of snowballing in systematic literature studies.

After reading this section, the systematic literature search and classification process results can be found in Section 4.1.2. In that section, the included papers are

summarized and presented in tables according to their classification.

3.1.1 Justifying the Systematic Literature Review

A systematic literature review was conducted to gather all the relevant information on the field of adversarial attacks and defenses in object detection. The aim of the systematic literature review was to gain comprehensive and unbiased knowledge of the current state of adversarial- attacks and defenses in the object detection task.

It was beneficial to perform a systematic literature review on the topic, as adversarial attacks are a relatively new field of study. Furthermore, most research on adversarial attacks is aimed at the image classification problem. Some surveys on adversarial attacks against object detectors exist, but their focus is on future research rather than the industry. Therefore, this systematic literature review was conducted to collect the desired information to give recommendations for the specific use case of unmanned systems in the defense industry.

3.1.2 Database Search Strategy

Scopus[1] was selected as the scientific database for this systematic literature review. Scopus has a collection of titles, abstracts, and keywords to peer-reviewed journal articles, conference papers, and other kinds of scientific papers. Google Scholar[2], which is an alternative to Scopus, was not used as Scopus provides better quality assurance by only including peer-reviewed literature.

Search String

The search string for finding relevant literature was crafted with a combination of words relating to adversarial- attacks, defenses, and object detection. It was searched for in the articles' titles, abstracts, and keywords only to exclude articles not focusing on the respective fields. The extra work required to review all articles mentioning the keywords in any section of the article would have been very time-consuming and was therefore not done. E.g., many articles mention object detection as a common use of deep learning in their introduction section. Some relevant articles might have been missed due to the strict search. However, one should expect that articles focusing on these fields have included these essential keywords in their title, abstract, or keywords.

Some exclusion criteria were used in the initial search string to limit manual exclusion using Scopus' advanced search functionality. Therefore, articles written in

[1] Scopus: https://www.scopus.com

[2] Google Scholar: https://scholar.google.com/

languages other than English were not included in the search results, with the same applying to the conference review document type. The conference review document type in Scopus provided no value as it was merely a list of conference papers already included in the same search. Non-English written papers were excluded to remove the possibility of translation errors, and due to the search providing sufficient search results without them.

The following search string was used in Scopus[3] for gathering research to be considered for inclusion in the start set of papers:

> *Search String: TITLE-ABS-KEY ("adversarial attack*" OR "adversarial example*" OR "adversarial patch*" OR "adversarial defen?e" OR "adversarial robustness" AND "object detect*") AND (EXCLUDE (DOCTYPE , "cr")) AND (LIMIT-TO (LANGUAGE , "English"))*

Inclusion- and Exclusion Criteria for the Start Set of Papers

The criteria for being further evaluated for inclusion in the review and being used for snowballing were that the articles needed to focus on adversarial attacks against object detectors or how to defend object detectors against adversarial attacks. If the articles did not fit these inclusion criteria, they were not included in the start set of papers. As previously described, two exclusion criteria were used: excluding conference reviews and papers written in non-English languages.

3.1.3 Snowballing

Combining the database search with additional methodologies for systematic literature reviews can be beneficial for a more comprehensive search. When performing a systematic literature review, relevant articles can be missed when only doing a single database search. Therefore this systematic literature review followed the snowballing technique described by Wohlin [40] to ensure a more comprehensive search. Snowballing helped include relevant literature missed by the initial database search because some authors used a different terminology than what was specified in the search string.

Snowballing is a technique to gather relevant papers for systematic literature reviews by iteratively identifying relevant papers using backward- and forward snowballing on a start set of papers. Snowballing starts by defining a search string for one or several research databases to gather relevant papers. The search string is usually a combination of relevant keywords, with optional inclusion and exclusion

[3]Scopus: https://www.scopus.com

criteria depending on the database. The number of results of the initial search should be an amount that thoroughly covers the topic of choice but also few enough for it to be feasible for the researcher to perform the literature review. Diversity is wanted in authors, publishers, and countries for a nonbiased review. When the database and search string are decided, all papers in the search results will be assessed for inclusion or exclusion to form the start set. The start set of papers in this literature review resulted from the initial database search using the previously described search string and inclusion- and exclusion criteria.

Backward snowballing is assessing the papers referenced in the included papers. It was done by reading the titles of the papers, and when in doubt if the paper should be included or not, the keywords and abstracts were read. If that was insufficient, the paper was read, but this was generally avoided as it makes an already time-consuming process even more time-consuming. Forward snowballing followed the same procedure as backward snowballing, with the only difference being that the papers cited by the included papers were assessed instead of the referenced papers.

As snowballing is an iterative process, every new paper included by backward- and forward snowballing on the start set of papers also needed to undergo backward- and forward snowballing. The same applied to the new papers included in the second iteration and for every iteration until no new relevant papers were found. The iteration that resulted in no new relevant papers was considered the last iteration and marked the end of the snowballing process.

3.1.4 Data Extraction

When all the relevant papers had been collected, it was time to extract their relevant information. Before the data extraction could begin, further evaluation was conducted on whether the collected papers would be included in the final review. The inclusion- and expulsion criteria are described in this section, and the method used to categorize the included papers is summarized.

Inclusion- and Exclusion Criteria

The inclusion- and exclusion criteria used to decide which papers should be included in the literature review are shown in Table 3.1. The first criterion for inclusion is that all papers must be peer-revied. This criterion was ensured by only including papers available in the Scopus research database. All papers must also focus on adversarial attacks, adversarial defenses, or both and conduct at least one experiment with an object detector. The datatype used for object detection needed to be either images or videos.

Inclusion Criteria	
1	The paper must be peer-reviewed.
2	The paper must focus on adversarial attacks, adversarial defenses, or both.
3	The paper must include experiments with at least one object detector.
4	The datatype used for object detection must be either images or videos.
Exclusion Criteria	
1	The paper is not peer-reviewed.
2	The paper does not conduct any experiments and is solely a theoretical study, e.g., a survey- or review paper.
3	The paper focuses on salient object detection.
4	The papers use face detection or face recognition as the experimental use case.
5	The object detectors use 3D data like point clouds or others.

Table 3.1: Inclusion and exclusion criteria for data extraction.

Any papers that were not peer-reviewed were excluded without further consideration, regardless of popularity. Any papers not conducting experiments with adversarial attacks or defenses in object detection were also excluded. This exclusion criterion also covers review- and survey papers. Papers focusing on salient object detection were also excluded due to the significant difference from traditional object detection. The same applied to papers focusing on 3D object detection using 3D data like point clouds. Videos are not considered 3D data in this study, as object detection in videos usually runs on individual video frames. The use case examples of face- recognition and detection were also excluded due to their difference from object detection in more relevant use cases like pedestrian detection.

Classification Criteria of the Included Papers

After performing the systematic literature review, the papers and findings needed to be classified according to the use case of safely deploying AI on the edge of unmanned military vehicles. The result of this classification is the tables presenting the state-of-the-art in adversarial attacks and defenses, Table 4.1 on Page 27 and Table 4.2 on Page 28, respectively. First, the papers were divided based on the topic of the paper, being either adversarial attacks or adversarial defenses. This division was done as categorization within these main themes is naturally different as they follow different methodologies and aim for different results.

Then the different topic-specific categories were decided based on the standard practices for categorization found in the collected survey and review papers summarized in Section 2.3.1. However, modifications were made to fit better with the use

case of unmanned military vehicles, which is the target industry use case of this thesis. These modifications were considered beneficial as most previous review papers are very technical, mainly aimed at future research rather than the industry. Therefore, some of the more detailed technical classifications were not carried over to this thesis in favor of a tailored classification approach in line with the use case of object detection in unmanned military vehicles.

The following lists describe more detailed information on the classification techniques used within the two topics of adversarial attacks and defenses and their shared classification techniques.

Common Classifications

Year of Publication The year of publication was noted for every paper to present the trend in the research area and show how adversarial attacks and defenses have evolved over time.

Detector The object detectors used in the experiments were documented to gain insight into how thoroughly the attacks and defenses were tested. The different types of object detectors, including one-stage detectors, two-stage detectors, and transformers, are different in functionality and may require separate attacks. Therefore, if an attack shows positive attack results on different detector types, it should be considered a higher threat due to its cross-model transferability.

Attack-Specific Classifications

Attack Method The papers on adversarial attacks were classified based on their attack method, like adversarial- examples, patches, and camouflages. The attack methods were partially found during the initial background study and during the extraction phase of the literature review when reading the included papers.

Attack Type Every paper on adversarial attacks was classified based on their attack type. In the context of this thesis, the attack type includes whether an attack is a white box or black box attack and whether the attack is targeted or untargeted.

Domain As the literature review results were used to assess the threat of adversarial attacks against object detection in unmanned vehicles, it was highly beneficial to know if the attacks had been tested in the digital or physical domain. Adversarial attacks in the digital realm are only effective if the attacker somehow gains access to the system, but physical adversarial attacks can be made regardless. Therefore, every attack was classified based on the domain they were tested in

to make it more intuitive to decide on suitable defense methods for a secure AI system against adversarial attacks.

Defense-Specific Classifications

Defense Method The defense methods were classified on two levels: on a higher level as either a proactive or reactive method and on a more detailed level. The more detailed defense methods were found during the extraction phase of the literature review when reading the included papers and are presented in Section 4.1.2.

Defense Against The defense methods were also classified by which adversarial attack they were designed to safeguard against. This classification connected each defense method to at least one of the included adversarial attack methods.

3.2 Securing the Edge-Based AI System for Unmanned Vehicles

This section describes the methodology used to derive the system architecture and requirements for a robust and secure edge-based AI system for unmanned military vehicles (Goal-02). The methodology described is more similar to a requirements engineering process than a formal scientific method as the systematic literature review. As described in Section 1.1, restrictions that followed the military use case, like limitations in communication technology and computing power, and the strict requirements for confidentiality, integrity, and availability, were key focus points in the design process of the system.

First, the stakeholder needs and requirements were discussed, understood, and documented. The stakeholder needs describe what is needed or wanted from the system, while the stakeholder requirements are specific, measurable requirements for the system. The stakeholder needs and requirements naturally needed to be defined before the work on deriving the system architecture and requirements could begin. The following subsections summarize these respective processes.

3.2.1 Understanding the Stakeholder Needs and Requirements

For practical reasons, KDA was the only stakeholder in this engineering process. Drafting the stakeholder needs and requirements was the first task of this part of the thesis and took place at the very beginning of the project. It started as verbal discussions with KDA before the discussions turned into a written project proposal. The document containing the project proposal was sent back and forth and was revised a

few times to clarify what was wanted of the thesis and what was considered feasible. The product of that process is described in Section 1.1 - Project Description.

The discussion to further define the stakeholder needs and requirements was picked up during weekly supervision meetings with KDA, which remained a continuous process through the thesis timeframe. These definitions became the written stakeholder needs and requirements shown in Section 4.2.2 and Section 4.2.3.

A more extensive workshop was also conducted with the involved employees at KDA towards the end of the thesis timeframe. This workshop was used to present the findings of the thesis so far and exchange and discuss ideas regarding the design of the system. This workshop differed from the weekly supervision meetings as the presented material was more prepared, and more people attended.

3.2.2 Derviving the System Architecture and Requirements

The process of deriving the system architecture and the list of requirements was heavily based on the stakeholder needs and requirements. It was essential to understand the technical context of unmanned vehicles, the platform and its communication infrastructure, and the placement of the AI functionalities on the platform to meet and fulfill the stakeholder needs and requirements. This understanding was achieved through weekly meetings and the aforementioned workshop with KDA.

3.2.3 Deriving the Security Requirements

It was crucial to ensure the safe deployment of AI in unmanned military systems. Therefore, KDA requested a summary of the risks of adversarial attacks that should be treated, alongside suggestions for security measures for risk reduction. The result of the systematic literature review on adversarial attacks and defenses in object detection played a massive part in deriving the main threats against the AI functionality of the system, and the recommended defenses to mitigate those threats.

The scope of the systematic literature review covered AI-specific defenses for robustifying object detection. However, it did not cover any general security methods required to secure the system. Therefore, the collaboration with security experts from both OsloMet and KDA also played an essential part in securing the system against adversarial attacks.

Chapter 4

Results

This chapter presents the findings from the research activities that were performed in order to address the two goals of the thesis. First, the findings of the systematic literature search leading to research Goal-01 - understand the status quo of attacks and defenses in deep learning-based object detection that needs attention in KDA's context - are summarized. The main results are presented in two tables categorizing the papers found during the review process. Also presented are additional statistics of the covered adversarial attack methods, attack types, domain, adversarial attack method's domain distribution, object detector families, defense methods, and the defense methods' targeted attack methods.

Following the results of the literature review are the results of Goal-02 of this thesis: deriving a system architecture and requirements for the safe deployment of AI in an unmanned military setting. These results include the stakeholder needs and requirements, the AI pipeline, threats against the AI pipeline, a defense strategy to mitigate those risks, and a system architecture with system- and security requirements.

4.1 State-of-the-Art Adversarial Attacks and Defenses in Object Detection

This section presents the results of the systematic literature review on attacks and defenses in deep learning-based object detection. The results of this literature review bring together various research works, methodologies, and approaches related to adversarial attacks and defenses in object detection. The results include a comprehensive overview of the topic by covering different attack methods, their impact on object detectors in the use case of unmanned military vehicles, and the various defense methods that implementers can use to mitigate such attacks. By providing a broad overview of the topic, readers are enabled to gain a comprehensive

understanding of the field.

A comprehensive list of references to relevant research articles is provided, giving interested readers the ability to dive deeper into the details of specific adversarial attacks or defenses. By doing so, the results of this thesis can be a starting point for further exploration and research in the field.

4.1.1 Results From the Literature Search

This section presents the results of the systematic literature search. The Scopus[1] database search using the search string defined in Section 3.1.2 resulted in 268 peer-reviewed papers. After reviewing these papers using the inclusion and exclusion criteria defined in Section 3.1.2, 163 papers remained and formed the start set for snowballing in the systematic literature review. The snowballing iterations went as follows:

The first iteration of snowballing resulted in 14 peer-reviewed papers found through backward snowballing and 12 peer-reviewed papers found through forward snowballing. The search string used in the Scopus database search covered the field of adversarial attacks and defenses in object detection well, as only 26 new papers were found when snowballing the start set of 163 papers.

The second iteration of snowballing was conducted on the newly acquired 26 peer-reviewed papers found during the first iteration. This iteration resulted in no new papers through backward snowballing and only one new paper through forward snowballing. Thereby, the second iteration of snowballing resulted in only one new paper.

The third iteration of snowballing was the final one, as no new relevant papers were found when snowballing the one article found in the second iteration. This iteration marked the end of the snowballing process, which resulted in a new total of 190 peer-reviewed papers to be considered for inclusion in the final review.

The 190 papers were then reviewed for inclusion in the review using the inclusion and expulsion criteria defined in Section 3.1.4. These are different criteria from the inclusion and exclusion criteria from the start set as they required a finer look at the content of papers. The papers that made it through got further reviewed using the classification criteria defined in Section 3.1.

The main result of the systematic literature review is presented in two tables: Table 4.1 presenting the classifications of adversarial attacks, and Table 4.2 presenting the classification of defense methods against adversarial attacks.

[1] Scopus: https://www.scopus.com/

Attack	Year	Method	Attack Type				Domain		Detector		
			WB	BB	T	UT	D	P	OS	TS	VT
DAG [41]	2017	Adversarial Example	✓	✓	✓		✓			✓	
Yang et al. [42]	2018		✓		✓		✓		✓		
Zhao et al. [43]	2019			✓		✓	✓		✓	✓	
Balda et al. [44]			✓			✓	✓		✓		
Co et al. [45]				✓		✓	✓		✓		
RAP [46]			✓	✓		✓	✓			✓	
Wei et al. [47]			✓	✓		✓	✓		✓	✓	
Evaporate Attack [48]	2020			✓		✓	✓		✓	✓	
Wang et al. [3]			✓	✓		✓	✓			✓	
PPAA [49]				✓		✓	✓		✓		
TOG [50]			✓	✓	✓	✓	✓		✓	✓	
Yang et al. [51]			✓	✓		✓		✓	✓	✓	
CAP [52]			✓			✓	✓		✓		
EfficientWarm [53]			✓	✓		✓	✓		✓	✓	
Lu et al. [54]				✓		✓	✓		✓	✓	
FA [23]			✓	✓		✓	✓		✓	✓	
Liao et al. [55]			✓	✓		✓	✓		✓	✓	
HNM-PGD [56]			✓		✓		✓		✓	✓	
Chen et al. [57]				✓		✓	✓		✓		
MI-FGSM [58]			✓	✓		✓	✓			✓	
Zhang et al. [59]				✓		✓	✓		✓	✓	
Huang et al. [60]			✓		✓	✓	✓			✓	
Wang et al. [61]	2021		✓			✓	✓			✓	
Haoran et al. [62]				✓		✓	✓		✓	✓	
Kuang et al. [63]				✓	✓		✓		✓		
Huang et al. [64]			✓		✓			✓		✓	
Xiao et al. [65]			✓		✓	✓	✓		✓	✓	
Yuan et al. [66]			✓			✓	✓			✓	
NaturalAE [67]			✓	✓		✓		✓	✓	✓	
PRFA [68]				✓		✓	✓		✓	✓	
Wang et al. [69]			✓			✓	✓		✓	✓	
Nezami et al. [70]			✓		✓	✓	✓		✓		
SocialGuard [71]				✓		✓	✓			✓	
Liao et al. [72]			✓	✓		✓	✓		✓	✓	
Li et al. [73]			✓	✓		✓	✓		✓	✓	
U-DOS [74]			✓	✓		✓	✓		✓	✓	
GARSDC [75]	2022		✓	✓	✓	✓	✓		✓	✓	✓
ADC [76]			✓		✓		✓				
Choi et al. [77]			✓			✓	✓		✓		
Raja et al. [78]			✓	✓		✓	✓		✓		
Re-AEG [79]			✓			✓	✓		✓		
AT-BOD [80]				✓		✓	✓		✓	✓	
RBM [81]				✓	✓	✓	✓		✓	✓	
Kang et al. [82]			✓			✓	✓		✓		

Attack	Year	Method	Attack Type				Domain		Detector		
			WB	BB	T	UT	D	P	OS	TS	VT
Daedalus [83]	2022	Adversarial Example	✓	✓	✓		✓	✓	✓		
DMA [84]				✓		✓	✓		✓	✓	
EvoAttack [85]				✓		✓	✓	✓	✓		
Zanddizari et al. [86]				✓		✓	✓		✓		
Wang et al. [87]			✓			✓	✓			✓	
JND [88]			✓		✓	✓	✓		✓		
Li et al. [89]			✓	✓		✓	✓		✓	✓	
RAD [24]			✓	✓		✓	✓		✓	✓	
Hu et al. [90]			✓	✓		✓	✓	✓	✓	✓	
Cai et al. [91]			✓	✓		✓	✓		✓	✓	
Ye et al. [92]				✓		✓	✓		✓	✓	
Eykholt et al. [93]	2018	Adversarial Patch	✓	✓		✓		✓	✓	✓	
Chambers et al. [94]	2019		✓		✓		✓	✓	✓	✓	
DPatch [25]				✓	✓	✓	✓		✓	✓	
Thys et al. [95]			✓			✓	✓	✓	✓		
Zhao et al. [96]			✓	✓		✓		✓			
Adharaki et al. [97]	2020		✓			✓	✓		✓		
TOG-Patch [50]			✓		✓	✓	✓		✓		
Li et al. [98]			✓		✓		✓		✓	✓	
Li et al. [99]			✓	✓		✓	✓		✓	✓	
Huang et al. [100]			✓		✓	✓	✓	✓		✓	
Zhao et al. [101]			✓	✓		✓	✓		✓	✓	
SAA [102]			✓	✓		✓	✓		✓	✓	
Shi et al. [103]	2021		✓		✓		✓		✓	✓	
AGAP [104]			✓	✓		✓	✓		✓	✓	
Raja et al. [22]			✓	✓		✓	✓			✓	
Kim et al. [35]			✓			✓	✓		✓		
DAAC [34]			✓		✓	✓	✓		✓	✓	
eSLP-GAN [105]			✓	✓	✓	✓	✓		✓	✓	
LAPs [106]			✓			✓	✓	✓	✓		
Hu et al. [107]			✓			✓	✓	✓	✓	✓	
Wang et al. [69]			✓			✓	✓		✓	✓	
RPAttack [26]			✓			✓	✓		✓	✓	
Patch-Noobj [108]			✓	✓		✓	✓		✓	✓	
Zolfi et al. [109]			✓	✓		✓	✓	✓	✓	✓	
Wang et al. [110]			✓	✓		✓	✓	✓	✓	✓	
Zhang et al. [111]	2022		✓			✓	✓	✓	✓		
Re-AEG [79]			✓			✓	✓		✓		
Dong et al. [112]			✓			✓	✓		✓	✓	
AP-PA [113]			✓	✓		✓	✓		✓	✓	✓
CAMA [114]			✓		✓	✓	✓		✓		
Cai et al. [115]			✓	✓	✓		✓		✓	✓	✓
Attention-Fool [116]			✓		✓	✓	✓				✓
MACA [117]				✓		✓	✓	✓	✓	✓	

Attack	Year	Method	Attack Type				Domain		Detector		
			WB	BB	T	UT	D	P	OS	TS	VT
Jia et al. [118]	2022	Adversarial Patch	✓		✓	✓		✓	✓		
Toheed [119]				✓		✓	✓		✓	✓	
Du et al. [120]			✓			✓	✓	✓	✓		
FRAN [121]			✓			✓	✓		✓		
CAMOU [27]	2019	Adversarial Camouflage		✓		✓	✓		✓	✓	
Alfimtsev et al. [33]				✓		✓	✓	✓		✓	
Xi et al. [122]	2020			✓		✓		✓	✓		
CCA [123]				✓		✓		✓	✓		
UPC [124]			✓	✓	✓	✓	✓	✓	✓	✓	
Deng et al. [125]	2021		✓	✓		✓	✓		✓	✓	
DAS [126]				✓		✓	✓	✓	✓	✓	
DAAC [34]			✓			✓	✓	✓	✓		
DTA [127]	2022		✓	✓		✓	✓	✓	✓	✓	
FCA [128]			✓	✓		✓	✓	✓	✓	✓	
CAC [129]				✓	✓		✓	✓	✓	✓	
Xu et al. [14]	2020	Advesarial Wearable		✓		✓	✓	✓	✓	✓	
LAPs [106]	2021		✓			✓	✓	✓	✓		
Hu et al. [107]			✓			✓	✓	✓	✓	✓	
Hu et al. [130]	2022			✓	✓		✓	✓	✓	✓	✓
InvisibiliTee [131]				✓		✓	✓	✓	✓	✓	
meshAvd [132]	2019	Adversarial Mesh	✓	✓	✓	✓	✓		✓		
Huang et al. [133]	2019	Adversarial Boarder	✓	✓		✓	✓	✓	✓	✓	
ShapeShifter [134]	2019	Adversarial Sign	✓		✓	✓	✓	✓		✓	
Huang et al. [135]	2020		✓	✓		✓	✓	✓	✓	✓	
Chen et al. [136]	2020	Content Distinguish	✓	✓	✓		✓		✓		
BBGAN [137]	2020	Adversarial Environment		✓		✓	✓	✓	✓		
SLAP [138]	2021	Adversarial Projection	✓	✓		✓		✓	✓	✓	
Deep-Dup [139]	2021	Adversarial Weights		✓	✓	✓	✓		✓		
Chow et al. [140]	2021	Perception Poisoning		✓	✓	✓	✓			✓	
Poltergeist [141]	2021	Accoustic Waves		✓	✓	✓	✓	✓	✓	✓	
Xu et al. [142]	2022	Adversarial Background Image		✓		✓	✓	✓	✓	✓	
Wei et al. [143]	2022	Advesarial Sticker		✓		✓	✓	✓	✓		

Table 4.1: **Adversarial attacks in object detection.** WB: White box, BB: Black box, T: Targeted, UT: Untargeted, D: Digital, P: Physical, OS: One-stage, TS: Two-stage, VT: Vision transformer.

Defense	Year	Method		Defense Against		Detector	
				AE	AP	OS	TS
Zhang et al. [28]	2019	Proactive	Adversarial Training	✓		✓	
Hu et al. [144]	2020			✓		✓	
Saha et al. [145]					✓	✓	
Det-AvdProp [146]	2021			✓		✓	
GAT [147]				✓		✓	
Chen et al. [148]				✓		✓	
Xu et al. [149]				✓		✓	✓
QUAT [150]	2022			✓		✓	✓
UDFA [151]				✓		✓	✓
VHAT [152]				✓			✓
Choi et al. [77]				✓		✓	
Raja et al. [78]				✓		✓	
Xue et al. [153]					✓	✓	
Naseer et al. [31]	2020		Input Transformations	✓			✓
Cheng et al. [154]				✓		✓	
Chiang et al. [155]					✓	✓	
Zhou et al. [156]					✓	✓	
Bao et al. [157]	2021				✓	✓	
Yu et al. [158]	2022				✓	✓	✓
Wang et al. [159]				✓		✓	
Zhang et al. [160]							✓
TEDM [161]	2020		Robustifying the Model Architecture	✓			✓
Saha et al. [145]					✓	✓	
Karimi et al. [30]	2021			✓		✓	
FUSE [162]				✓	✓	✓	✓
Xu et al. [149]				✓		✓	✓
Amirkhani et al. [163]	2022			✓		✓	✓
Dong et al. [164]				✓		✓	
Searle et al. [165]			System Security		✓	✓	✓
AdvIT [29]	2019	Reactive	Adversarial Detection	✓		✓	
CARAMEL [166]	2020			✓			
SCEME [167]				✓			✓
Chai et al. [168]	2021			✓		✓	
Yin et al. [169]				✓		✓	✓
Xiang et al. [170]					✓	✓	✓
APM [171]					✓	✓	
Liu et al. [172]	2022				✓		✓
LanCeX [173]					✓	✓	
Xue et al. [153]					✓	✓	

Table 4.2: **Adversarial defenses in object detection.** AE: Adversarial Example, AP: Adversarial Patch, OS: One-stage, TS: Two-stage.

An additional table was created, presenting the tools and datasets found during the systematic literature search that are beneficial for practitioners and for future research on adversarial attacks and defenses. Table A.1 covering the tools and datasets can be found in Appendix A. The tools and datasets table was placed in the appendix as it does not strictly cover adversarial attacks or defenses, as the research question requests. However, they are worth noting for future research and development.

4.1.2 Categorization and Analysis of the Reviewed Literature

This section presents the categorization results and analysis of the reviewed literature. The results are a combination of statistics from the review and various observations made throughout the review process. It starts by explaining how the grouping of adversarial attack methods facilitated the threat modeling for the system proposed in Section 4.2. Following that are various results related to the attack types found in research. The importance of the domain in which the adversarial attacks were tested is then stated, along with the domain distribution across the adversarial attack methods. The object detectors covered in the research are then presented to gain insight into which detectors are usually attacked and defended in studies. Lastly, the grouping of defense methods against adversarial attacks is shown, and the attack methods targeted for mitigation are presented.

Grouping the Adversarial Attack Methods

One of the discoveries of the literature review was that many attacks have similar methods of attacking the functionality of object detectors, like the three most common attack methods: adversarial examples, adversarial patches, and adversarial camouflages. These similarities made grouping the attacks under the same attack method possible, making evaluating their threat against unmanned military vehicles much more feasible given the high number of attacks. As an example, this grouping made it possible to evaluate the category of "adversarial examples" as a group instead of evaluating the threat level of 55 individual attacks. Grouping by attack method, in turn, improved the results' readability and removed unnecessary repetitive work. This grouping of attack methods can be seen in the method column of the summary table for adversarial attacks in object detection, Table 4.1.

As seen in Table 4.1 and Figure 4.1, the most frequently covered attack method is the original adversarial examples. Adversarial example generation is the method of 46% of the covered adversarial attacks, which clearly makes it the most common adversarial attack type in object detection. As previously described in Section 2.2 but put in the context of object detection, adversarial examples aim to be imperceptible

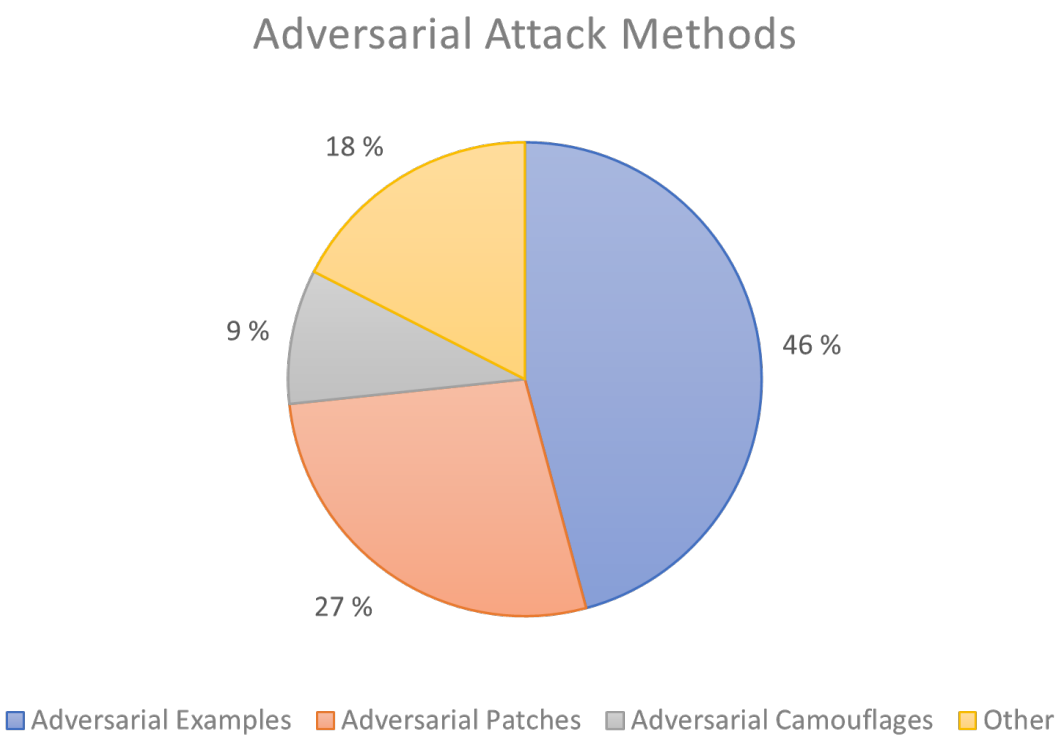

Figure 4.1: The distribution of the covered adversarial attack methods.

noise added to images to make object detectors wrongly predict, misclassify, or miss objects.

Adversarial patch generation is the method of 27% of the covered adversarial attacks, as shown in Figure 4.1. This amount of coverage makes it the second most frequently used adversarial attack method in object detection. Unlike the typical approach of adversarial examples by applying noise to an entire image, adversarial patches aim to generate a patch that is typically placed on the object to be hidden or misclassified. Adversarial patches are also more physically realizable than adversarial examples, as they can easily be printed as stickers or posters.

The following method, called adversarial camouflage, extends the concept of adversarial patches by generating an adversarial camouflage that covers the entire object. As seen in Figure 4.1, adversarial camouflages are the third most common attack type in object detection, covering 9% of the included papers on adversarial attacks in object detection.

The vast majority of the papers fit into one of these three main categories, and the ones that did not were given a descriptive method name, e.g., adversarial wearables. Most of these papers were considered less critical during the threat analysis when deriving the system for unmanned military vehicles for Goal-02 of the thesis. This assessment was primarily made due to the fact that the particular use case experiments of these studies do not fit the use case of unmanned military vehicles.

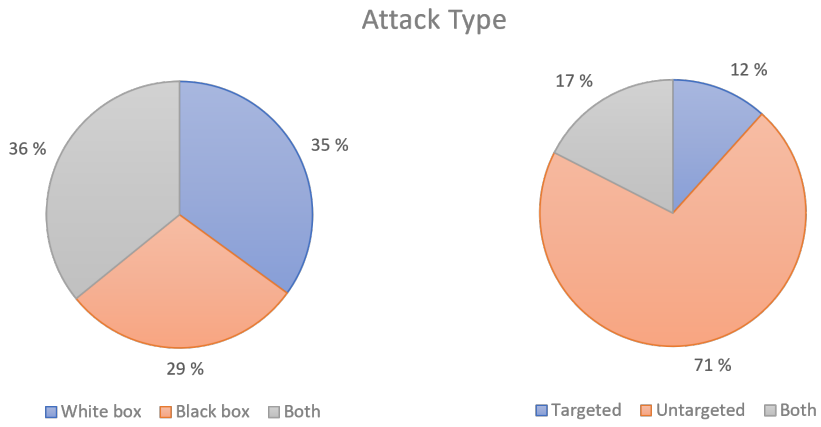

Figure 4.2: The distribution of attack types.

Multiple Adversarial Attack Types

Originally it was planned to classify each attack as either a white box or black box attack, and as either a targeted or untargeted attack. However, reading the literature made it clear that many white box attacks were also tested in a black box scenario to test the transferability of attacks. Some attacks also had different versions, e.g., one that is targeted and one that is untargeted. As seen in Figure 4.1, the TOG attack [50] is an example of an attack with a targeted and an untargeted variant and white box and black box experiments. Therefore, the classification rules regarding the attack types were changed so that each attack could have multiple classifications to account for such scenarios. This classification can be seen in the attack type column in Figure 4.1.

White Box vs. Black Box Attacks

The adversarial attacks included in the systematic literature review were classified based on their attack type to indicate their threat level against unmanned military vehicles. One of the classifications was whether the attackers assumed white box or black box knowledge of the target model when attacking.

As white box attacks have full access to the targeted model, they are often more effective in attacking the model than methods that do not have this privilege. However, their limitation is that the attacker must somehow access the targeted model. On the other hand, black box attacks do not need access to the target model. Not having access to the target model makes it harder to create effective attacks. Still, attacks that are effective without model access are much more dangerous in real-world scenarios where access to the target model is unlikely.

As seen in Figure 4.2, 35% of attacks were tested in a white box scenario only,

29% of attacks were tested in a black box scenario only, and 36% of attacks were tested in both scenarios. Adversarial attacks tested in both white box and black box experiments are usually white box attacks at function. I.e., They exploit the advantage of being white box attacks and test their performance against other black box models in what is referred to as black box transfer attacks.

When reviewing the papers, it was observed that it is common for researchers to test their white box attacks as black box transfer attacks. This observation is represented in Figure 4.2, with half of the white box attacks also being tested as black box attacks. This focus on black box attacks is primarily because attackers can not expect full access to the target model in real-world attack scenarios. This realization is especially true for our military use case due to the strict requirements for security, making black box attacks the more likely threat.

Targeted vs. Untargeted Attacks

A second attack type was used to classify the adversarial attacks included in this literature review. This classification shows whether the proposed attacks are targeted or untargeted. Both attack types are considered dangerous if performed successfully but in different ways. In some cases, the attacks have different versions to cover both targeted and untargeted purposes.

As the targeted attacks can guide predictions towards a class of choice, they can be used more intellectually. E.g., in a military setting, an adversary can camouflage their vehicles to be classified as friendly by their opponent's AI. Such attacks can have dangerous consequences in the future with an increasing amount of unmanned systems. These attacks can put human lives at risk, e.g., if a UAV on a reconnaissance mission misclassifies the opponent as friendly, and the human troops move along with this misinformation.

Untargeted attacks also pose a threat to the military use case. Their main difference from targeted attacks is that they do not target a specific class when attacking. The goal of untargeted attacks is to make the target model misclassify the object as any class other than the correct one. However, in object detection, untargeted attacks also cover attacks that make the model predict many more objects than usual, acting as denial-of-service attacks [50]. Untargeted attacks cover a broader range of attacks than targeted attacks, but their common goal is to render the attacked model useless.

Most attacks covered in the reviewed articles were classified as untargeted attacks. As seen in Figure 4.2, 71% of attacks only provide untargeted versions. Only 12% of the covered attacks are solely focused on targeted attacks, and 17% have both targeted and untargeted versions. This distribution does make sense as the optimization problem of untargeted attacks is simpler than targeted attacks. Untargeted attacks

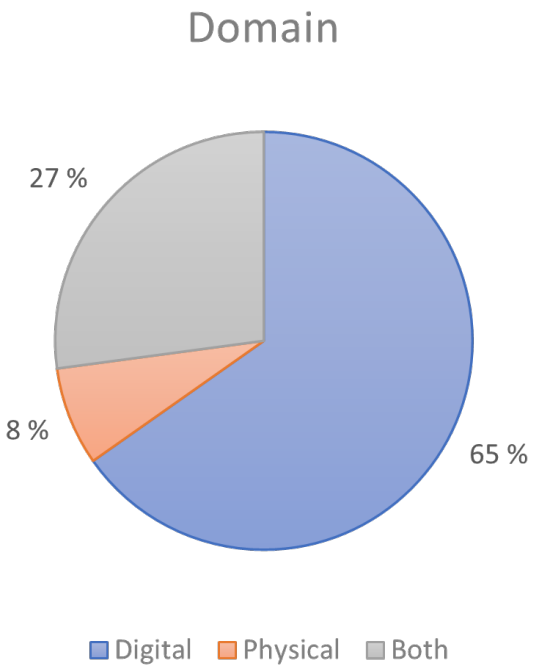

Figure 4.3: The distribution of the domain used in adversarial attack experiments.

only need to reduce the accuracy of the correct class, while targeted attacks also need to increase the accuracy of the targeted class.

Domain Importance of Adversarial Attacks

The distribution of the domain used in experiments with adversarial attacks in the included papers is shown in Figure 4.3. An entire 92% of the studied attacks were experimented with in the digital domain, with 65% of them being tested in the digital domain only. That leaves 8% of papers tested in the physical domain only. However, 35% of the total attacks were experimented with in the physical world.

The domain of the attack, either digital or physical, was considered highly relevant to the threat evaluation in the second part of the thesis. For digital attacks to be effective, they need access to the target system in the context of unmanned military vehicles, as unmanned vehicles operate in and perceive the physical world and not the digital one. Therefore, digital adversarial attacks can not be used against unmanned military vehicles unless the attacker somehow manages to sneak the digital attacks into the system. However, this is unlikely due to the strict security requirements of military systems, in contrast to civil systems like the widely covered use case of self-driving cars. The scenario of digitally attacking the system will be further described during the threat evaluation of adversarial attacks against the AI-based system for unmanned military vehicles.

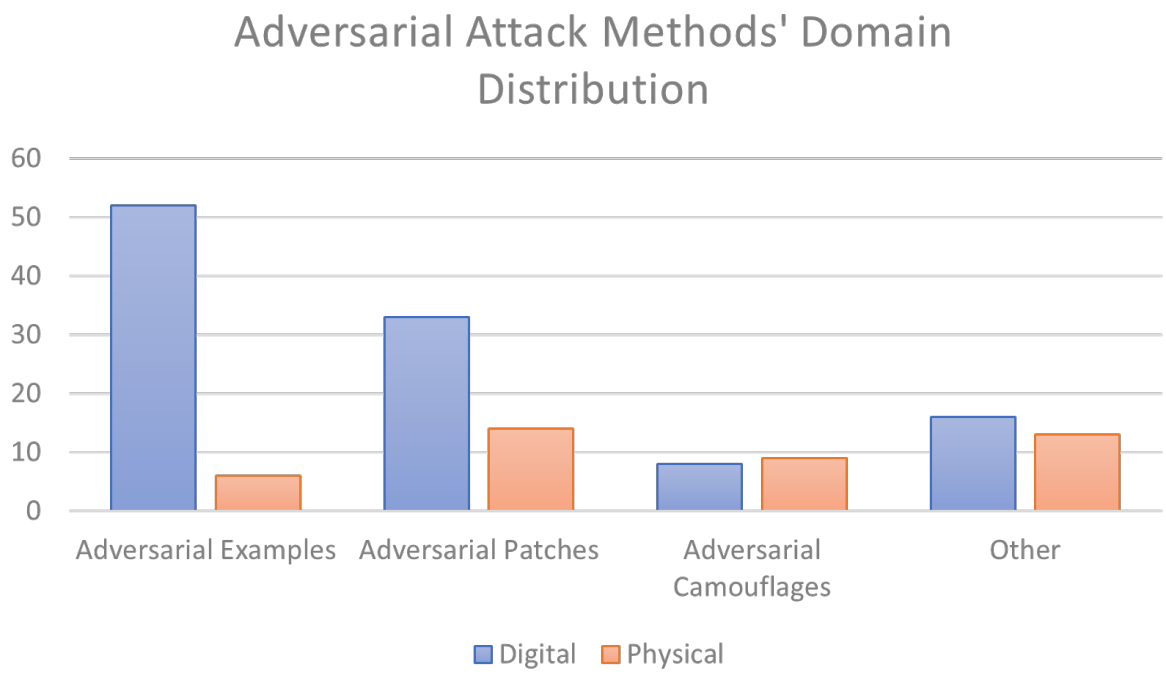

Figure 4.4: The covered adversarial attack methods' domain distribution.

On the other hand, physical adversarial attacks are of much higher interest as these adversarial attacks can be effective in the physical world. Because of that, they do not require access to the target system. This ability to attack systems operating and perceiving the physical world makes them especially dangerous towards all cyber-physical systems, including KDA's use case of unmanned military vehicles.

Adversarial Attack Method's Domain Distribution

When looking at the domain distribution of adversarial attack methods in Table 4.4, it is clear that the dominant experimental domain varies between the covered attack methods. Adversarial examples are rarely tested in the physical domain compared to the other adversarial attack methods. However, this makes sense as the attack method mainly focuses on the digital domain by making minor pixel-wise image adjustments.

Adversarial patches, on the other hand, have a distribution closer to half of the adversarial patch attacks being tested in the physical domain. This more significant percentage of physical experiments is justified as adversarial patches are more easily convertible to the real, physical world, e.g., by printing them on posters to be placed on physical objects.

There are more documented experiments in the physical domain compared to the digital domain for adversarial camouflages, indicating a shift in focus towards the former. This shift is understandable as adversarial camouflages take the concept of

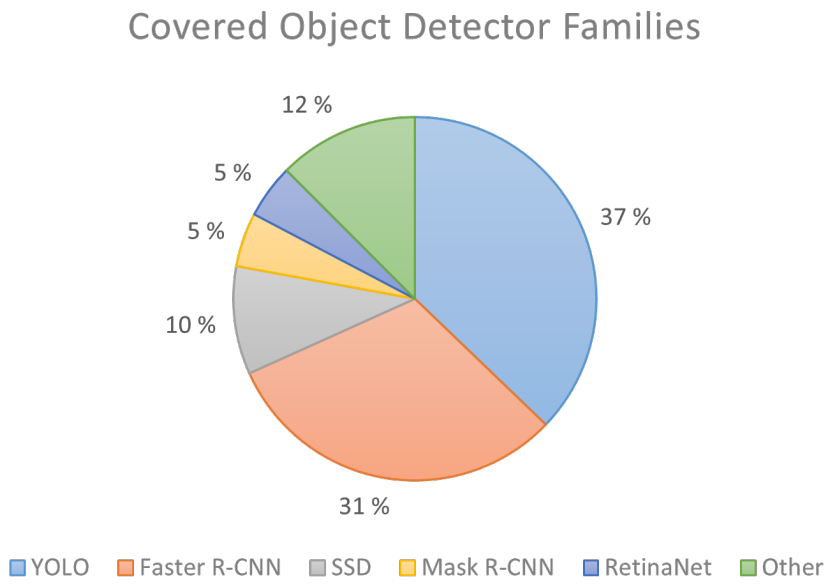

Figure 4.5: The distribution of the covered object detector families.

physically realizable adversarial patches to the next level by aiming to be effective from all viewing angles. The initial testing of the adversarial camouflage attacks is usually performed in simulation before being carried out in the physical world. This testing order is why most adversarial camouflages are also tested in the digital domain, as seen in Figure 4.4.

The other attack methods not covered by the three main methods are about equally covered in terms of experimentation in the two domains. As seen in Table 4.1, these attacks cover attack methods such as adversarial wearables, adversarial projection, adversarial sign, adversarial sticker, and adversarial environment. These attack methods, and most others not mentioned, aim to attack specific use cases in the physical world. Therefore, the experiments of these attacks are commonly performed in the physical world, with initial testing done digitally as with adversarial camouflages.

Covered Object Detector Families

All the adversarial attacks and defenses were classified based on the object detectors used in their experiments. This classification was done to gain insight into which object detectors were frequently used in research. Only the detector types, one-stage detectors, two-stage detectors, and transformers, were documented in Table 4.1 and Table 4.2.

However, specific detector families were also noted during the review process. This information can be helpful for practitioners when deciding what detector to use. As seen in Figure 4.5, the YOLO [10] family of detectors is the most commonly used in

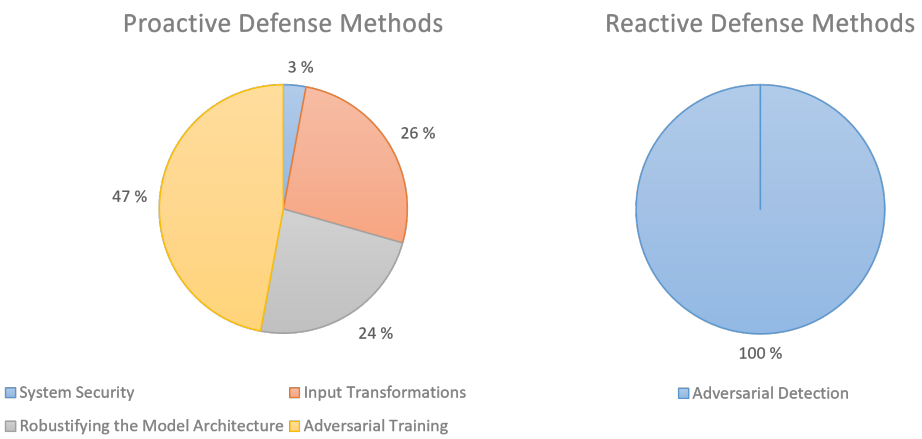

Figure 4.6: The distribution of the covered defense methods against adversarial attacks.

research, closely followed by Faster R-CNN [9]. YOLO being a one-stage detector, and Faster R-CNN being a two-stage detector, front each their detector type in research on adversarial attacks and defenses in object detection. Another one-stage detector family, SSD [11], is the third most covered, followed by a tie between the two-stage detector Mask R-CNN [174] and the one-stage detector RetinaNet [175].

The remaining 12% of detectors are covered less than five times each and are therefore collected in the "other" category in Figure 4.5.

Grouping the Adversarial Defense Methods

The higher level categories, proactive- and reactive defenses, describe how the defense method approaches adversarial attacks. To categorize the defense methods as proactive- and reactive defenses were inspired by the survey on adversarial attacks for object detection in autonomous vehicles by Amirkhani, Karimi, and Banitalebi-Dehkordi [19]. The proactive methods aim to neutralize attacks without the system being aware of them, while the reactive defenses act against an attack once it happens. Figure 4.6 shows the defense methods covered in this systematic literature review.

The proactive methods covered in this systematic literature review are input transformations, robustifying the model architecture, adversarial training, and system security. Adversarial training is a common term within the field, while the terms "input transformations" and "robustifying the model architecture" were introduced specifically for this thesis. The system security classification was used for defense methods more in line with traditional security than adversarial defenses. The following list describes the newly introduced categories for defense methods.

Input transformations is an umbrella term for defense methods that modify the input data for the object detector in some predefined way. They aim to reduce the effect of adversarial attacks by applying various transformations to the input data.

Robustifying the model architecture is an umbrella term for the defense methods that aim to create more robust object detectors by improving their model architecture.

Adversarial training is the most popular proactive defense method against adversarial defenses, as seen in Figure 4.6. Adversarial training counts for 47% of the proactive defenses, while input transformations and techniques for robustifying the model architecture are almost equally covered at 26% and 24%, respectively. The least covered proactive defense method against adversarial attacks is classical system security, with only one article included.

The only reactive method covered in this review is adversarial detection. Adversarial detection methods aim to detect adversarial attacks at runtime. The procedure for what happens when an attack is detected varies between adversarial detection methods, but they often include an attempted removal of the adversarial attack by data modifications. Adversarial detection is the second most common defense method, following adversarial training, as seen in Table 4.2.

Defense Methods' Targeted Attack Methods

The defense methods covered in this review got classified by which attack method they aim to protect against. This information is valuable when defending object detectors against different methods of adversarial attacks. As seen in Figure 4.7, all defenses were targeting either adversarial examples or adversarial patches, the two most widely used attack methods shown in Figure 4.1.

However, this systematic literature review recorded no defense methods against adversarial camouflages. This lack of defense is worrying as adversarial camouflages are potentially the most dangerous adversarial attack in real-world scenarios like unmanned military vehicles. The research community and users of detectors in safety-critical systems must therefore develop viable defenses against adversarial camouflages to ensure the safety and viability of their systems.

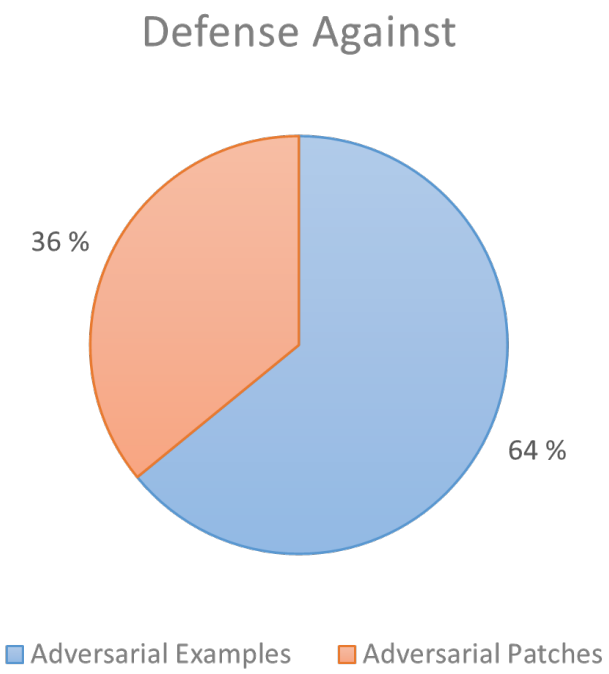

Figure 4.7: The defense methods' targeted attack methods.

4.2 Ensuring the Safe Deployment of AI Against Adversarial Attacks in Unmanned Systems

This section presents the results of Goal-02 of this thesis: to derive a system architecture and requirements for the safe deployment of AI in an unmanned military setting. Before getting there, the threat model is described. The threat model is based on the results from the systematic literature review on adversarial attacks and defenses in object detection, which address Goal-01 of the thesis. Following the threat model are the stakeholder needs and requirements used for defining the system. Then the developed AI pipeline is shown, coupled with its possible security risks against adversarial attacks. Mitigation techniques against these attacks are then presented. Lastly, the designed system architecture with system- and security requirements are shown.

4.2.1 Threat Model

The threat model described in this section is based on the comprehensive systematic literature review on adversarial attacks and defenses in object detection presented in Section 4.1. The domain of the attack, the attacker's knowledge about the target model, and the different attack methods are considered in the threat model.

The domain of the adversarial attacks plays a critical role when considering their threat against the use case of unmanned military vehicles, as described in Section 4.1.2. Digital adversarial attacks need access to the system to be effective, while physical adversarial attacks are realizable without system access. The threat analysis will be based on these domains as they require different defenses.

The attackers' knowledge of the target object detection model will be considered when evaluating the threat level of the system, including white box and black box attacks.

The attack methods and their applicability in the use case of unmanned military vehicles will also be taken into consideration. The threat evaluation will cover the three most frequent attack methods from the systematic literature review on adversarial attacks and defenses in object detection: adversarial- examples, patches, and camouflages. Together they cover 82% of the attacks, and the remaining 18% mainly cover special use cases not as relevant to the use case of unmanned military vehicles.

4.2.2 Stakeholder Needs

This thesis aimed to design a secure AI-based unmanned system against adversarial attacks in the military setting. With the military use case comes strict security requirements that result in limitations that affect how the AI can operate in the system. Communication and network bandwidth limitations are such examples, limiting the amount of data that can be sent between systems. Therefore, such limitations and strict confidentiality, integrity, and availability requirements had to be accounted for during the design process.

Another major limitation of the military use case is the lack of publicly available datasets for training machine learning models. Therefore, the system needed solutions accounting for the need for more data for training the models. Both in terms of initial model training and model improvement during the life cycle of the system.

It was assumed that similar systems to the one designed for this thesis already exist, therefore keeping the integration of AI in focus. The following needs were defined to deploy AI in unmanned military systems safely:

Need-01 The concept for a general system capable of using AI on the edge in an unmanned military setting where the limitations to network bandwidth and computing power are considered.

Need-02 The functionality for generating new training data on the fly during missions to be used for training new and better AI models in the future.

Need-03 Protection of the integrity of the system from adversaries by inhibiting or reducing their possibilities to manipulate information (e.g., training data) in the system.

Need-04 Protection of the availability of the system from adversaries by inhibiting or reducing their possibilities to make information (e.g., sensor data) from the system unavailable.

Need-05 Protection of the confidentiality of the system from adversaries by inhibiting or reducing their possibilities to extract information (e.g., the AI model) from the system.

4.2.3 Stakeholder Requirements

The following stakeholder requirements presented in this section are considered extraordinarily important and are based on the above needs.

One crucial part of safely deploying AI in an unmanned military vehicle is to ensure system reliability. People do not trust AI to make the correct prediction every time in safety-critical systems, which is crucial in such scenarios. Moreover, there is no guarantee that the AI will make correct predictions with 100% accuracy. Therefore, manned and unmanned systems using AI require functionality to deal with wrong predictions. This functionality is reflected in Req-01 and Req-02 below.

As adversarial attacks pose a threat to object detection, the system requires functionality to shut down AI-related processes in cases of adversarial attacks. This functionality is reflected in Req-03 below and is required as wrong predictions can lead to undesired outcomes.

It was a priority to limit the possibility that the AI models and relevant information could land in the hands of adversaries. This security was critical as white box attacks are way more effective at attacking detectors than their black box counterparts. Therefore, Req-04 was decided to prevent unmanned vehicles from using AI trained on classified information, as unmanned vehicles are more easily lost than manned vehicles. Then, if an unmanned vehicle and its AI model get lost, adversaries can not white box attack the models trained on classified information running on the more secure manned vehicles.

The entire system needs to run smoothly and uninterrupted to perform at an operational level. Not only must the system run smoothly, but the AI use case of this thesis, object detection, needs to run effectively in real-time to be able to

detect potential threats in time. Therefore, some requirements were aimed at the hardware to be used by the systems. Req-05 stated that the AI must not slow down other subsystems sharing the same resources, while Req-06 ensures that the object detection runs at an acceptable speed.

Every prediction made by the AI must be logged for debugging and model training purposes - reflected in Req-07. It was derived from the stakeholder need to generate new training data on the fly during missions to be used to train new and better future AI models - Need-02.

The following list defines the stakeholder requirements:

Req-01 The unmanned system must have the functionality to hand over control to a human operator at a remote location.

Req-02 An operator must be able to monitor and, if needed, override or revert decisions made by the AI.

Req-03 If the AI functionality is under attack or a previous attack has been discovered, the system must support the disconnection of the AI from the rest of the system. If the AI is to be disconnected, and the AI is essential for the system to be functional, the system must automatically request manual control.

Req-04 AI models trained on classified data must not be stored locally on unmanned devices.

Req-05 The AI shall not slow down other subsystems in conjunction with computing power.

Req-06 The object detector must run in at least 10 frames per second.

Req-07 Every AI prediction must be logged.

4.2.4 System Architecture – More than Just Unmanned Vehicles

The system architecture proposed in this thesis shows how AI can be integrated into unmanned military vehicles. The system architecture describes not only the unmanned vehicles themselves but also the surrounding enabling systems. Therefore, the proposed system architecture also includes the manned military vehicles assisted by the unmanned vehicles and the back-end server where the AI models will be trained before deployment. Why this was done will be further described below.

Back-End Server

The reason for including a back-end server was the limitations to computing power on the unmanned vehicles, as described in stakeholder Need-01. It could not be assumed that any AI model could be trained using local hardware on unmanned vehicles. Model training was therefore off-loaded to a back-end server equipped with appropriate hardware. Including a back-end server is also beneficial for developing AI models, as experts with access to the server can develop AI models for different platforms and uses. The same server should also be the central storage of all the training data generated over time, as requested by stakeholder Need-02. By that, AI experts can choose and combine relevant data from different systems to train new and improved models for other use cases.

Manned Vehicles

The interactions between the manned and unmanned vehicles are essential to the system. As shown in Figure 1.2 on Page 4 that explained the use case example of this thesis, the objective of the unmanned vehicles is to assist the manned vehicles, e.g., by reconnaissance of the surroundings. The manned vehicles were essential in meeting several stakeholder requirements, including Req-01, Req-02, Req-03, and Req-04. Requirements 1-3 require a human operator to be able to monitor, revert decisions, or take control of unmanned vehicles. As stated in Need-01, wireless communication in the military is limited. Therefore, such tasks needed to be given to nearby human operators, preferably in vehicles already involved with the unmanned vehicles.

Stakeholder Req-04 states that unmanned vehicles can not use AI models trained on classified data due to the high risk of getting astray. As a solution, the connected manned vehicles can run the models trained on classified information, while the unmanned vehicles run unclassified alternatives. This scenario can be better explained with an example: The manned vehicle commands the unmanned vehicle to reconnaissance a nearby location. The unmanned vehicle runs an object detector trained to detect and tell the difference between military and civilian vehicles. However, the unmanned vehicle can not distinguish between different types of military vehicles. When detecting a military vehicle, the unmanned vehicle can send the prediction with one or several video frames to the closest manned vehicle. As the manned vehicle is installed with a more advanced AI trained on classified information, it can decide what type of military vehicle is detected and share that information with the entire team.

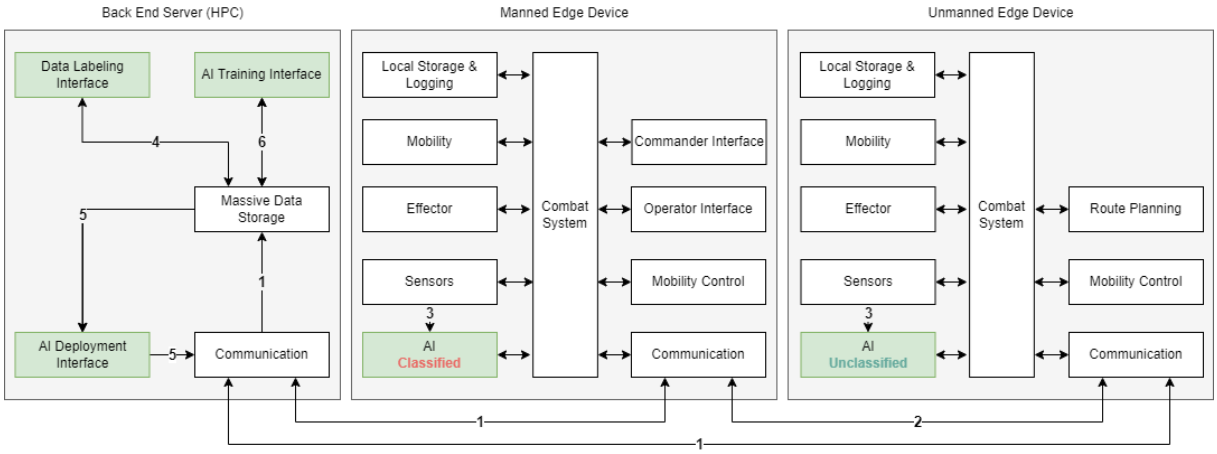

Figure 4.8: The logical system architecture of the proposed system, including the server side and both manned- and unmanned edge devices. The numbered connection will transmit the following data: **1:** Trained AI models (server $\rightarrow$ edge), sensor data coupled with AI predictions and human interventions (edge $\rightarrow$ server), **2:** AI predictions and sensor data, **3:** Sensor data (e.g., video feed) fed directly to the AI model, **4:** Raw, unlabeled data (storage $\rightarrow$ interface), labeled data (interface $\rightarrow$ storage), **5:** Trained AI models, and **6:** Training data and trained AI models (storage $\rightarrow$ interface), Trained AI models (interface $\rightarrow$ storage).

4.2.5 Logical System Architecture

The proposed system architecture is shown in two different figures: one being the logical system architecture and the other being the physical system architecture. The logical system architecture describes how the various components of the system interact with each other, while the physical system architecture also includes the physical placement of components in the system. This section describes the proposed logical system architecture, focusing on the AI implementation.

Figure 4.8 shows the proposed logical system architecture. It is divided into three different types of devices: a back-end server/high-performance computer (HPC), manned edge devices, and unmanned edge devices. This thesis focuses on the AI implementation in the system, represented by the green blocks in the diagram, with the accompanying numbered connections. The caption of Figure 4.8 describes what is to be transmitted in the connections to the AI-related parts of the system.

Back-End Server

The back-end server is responsible for training and deploying AI models for manned and unmanned vehicles, using AI- training and deployment interfaces. The back-end server is also responsible for storing all the data collected through the life-cycle of the

system to support stakeholder Need-02. This data is mainly sensor data from manned- and unmanned vehicles, coupled with AI predictions and human interactions. This data can then be used to train new and improved AI models for future missions. A data labeling interface is implemented to support the process of turning the raw data into usable training data.

Manned and Unmanned Edge Devices

Both the manned- and unmanned vehicles shall use a combat system that connects the different parts of the systems and will control the flow of information within the system. All information flows through the combat system with only one exception: the direct flow of sensor data from the sensors to the AI to limit the delay, to meet stakeholder Req-06 of real-time performance. The combat system shall also be responsible for the situational awareness of the system by sharing information between manned- and unmanned devices when needed.

As seen in Figure 4.8, the system is designed with modularity in mind, with everything being connected to the combat system. This modularity makes the system scalable by enabling the removal- and addition of new parts. This design choice was also made with system security in mind. If the AI gets exposed to adversarial attacks, the system can isolate the AI by terminating its connection to the combat system if no other mitigation technique is sufficient. This functionality directly addresses stakeholder Req-03. If the system depends on the terminated AI, the combat system is responsible for requesting manual control.

4.2.6 Physical System Architecture

The physical system architecture takes the modeling further and includes the physical components of the system. It comprises the hardware, software, interfaces, and other relevant components and their connections. Figure 4.9 shows the proposed physical system architecture. It extends the system presented in Figure 4.8 by including the hardware components of the system. Again, the focus was on the AI functionality of the system. The caption of Figure 4.9 describes what AI-related data is to be transmitted in the cross-platform connections of the system.

Manned and Unmanned Edge Devices

Stakeholder Req-05 and Req-06 required the AI not to slow down other subsystems and for object detection to be able to run in real-time, respectively. Therefore, the system was designed to support graphical processing units (GPUs) in all edge devices, including manned and unmanned vehicles. GPUs are exceptionally good at performing

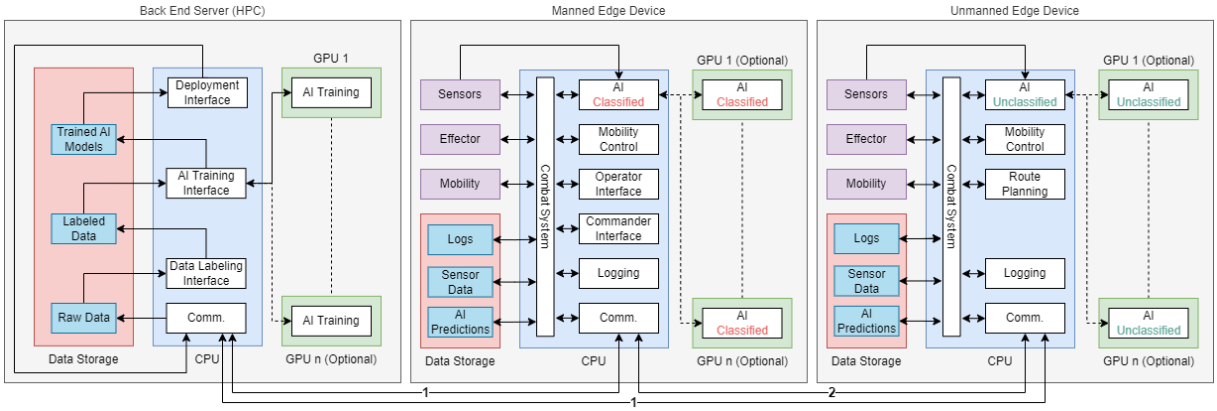

Figure 4.9: The physical system architecture of the proposed system, including the server side and both manned- and unmanned edge devices. The numbered cross-platform connections will transmit the following data: **1:** Trained AI models (server $\rightarrow$ edge), sensor data coupled with AI predictions and human interventions (edge $\rightarrow$ server), and **2:** AI predictions and sensor data.

computations on AI-related tasks and can therefore unload the central processing unit (CPU). By offloading the AI's calculation to the GPU, the AI will not slow down the CPU and negatively affect other subsystems.

However, one can only assume that some edge devices in question can have one or several GPUs installed due to cost, weight limitations, and system support. Due to their weight restrictions, this claim is especially valid for UAVs, a prime example of our use case presented in Figure 1.2 on Page 4. Therefore, including GPUs on the edge platforms was designed to be optional, as visualized by the dotted lines between the AI process on the CPU and the GPUs in Figure 4.9. Another reason for keeping the use of GPUs optional is that the computing power needed by the AI is highly dependent on the model's complexity and use case. This design choice is also supported by stakeholder Need-01, which requests a general system applicable in several scenarios.

Back-End Server

The back-end server responsible for AI model training will have at least one GPU. AI training is computationally demanding compared to the AI inference run on edge devices. The number of GPUs to be installed depends on the complexity and size of the models to be trained, and the type of GPU to be used. Therefore the additional GPUs shown in Figure 4.9 were designed to be optional.

Figure 4.9 shows how the unlabeled raw data enters the back-end server and the process of turning it into new and improved AI models ready for deployment to the edge platforms.

4.2.7 AI Pipeline / Functional System Architecture

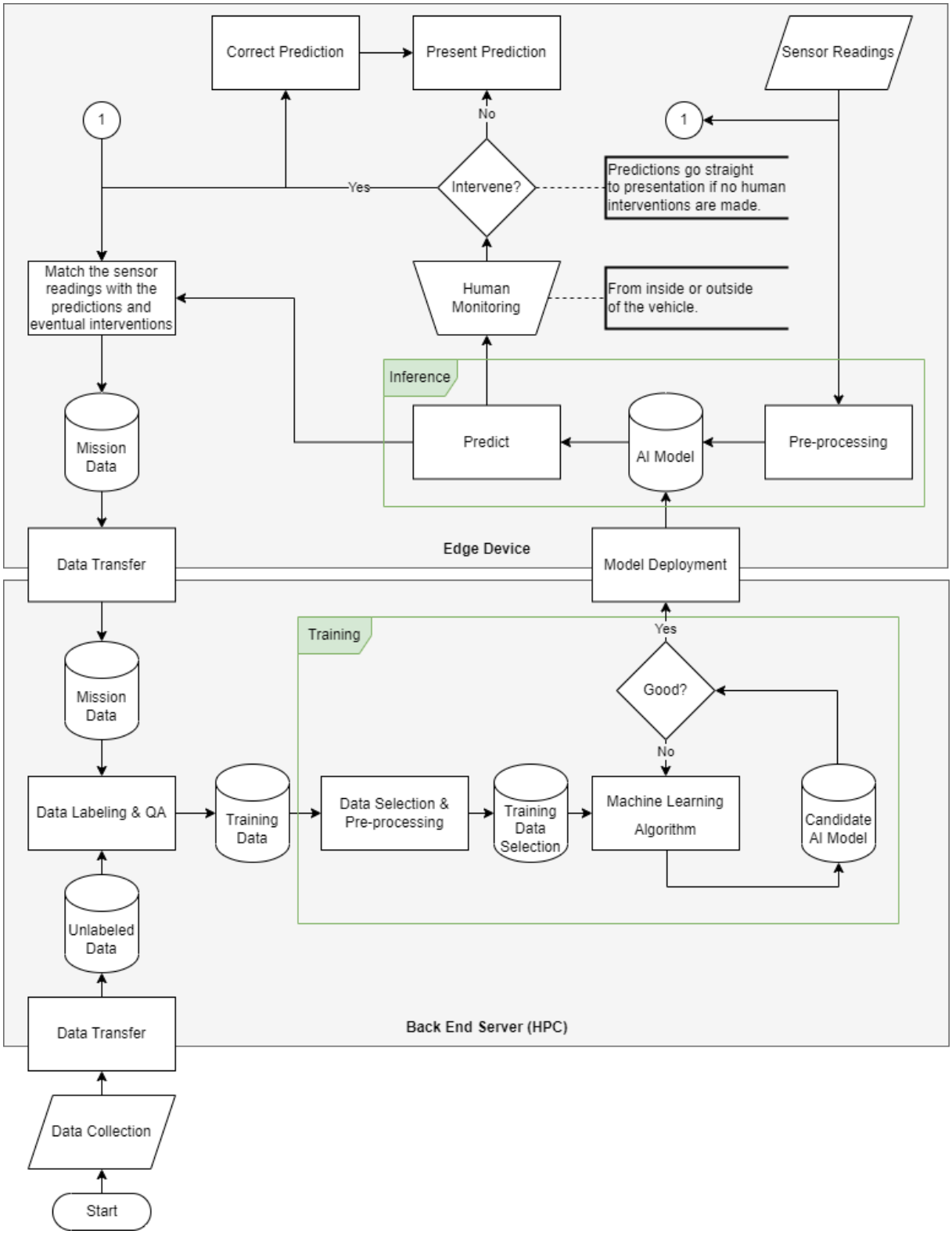

Figure 4.10: The workflow of the AI-related processes in the system. Read Section 4.2.7 for further explanations.

This section aims to describe the AI pipeline of the system. In contrast to the logical- and physical system architectures presented in Figure 4.8 and Figure 4.9, the AI pipeline proposed in Figure 4.10 describes the AI-related tasks in more detail. When reading this section, it is advised to follow along in Figure 4.10.

Initial Data Collection and Labeling

The first process in the AI pipeline is to collect relevant data that can be used to train the initial model, assuming that the system user does not have any model for the desired use. After collecting the data, it must be transferred to the back-end server, where it will be stored, labeled, and used for training AI models. The collected unlabeled data shall be stored in a separate database before it is used in the data labeling and quality assurance (QA) process. The data labeling and QA process are accessed by operators using the data labeling interface presented in Figure 4.8 and Figure 4.9. After the data is labeled, it will be stored as training data in a separate database to separate it from the unlabeled data.

Training

The training data can then be used for training AI models in the training section encapsulated by the green frame labeled "Training" in Figure 4.10. The data will first be selected from the training data database and pre-processed before being used to train the machine learning algorithm. The loop within the training section in Figure 4.10 represents the training and evaluation of models until a model with the desired accuracy is achieved. The training process will be accessible to operators using the AI training interface presented in Figure 4.8 and Figure 4.9.

Deployment and Inference

The final model will then be deployed to the edge devices, where it will be used for inference in the use case of object detection. The model will receive pre-processed sensor data to run inference on, and the model will output a prediction. The prediction of the AI model will be used to perform some action depending on the system and use case, e.g., to mark possible dangers on a map.

Human Interventions and Data Collection

As stated by stakeholder Req-02, operators inside or outside of the edge device must be able to monitor, evaluate and intervene with the AI predictions. Figure 4.10 also accounts for this functionality, as seen in the steps following AI inference. As stated by

stakeholder Need-02, new training data shall be generated on the fly during missions to be used for training new and improved AI models for future missions. This need is accounted for by saving sensor data coupled with the AI's predictions to local data storage - stakeholder Req-07. Another vital feature is that every intervention made by human operators will be saved as critical training data to improve future models.

Training New and Improved Models

When a mission is ended, the collected data will be transferred from the edge devices to the back-end server for labeling, QA, and model training. The collected data from the time before human interventions will be used to improve current models and make them more robust. The collected data also brings a new range of cases that result in a broader range of cases in the training data database. This broader range of cases will result in more general and robust models for future missions. The newly collected data can also be used to train entirely new AI models for different use cases not previously possible due to the need for more data. Then the loop continues, resulting in more accurate and robust models for a broader range of use cases over time.

4.2.8 Acknowledging and Mitigating the Threats of Adversarial Attacks

The following sections show the threat analysis of adversarial attacks against the proposed system. First, the threat of digital adversarial attacks is presented, followed by the recommended defense methods to mitigate the risk of digital adversarial attacks. The threat analysis of physical adversarial attacks is then presented, followed by adversarial defense recommendations to mitigate those threats. The adversarial attack threats are based on the results from the literature review presented in Section 4.1 and are shown in the context of the AI pipeline presented in Figure 4.10.

Threat Analysis of Digital Adversarial Attacks

With object detection in unmanned military vehicles operating in the physical world, digital adversarial attacks have limited effect as they can not alter the physical world. However, digital adversarial attacks can threaten AI functionality if adversaries somehow gain access to the system. With access to the system, adversaries can inject digital adversarial attacks into the AI pipeline to attack its functionality. Therefore, this scenario of an attacker gaining access to the system to inject digital adversarial attacks plays a vital role in the threat analysis presented in this section.

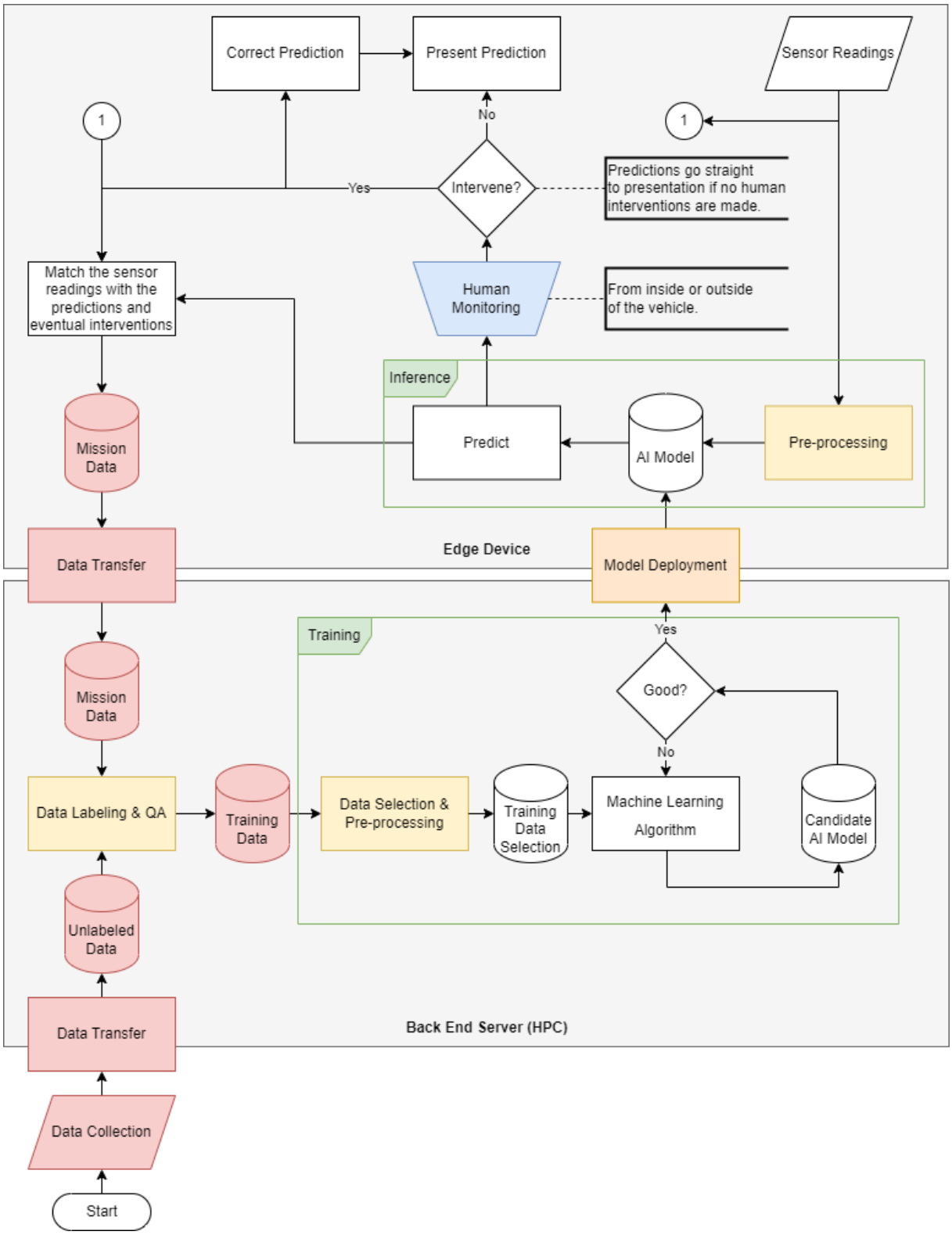

Figure 4.11: Threat analysis of digital adversarial attacks. The colors represent different threats against the AI functionality: **Red:** Data poisoning, **Yellow:** Malware, **Orange:** Model extraction and model swapping, **Blue:** Unauthorized access. Read Section 4.2.8 for further explanations.

Data Poisoning Data poisoning is regarded as digital adversarial attacks injected into the system in this threat analysis. These digital attacks can be adversarial examples of imperceptible image changes that cause misclassification or mislocalization in the object detector. The red blocks in Figure 4.11 represent the parts of the AI pipeline prone to data poisoning attacks.

Data poisoning attacks can happen at different locations in the system, including during data- collection, transfers, and storage. Data poisoning can occur during data collection if an adversary plants adversarial examples in the data source, e.g., in a public image database. Adversarial examples can also be injected into the system during the data transferring stage, given that the adversary gains access to that process. Given system access, adversarial examples can also be injected directly into the databases. The injected adversarial examples will then be used during model training and reduce the accuracy of the trained model. If the adversarial attacks are targeted, they can even impact the learning of objects by tricking the model into thinking an object is entirely different from what it is.

Malware The yellow blocks in Figure 4.11 represent the parts of the AI pipeline that can be affected by malware attacks. Malware in the context of this threat analysis is adversarial attack methods integrated with the system processes to create adversarial examples within the system. The adversarial attack methods can be placed in various processes, including data labeling and quality assurance, data selection and pre-processing before model training, and sensor data pre-processing before inference. While remaining undetected, this attack can continuously create adversarial examples and poison the data over an extended period compared to manual data poisoning.

Model Extraction and Model Swapping The orange block in Figure 4.11 shows which process where model extraction and model swapping can happen. This threat is considered the most crucial as trained AI models can get astray, enabling adversaries to white box attack the models. The same model knowledge can be achieved if an adversary swaps the model with one of their own. Such white box knowledge can result in devastating attacks rendering the AI dangerous to use in operations. Therefore, the model deployment process must be secure, and the personnel handling the model transfer must have good intentions.

Model swapping must not only be used for white box attacks, as the model trained by the adversary can be trained to misclassify or mislocalize vital objects. E.g., an opponent vehicle gets classified as friendly.

Unauthorized Access The blue block in Figure 4.11 shows where unauthorized access to the AI live monitoring can pose a risk. If an adversary gains access

to this feature, interventions can be made by the adversary to trigger events to be stored for future model training. E.g., the AI predicts an opposing vehicle to be an opponent, but the adversary intervenes with the prediction and changes the prediction of the vehicle to be friendly. Then the misclassified event will be used for later model training if not detected during quality assurance.

Mitigation of Digital Adversarial Attacks

This section aims to mitigate the threats of digital adversarial attacks presented in Section 4.2.8. Since most digital attacks require system access, traditional system security mechanisms can mitigate those scenarios requiring system access. However, adversarial defense methods will also be implemented to reduce the threat of adversarial attacks if an adversary gains access to the system regardless.

System Security Stakeholder Need-3, Need-4, and Need-5 describe that the integrity, availability, and confidentiality of the system must be protected from adversaries. Ensuring that these needs are met will, in turn, protect the system against most of the digital adversarial attack threats presented in Section 4.2.8. As discussed during the workshop with KDA, it is assumed that the military vehicles implementing the system proposed in this thesis must follow the Norwegian law of national security (Sikkerhetsloven) [176] due to their handling of classified information (Sikkerhetsloven, 2019, § 9–1). The Norwegian National Security Authority (NSM) [2] has derived guidelines [3] (exempt from public disclosure) to ensure that the law of national security is followed, and it is assumed that the proposed system will follow those guidelines to protect the integrity, availability, and confidentiality of the system against adversaries. With this comes protection from digital adversarial attacks.

These security mechanisms will also protect the AI against white box attacks, as the chance of AI models getting astray and being accessed by adversaries gets significantly reduced. By that, the overall threat of adversarial attacks was considered considerably mitigated as white box attacks were more widespread and easier to use effectively than their black box counterparts.

Caution During Data Collection Given the above assumption to system security, all threats presented in Section 4.2.8 are mitigated, except for adversarial examples in the data collection step. Therefore, the initial data collection must be met with caution if the data is collected from publicly available databases, as adversaries can plant adversarial examples that can sabotage the model training.

[2] The Norwegian National Security Authority (NSM): https://nsm.no/

[3] Referred to as "G-guides" for the remaining of the thesis.

It is therefore recommended only to use data collected from reliable sources. However, this thesis did not investigate what is considered reliable sources and is left for future work on the system.

With the system secured, it is assumed that the threat of digital adversarial attacks and physical white box attacks have been considerably mitigated. However, adversarial defense methods can be implemented to protect the object detector against digital adversarial attacks if an attack happens regardless. The recommended adversarial defense methods from Table 4.1 are as follows:

Adversarial Training The system must implement adversarial training methods to protect the AI against digital adversarial- examples and patches. Adversarial training uses samples of adversarial attacks during model training to make the model robust against these attacks. The limitation of this defense method is that the model will only be robust against the attacks used during training. Therefore, the adversarial training schedule must be updated to include new attack methods once they emerge.

Input Transformations The system must also implement input transformations as steps in the pre-processing before model training and inference. Such input transformations try to reduce the impact of adversarial attacks by performing actions such as smoothing the input data. This defense method is a simple and effective way of reducing the risk of digital adversarial examples.

Robust Model Selection Selecting robust models should also be considered to increase the general robustness of the model.

Threat Analysis of Physical Adversarial Attacks

It is now assumed that the threats of digital adversarial attacks against the system are mitigated, which leaves the physical adversarial attacks. Physical adversarial attacks in our use case of object detection manipulate the real world instead of the digital one to trick object detectors into making wrong predictions. Such attacks do not require system access and are therefore not protected against by the defenses proposed in Section 4.2.8. Therefore, this section presents the remaining risks of physical adversarial attacks against the proposed system.

Physical Adversarial Attacks The red blocks in Figure 4.12 show the system parts prone to physical adversarial attacks like physical adversarial patches or physical adversarial camouflages. These system parts are the only input nodes to the system, making them prone to physical adversarial attacks. These attacks can

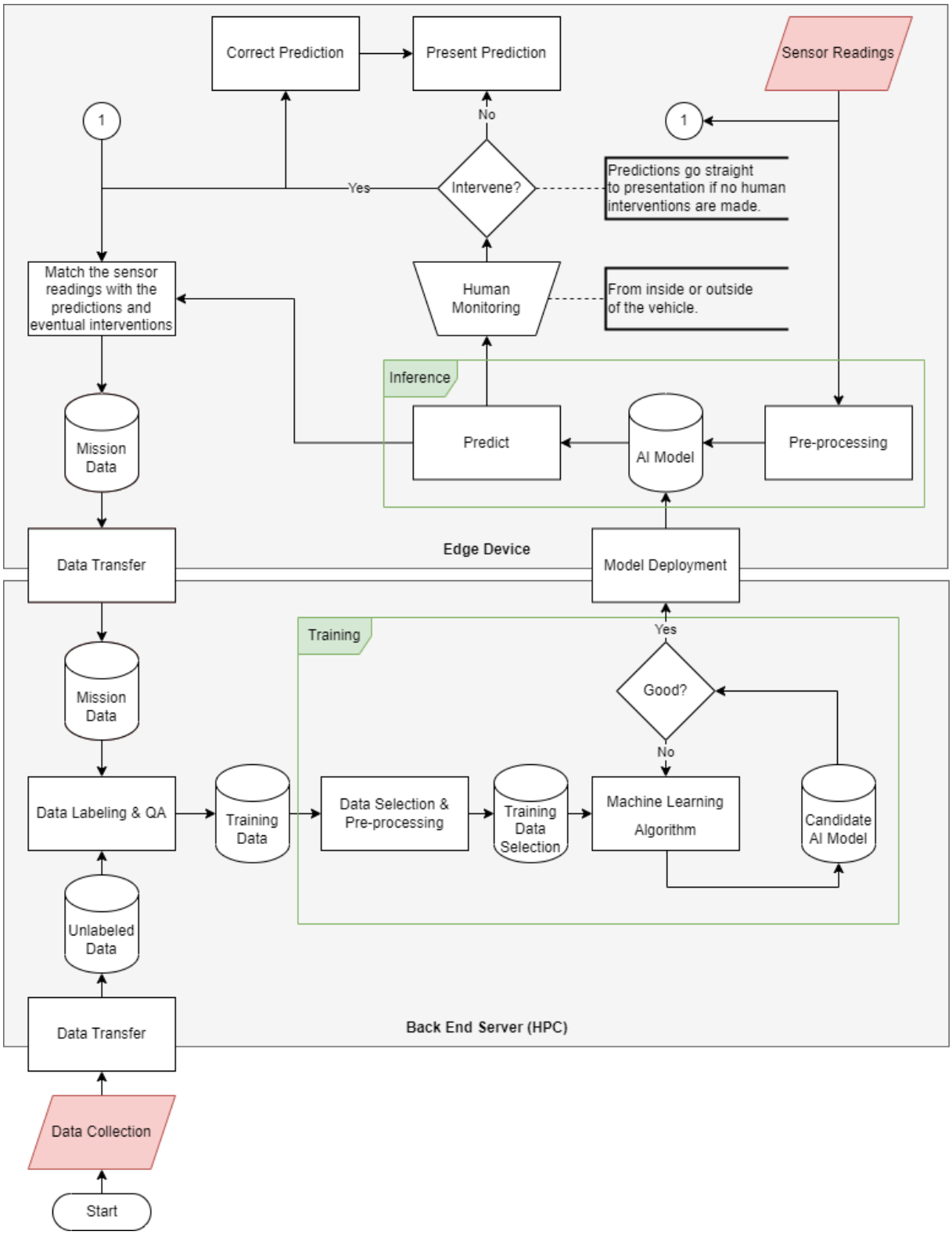

Figure 4.12: Threat analysis of physical adversarial attacks. The colors represent different threats against the AI functionality: **Red:** Physical Adversarial Attacks. Read Section 4.2.8 for further explanations.

happen during the initial data collection and sensor readings. An adversary can use physical adversarial attacks to poison the data during the initial data collection. However, this can be avoided by only using reliable sources as described in Section 4.2.8. That leaves the physical adversarial attacks against the system during operations. E.g., an adversary covers their vehicles in adversarial camouflages to be misclassified as friendly by the object detector. Such attacks are critical as traditional security mechanisms do not mitigate them and need dedicated adversarial defenses.

Mitigation of Physical Adversarial Attacks

Adversarial defense methods can be implemented to robustify and secure AI against physical adversarial attacks. The recommended adversarial defense methods against physical adversarial attacks from Table 4.1 on Page 27 are as follows:

Adversarial Detection The system must implement adversarial detection to be able to detect physical adversarial attacks when they happen in real-time. This defense method has the same limitation as adversarial training by only learning to detect the attack methods it was trained to detect. Therefore, other defense methods must also be included to account for this limitation. As adversarial detection is a kind of object detector, it can be implemented as the general AI as shown in the system architectures described in Figure 4.8 and Figure 4.9.

Adversarial Training The system must implement adversarial training methods to protect the AI against physical adversarial examples. This defense method was also mentioned as a defense against digital adversarial attacks. However, the adversarial training schedule must also include instances of physical adversarial attacks, similar to the ones provided in the APRICOT dataset [177]. Again, the adversarial training schedule must be updated to include new attack methods once they emerge.

Robust Model Selection Selecting robust models should also be considered to increase the general robustness of the model. This defense method was also mentioned as a defense against digital adversarial attacks but is also relevant for mitigating physical adversarial attacks.

4.2.9 System and Security Requirements

This section describes the system- and security requirements to ensure the safe deployment of AI in an unmanned military setting. These requirements were derived based on the threat analysis of the proposed system architecture with the goal of

meeting the stakeholder needs and requirements. As the subsystems serve different purposes in the overall system and require different security classification levels, different requirements are derived for the individual subsystems. The respective subsystems include the back-end server, manned edge devices, and unmanned edge devices.

Back-End Server

Req-01-01 Follow NSMs G-guides[4] to protect the confidentiality, integrity, and availability of the system from adversaries as the system will be handling classified information.

Req-01-02 Adversarial training must be included as a countermeasure to adversarial attacks. See Table 4.2 for adversarial training methods.

Req-01-03 Input transformations must be made to the training data to counter adversarial attacks. See Table 4.2 for input transformation methods.

Req-01-04 Robust model alternatives must be considered when deciding which object detector to train. See Table 4.2 for methods of robustifying the model architecture.

Req-01-05 A separate detector must be trained for adversarial detection. See Table 4.2 for adversarial detection methods.

Req-01-06 The adversarial detector must be updated when new attack methods arise.

Req-01-07 AI models must only be trained on data received from reliable sources.

Manned Edge Device

Req-02-01 Follow NSMs G-guides[5] to protect the confidentiality, integrity, and availability of the system from adversaries as the system will be handling classified information.

Req-02-02 Input transformations must be made to the sensor data before inference to counter adversarial attacks. See Table 4.2 for input transformation methods.

[4] Guidelines designed to ensure that systems handling or interacting with systems handling classified information comply with the law of national security [176]. The guidelines was derived by NSM: https://nsm.no/regelverk-og-hjelp/veiledere-og-handboker-til-sikkerhetsloven/. The referred documents are exempt from public disclosure.

[5] See Footnote 4.

Req-02-03 Adversarial detection must be included as a countermeasure to adversarial attacks. It shall run in parallel with the main object detector. See Table 4.2 for adversarial detection methods.

Unmanned Edge Device

Req-03-01 Follow NSMs G-guides[6] to protect the confidentiality, integrity, and availability of the system from adversaries as the system will be interacting with systems handling classified information.

Req-03-02 If the unmanned edge device is an UAV, the AI must run on a CPU due to weight limitations.

Req-03-03 Input transformations must be made to the sensor data before inference as a countermeasure to adversarial attacks. See Table 4.2 for input transformation methods.

Req-03-04 Adversarial detection must be included if the system can afford it in terms of computational resources. This decision must be made on an individual basis. See Table 4.2 for adversarial detection methods.

Figure 4.13 below shows the derived system- and security requirements in the context of the AI pipeline. The AI pipeline is the same for both edge devices, but their system- and security requirements differ. Therefore, the requirements for the edge device in Figure 4.13 are marked either “M” or “UM” for manned and unmanned, respectively.

Req-01-01 covers the entire back-end server and was placed in the bottom right corner to represent that it covers the entire subsystem. The same applies to Req-02-01 and Req-03-01 in the context of manned- and unmanned edge devices, placed in the top left of the figure. These requirements ensure that the system follows the guidelines derived by NSM for protecting the confidentiality, integrity, and availability of the system.

While the above requirements indirectly secure the AI using traditional security mechanisms, the remaining requirements directly aim at AI security. System- and security requirements Req-01-02 to Req-01-07 are related to the AI training process and protect the system against adversarial attacks by implementing adversarial training, input transformations, robust model selection, adversarial detection, and by limiting data collection to reliable sources.

The remaining system- and security requirements aim to protect the AI during runtime on edge devices. Req-02-02 and Req-03-03, for manned- and unmanned

[6] See Footnote 4.

edge devices, respectively, ensure that the input to the AI models must be transformed to limit the risk of adversarial attacks. To further protect the AI, Req-02-03 and Req-03-04 ensure the use of adversarial detection for manned- and unmanned edge devices, respectively.

Req-03-02 was not accounted for in Figure 4.13 as the requirement is aimed at the physical placement of AI inference, which is separate from the AI pipeline.

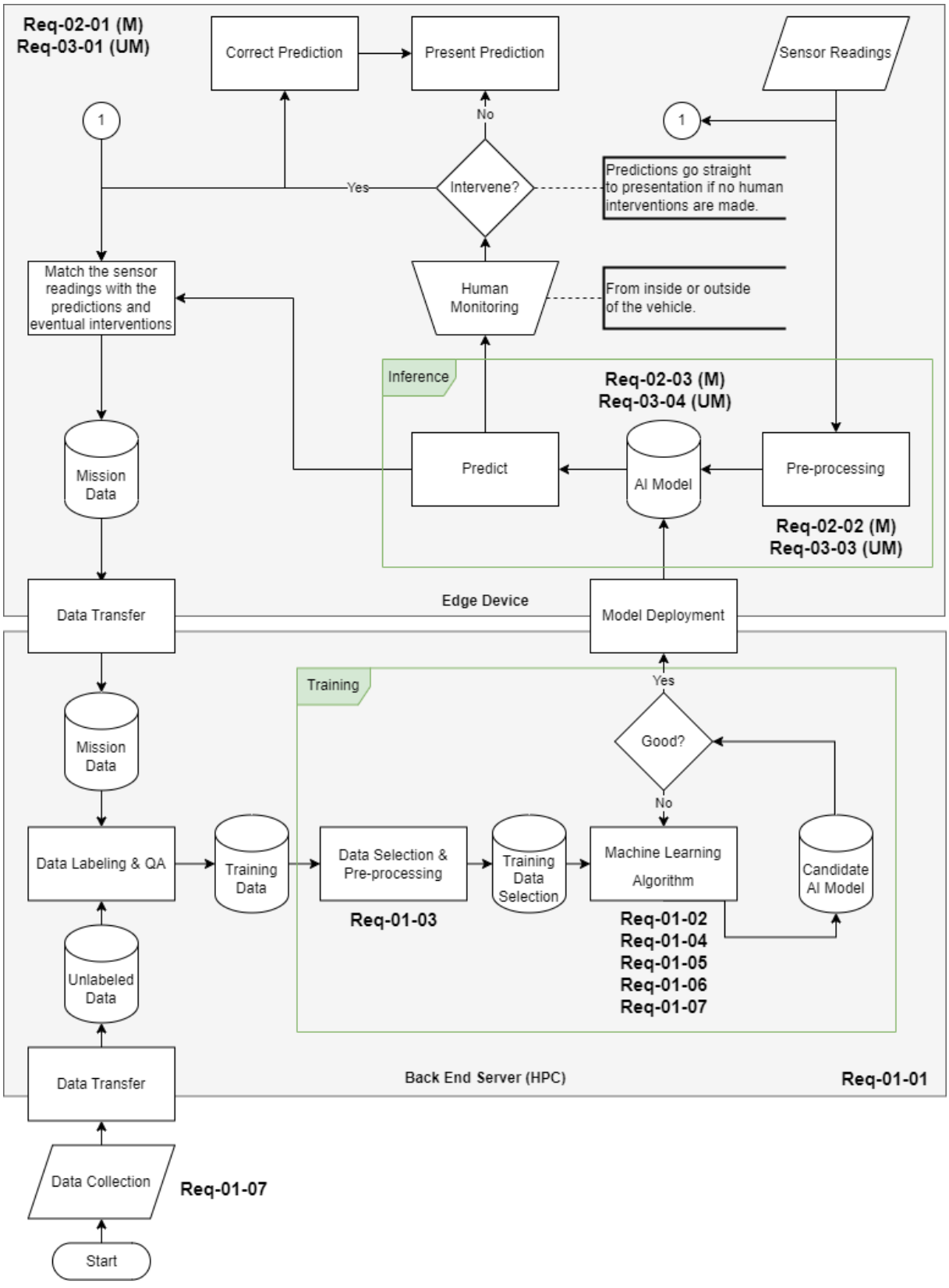

Figure 4.13: The system requirements starting on Page 55, placed in the context of the AI pipeline.

Chapter 5

Discussion

This chapter discusses the results of the thesis, including results from the literature review and the derived unmanned system against adversarial attacks. First, discussions on the current landscape of adversarial attacks and defenses are presented. The discussions include the transferability of attacks between computer vision tasks as a potential threat, the difficulty of creating physically realizable adversarial attacks, and conflicting terminology that needs attention in the research field.

The derived system architecture and requirements for the safe deployment of AI in unmanned military vehicles are also discussed. These discussions include how the stakeholder needs and requirements were met, whether adversarial attacks significantly threaten unmanned military systems, and the dilemma of which object detectors to use based on adversarial attack research.

The ethical considerations of attacking AI in an unmanned military setting are also discussed below, followed by several proposed directions for future research and development on the results of this thesis.

5.1 The Research Field of Adversarial Attacks and Defenses in Object Detection

This section discusses the literature review of this thesis on adversarial attacks and defenses in object detection and some relevant topics related to the research field. First, the strengths and weaknesses of the literature review are discussed. Then a discussion on the current state of physical adversarial attacks is presented, followed by discussions on conflicting terminology in the research field.

5.1.1 Discussion of the Systematic Literature Review

This section discusses the strengths and weaknesses of the systematic literature review on adversarial attacks and defenses in object detection. The systematic literature review has several strengths, including the broad and unbiased search strategy, the strategic classification of papers, and the resulting comprehensive research collection. The systematic literature review also has a few areas for improvement, including further research on the possible cross-task transferability of attacks, including non-English written papers, and providing more detail of the included papers.

Search Strategy

The search strategy covered the entirety of the object detection task, only excluding salient object detection as the task is closer to segmentation than pure object detection. Moreover, the research on adversarial attacks and defenses was kept unbiased by having limited knowledge of the field before doing the literature review. In turn, no attack- or defense methods were favored.

Additionally, papers from all countries were included, except a few papers written in Chinese. These papers could have been included with the use of translation technologies. However, a good amount of similar papers were already included. With that in mind, it was decided to exclude these papers to remove the chance of translation errors.

Classification Strategy and Results

A significant strength of the literature review was the classification strategy of the included papers to facilitate the threat modeling for industry use cases. Such a classification strategy was prioritized from the beginning, as most similar review papers in the field are strictly focused on research rather than the industry.

The resulting tables (Table 4.1, Table 4.2, and Table A.1) can be starting points for further research on the field as they include a comprehensive overview of the current state-of-the-art in adversarial attacks and defenses in object detection. No other review on the same topic matches the comprehensiveness and reproducibility of the systematic literature review of this thesis.

Cross-Task Transferability

A limiting factor of the literature review is the possible risk of cross-task attack transferability from related deep learning tasks like image classification and segmentation.

Image classification and segmentation are similar to object detection, and some attacks targeting these tasks have effectively attacked object detectors [44, 45, 54, 70, 114, 178]. Additionally, more research is done on adversarial attacks against image classification than object detection. Therefore, more research on cross-task transferability is needed, as successful cross-task transfer attacks will expand the range of threats against object detection further than what is covered in this thesis.

The Big Picture vs. Details

Lastly, more detailed explanations of the included papers could have benefitted some readers. However, rather than focusing on the details of a few papers, it was decided to focus on the overview of a more significant number of publications. Such a decision had to be made due to the time restraints of the short master's thesis. However, this decision resulted in a comprehensive literature review of the current state-of-the-art object detection, providing a great starting point for further research.

5.1.2 Physically Realizable Adversarial Attacks

Physical adversarial attacks were perceived as a premature research field needing more attention. Research on adversarial attacks primarily performs experiments in the digital domain rather than real-world scenarios, as shown in Figure 4.3 on Page 33. It is not given that adversarial attacks are viable in the physical world, even though they have successful experiments in the digital domain. For an attack to be effective in the physical world, it must be printed somehow, e.g., by printing on paper. However, the real world has several limitations, including the color differences before and after printing. Lighting is also different in the physical world, and the many possible viewing angles make it additionally difficult. These limitations, among others, make creating physically realizable adversarial attacks difficult and make the current state of physical adversarial attacks a lesser threat against real-world systems.

5.1.3 Conflicting Terminology

Several instances of conflicting terminology were observed when reviewing the young research field of adversarial attacks and defenses in object detection. The words "targeted" and "black box" had different meanings across papers within the same research field. These conflicting terminologies cause confusion for readers, especially those that are new to the field. Following is the discussion of the mentioned terms to guide the research field toward a common use of expressions.

Black Box

The "black box" term is used differently across papers. Some papers define black box attacks as having no access to the target model. In contrast, while modifying the attack, others define that their black box attack methods can query the target model to get feedback on the effectiveness of the attack. In the case of object detection, this feedback can be the predicted classes along with their respective bounding boxes. These different knowledge assumptions need better dividers, given the clear advantage of querying black box attacks over query-free black box attacks.

The different knowledge assumptions of black box attacks make it unfair to compare them under the same term. Therefore, researchers should use separate terms in future research on black box adversarial attacks. However, the black box term used for classifying attacks in this thesis covers both definitions, as the terminology conflict was discovered during the literature review and was unknown in advance.

Targeted Attack

Research on adversarial attacks originates from image classification and has been extended to object detection. With this extension to object detection came the conflict in using the expression "targeted attacks". In image classification, targeted attacks aim to change a classification to a target class. However, this definition can be inconsistent in object detection, with object detection being a multi-task problem. Moreover, object detection classifies and localizes several objects in images instead of classifying a single object in image classification.

Research on adversarial attacks in object detection has been using an additional description of the term: to target a specific class for misclassification or mislocalization. E.g., the attack targets only cars to be misclassified and does not affect the detected pedestrians. This second definition makes sense in object detection, but it does cause conflicts with the original definition. Therefore, the research community should define common terminologies to avoid further misuse and confusion.

This thesis used the original definition of the term, as the second description was discovered during the literature review. However, future research on the field should also include the other definition due to its relevancy in object detection.

5.2 Discussion of the Derived System Architecture and Requirements

This section discusses the derived system architecture and requirements. The discussion includes how well the stakeholder- needs and requirements were met,

how specialists should work together to develop better solutions, the choice of object detectors based on their occurrence in adversarial attack- and defense research, and the current threat of adversarial attack against the designed system.

5.2.1 How the Stakeholder Needs and Requirements were Met

Meeting the stakeholder needs and requirements was essential for successful project delivery. Therefore, this section discusses to which extent the designed system met the stakeholder needs and requirements presented in Section 4.2.2 on Page 39 and Section 4.2.3 on Page 40, respectively. The stakeholder needs and requirements are discussed one after the other in the following subsections.

Stakeholder Need-01 requested a concept for a general system capable of using AI on the edge in an unmanned military setting where network bandwidth and computing power limitations are considered. This stakeholder need was met by designing the logical and physical system architectures (Figure 4.8 on Page 43, and Figure 4.9 on Page 45, respectively) and the AI pipeline (Figure 4.10 on Page 46).

The limitations to network bandwidth were considered by not needing to transmit large amounts of data during operations. E.g., the transmission of collected training data from the edge to the server is done after operations when communication is no longer an issue. The communication between edge devices during missions is also kept to a minimum, only transmitting the AI predictions and their respective video frames.

As for computing power, the edge devices were designed to support GPUs. However, using GPUs on the edge was designed to be optional due to the different hardware limitations on different platforms. If the platforms can not support the installation of a GPU, like smaller UAVs, then the AI must be run on the CPU. The designed system was kept as general as possible to support the different scenarios. Therefore, the job of selecting and training appropriate models for the different platforms was left to the AI trainers of the system.

Moreover, the training of AI models was offloaded to a back-end server, regardless of whether the edge devices use GPUs. This offloading was done because AI training requires even more computing power than performing inference. Training AI models also require a large amount of labeled and pre-processed data, which is difficult to achieve on the edge when this data must be quality controlled by humans. AI training also requires personnel with specialized training, which is not expected from vehicle operators.

Stakeholder Need-02 requested that new training data can be generated on the fly during missions to be used for training new and better AI models in the future. This functionality was accounted for by the designed AI pipeline, presented in Figure 4.10 and described in Section 4.2.7.

Stakeholder Need-03, Need-04, and Need-05 requested the protection of the integrity, availability, and confidentiality of the system against adversaries. It was unnecessary to elaborate on the details of achieving this protection due to the assumption that existing systems incorporating the AI functionality of this theoretical system already have these accounted for. Moreover, this decision not to elaborate on system security mechanisms was appropriate when the thesis theme was AI, not cyber security.

Stakeholder Req-01 required the unmanned system to have the functionality to hand over control to a human operator at a remote location. This functionality is supported by the designed system architecture by having communication links between unmanned and manned edge devices. However, the exact implementation of the described functionality was left for future work as it was not strictly AI related and, by that, out of the scope of this thesis.

Stakeholder Req-02 required that an operator must be able to monitor and override or revert decisions made by AI. This functionality was accounted for by the designed AI pipeline, presented in Figure 4.10. However, the exact implementation of the described functionality was left for future work as it was not strictly AI related and, by that, out of the scope of this thesis.

Stakeholder Req-03 required that the AI must be disconnected from the system if the AI functionality is being attacked. Furthermore, if the subsystem using the AI, e.g., an unmanned vehicle, is dependent on the AI functionality to operate, the subsystem must automatically request manual control. The modular system design accounts for the first part of the requirement, shown in Figure 4.8 and Figure 4.9. The second part is accounted for by having direct communication with nearby subsystems. However, the exact implementation of the described functionality was left for future work as it was not strictly AI related and, by that, out of the scope of this thesis.

Stakeholder Req-04 required that AI models trained on classified data must not be stored on unmanned devices. This was accounted for when designing the logical- and physical system architectures, presented in Figure 4.8 and Figure 4.9.

Stakeholder Req-05, and Req-06 targeted the computation and performance of the AI models, respectively. These were accounted for by the general design of the

physical system architecture presented in Figure 4.9 on Page 45. If the AI slows down the system by running on the CPU, the AI model must be smaller, or the system must install an appropriate GPU. The same goes for whether the object detector can run at a stable 10 FPS by running on the CPU.

Stakeholder Req-07 required every AI prediction to be logged. This functionality was accounted for by the designed AI pipeline, presented in Figure 4.10 on Page 46. However, the exact implementation of the described functionality was left for future work as it was not strictly AI related and, by that, out of the scope of this thesis.

5.2.2 Combining Knowledge from Different Specializations

The previous section discusses how the stakeholder needs and requirements were met. Overall, all stakeholder needs and requirements were met regarding the AI implementation, which was the primary goal of this thesis. However, some stakeholder needs and requirements need different expertise, especially in cyber security. If this thesis were to cover the protection of the confidentiality, integrity, and availability of the designed theoretical system in detail, an expert in cyber security would be needed. New solutions for securing the system could then emerge instead of relying on the already implemented security mechanisms of existing systems.

The focus of this thesis on securing the AI functionality of the system against adversarial attacks is another statement as to why diverse teams are essential. If an expert in cyber security was tasked to secure an edge-based AI system against adversaries, they might have focused less on AI security than was done in this thesis. Moreover, the same applies the other way around. Therefore, assembling diverse teams with members from various specializations is strongly recommended to further develop this and future systems to ensure more innovative and comprehensive solutions to complex problems.

5.2.3 Choice of Object Detectors

Notes on the used object detectors were taken for every included paper on adversarial attacks and defenses. This data was collected to get an overview of the used object detectors and gain insight into the attacks' performance on different detectors. The overview in Figure 4.5 on Page 35 shows the used object detection families, which can be helpful to practitioners in deciding which object detectors to use.

However, is it better to use a publicly available detector that has been used in many adversarial attack experiments or to use a private detector that has yet to be the subject

of attacks? This question is rather difficult to answer and will largely depend on the use case and its requirements for security. With strict security requirements, using privately developed detectors is beneficial when factoring in the increased risk of white box attacks when using public detectors. On the other hand, practitioners get to know the exact security risks of the detector by using one of the widely used, publicly available detectors. Thereby they can be mitigated. However, if security is not of concern and the performance, cost, and efficiency of the detector are more important than security, the widely used detectors are recommended.

Object detectors should be monitored and tested for their performance and bugs before deployment, regardless of the choice of object detectors. Testing ensures that the detectors run at an acceptable speed and are not prone to known adversarial attacks. It is also an idea to deploy different detectors trained for the same purpose to the different edge devices. In that way, all devices need not be threatened if one of the devices is successfully degraded by adversarial attacks. The same device can also store several of these models to exchange for a different model if the one in use is attacked. Another idea is to run several detectors in parallel on the same platform to enable the detection of deviating classifications. However, this can also be achieved with a network of edge devices running different detectors while sharing information.

5.2.4 The Current Threat of Adversarial Attacks

This section discusses the current threat of adversarial attacks against the derived system while considering the recommended security and defense mechanisms summarized in Section 4.2.9 on Page 54.

The system security ensured by following NSMs G-guides[1] protects the AI against all digital adversarial attacks and white box adversarial attacks. This protection secures the AI against most adversarial attacks summarized in Table 4.1 on Page 27, leaving only black box physically realizable adversarial attacks.

Creating black box attacks with a high attack success rate is a difficult task in itself, as attackers have no access to the target model. Furthermore, the task gets even more complex when the attacks need to be physically realizable and not only effective in the digital domain. Combining these two complex problems creates a challenging task for researchers to explore to understand better the threat of adversarial attacks against cyber-physical systems.

The current threat of adversarial attacks against the system proposed in this thesis is relatively low, given the assumptions of system security and the derived recommendations toward AI security summarized in Section 4.2.9 on Page 54.

[1] Guidelines designed to ensure that systems handling classified information comply with the law of national security [176]. The guidelines are exempt from public disclosure.

However, the cross-task transferability discussed in Section 5.1.1 must be investigated further as it might broaden the threats covered in this thesis.

This conclusion is based on the current landscape of adversarial attacks covered in the systematic literature review of this thesis. Furthermore, the research field of physically realizable black box adversarial attacks was still somewhat premature in the writing of this thesis.

5.3 Ethical Considerations

The ethical issues of adversarial attacks on military systems are covered in this section. Adversarial attacks involve purposefully altering the data inputs to AI models to make them behave abnormally and make mistakes. Systems that rely on predictions from AI models run the risk of performing destructive behaviors and having unintended outcomes if the AI functionality gets manipulated by adversaries. In the worst-case scenarios, the effects of such adversarial attacks may endanger human lives. The threat of adversarial attacks must, therefore, be considered in the design process of such systems.

In the case of the system for unmanned military vehicles proposed in this thesis, the AI in question will perform object detection. For this use case, object detectors can be used to detect humans and vehicles on the battlefield. Moreover, the detected humans and vehicles can be classified as friendly, hostile, or civilian. The predictions from the object detectors can then be marked on a map for situational awareness on the battlefield. It is a relatively innocent task in itself; however, misclassifications can result in unwanted consequences if the system is not appropriately designed. Fortunately, the system was designed to always keep a man-in-the-loop for evaluating and overriding AI predictions to lower the risk of mistakes.

One of the dangers of adversarial attacks against object detection in military systems is the potential misclassification of civil objects as military targets. Such outcomes can be achieved intentionally using targeted attacks or at random using untargeted attacks. Fortunately, international humanitarian law [179] forbids civilian targeting. Therefore, such adversarial attacks should be defined as war crimes and be forbidden.

However, holding those performing adversarial attacks responsible for their actions can be challenging, as adversarial attacks are not only limited to resourceful adversaries. Everyone with the proper knowledge and intentions can carry them out due to their availability in research and cost-effectiveness. Therefore, it is of high importance for researchers and the industry to design proper defenses against adversarial attacks. Furthermore, as this thesis exemplifies, AI security must

be considered when designing AI-enabled systems to mitigate the risk posed by adversarial attacks.

Humans, as well as AI, are prone to make mistakes. Well-designed AI-based systems can be as good or better than humans to observe and perceive the environment. Therefore, using object detection to assist humans in their decision-making can reduce the risk of human error, which can save lives. Furthermore, the safe deployment of AI is an important research topic for reducing the chance of human error in safety-critical systems.

5.4 Future Research

This section proposes several directions for future work on the topics covered in this thesis. They cover the calls for more research on physically realizable adversarial attacks and universal defense methods, possible extensions of the thesis, including further research on cross-task attack transferability and adversarial attacks on 3D object detection, and viewing the thesis topic from other specialisations and disciplines like cyber security and law.

Physically Realizable Adversarial Attacks Adversarial attacks must be physically realizable to be effective against cyber-physical systems like the one presented in this thesis. The research covered in the systematic literature review includes experiments with physical adversarial attacks, but most lack performance and are limited to specific scenarios. It is challenging to create universally effective physical adversarial attacks due to the endless possible viewing angles and differences in lighting. However, it is important to continue this research to further assess the threat of adversarial attacks against cyber-physical systems.

Universal Defense Methods Many defense methods exist to mitigate the threat of adversarial attacks in object detection, seen in Table 4.2 on Page 28. However, they all have their limitations. While adversarial training and adversarial detection only learn to defend against the attacks they are trained with, input transformations and adversarial training can reduce the accuracy of detectors on clean data, and robustifying the model architecture can reduce performance. Therefore, research towards universal defense methods against adversarial attacks is needed to secure object detectors against all attacks without negatively affecting performance or accuracy.

Cross-Task Transferability As discussed in Section 5.1.1, adversarial attacks in image classification have proven effective in attacking object detectors. This discovery broadens the threat of adversarial attacks in object detection further

than what was covered in the systematic literature review of this thesis. Therefore, more research on the cross-task transferability of attacks must be done to evaluate the threat further. Researchers should extend this literature review if it is to be discovered that attack methods designed for similar tasks are effective in attacking object detectors.

Adversarial Attacks Against 3D Object Detection Appendix B includes an overview of adversarial attacks and defenses in object detection that uses different data types than images and video. These papers were excluded from this thesis, as the goal was to evaluate the threat against image and video object detection. However, other data types like LiDAR and infrared imagery are valuable sources of information in cyber-physical systems, like unmanned vehicles. The threats against object detectors using these datatypes should therefore be evaluated in future research.

Cyber Security Experts Perspective As discussed in Section 5.2.2, the goal of this thesis to design a secure AI-based system for unmanned military vehicles can also be done from the perspective of cyber security experts. Doing so can result in new ideas for securing the system instead of relying on the already implemented security mechanisms of existing systems. Future research should therefore be done toward securing the proposed system, e.g., in the form of a complementing master's thesis.

Lawyers Perspective Section 5.3 discussed the complications of using adversarial attacks in armed conflicts. Adversarial attacks can endanger civilians if performed against systems not designed to be resilient to adversarial attacks. Therefore, the possible implications of international humanitarian law [179] should be researched by lawyers to define whether existing laws cover adversarial attacks or if new ones should be made to restrict their use.

Chapter 6

Conclusion

The main objectives of this thesis were to understand the status quo of attacks and defenses in deep learning-based object detection that needs attention in KDA's context (Goal-01) and use that information to derive a system architecture and requirements for the safe deployment of AI in an unmanned military setting (Goal-02).

Goal-01 was achieved through a systematic literature review with the research question of whether the current state of adversarial attacks poses a threat to image and video object detectors in unmanned systems – and how they can be defended. The results showed that most adversarial attacks were made to attack object detectors in the digital domain rather than the real world. As unmanned systems operate with strict security standards in the real world, digital adversarial attacks become relatively unthreatening. The same applies to white box attacks, as the strict security standards prevent the AI model from getting astray. That leaves physically realizable adversarial attacks as the most likely attacks that can affect the AI in unmanned military systems. However, creating physically realizable black box adversarial attacks is a challenging task yet to be properly solved. With the added layer of protection from adversarial defenses like adversarial- training and detection, adversarial attacks become an even lesser threat against unmanned military vehicles.

Goal-02 was achieved through a requirements engineering process, including workshops, technical analysis, and expert interviews. The designed system met all of the derived stakeholder- needs and requirements relevant to the AI functionality. The designed system was expanded beyond only unmanned vehicles, as supporting manned vehicles and a high-performance back-end server were needed to cover the entire AI lifecycle. Furthermore, a thorough threat analysis of adversarial attacks against the designed system was delivered. It was shown that most adversarial attacks against object detection in manned- and unmanned military systems are already defended by the security mechanisms of similar existing systems. However, AI-specific defenses can be beneficial as an added layer of security to ensure system reliability and

protection against physically realizable adversarial attacks not covered by traditional defenses. Therefore, guidelines were made for securing the system against adversarial attacks by implementing adversarial defense methods: input transformations, robust model selection, adversarial training, and adversarial detection. The manned- and unmanned edge platforms were designed with different security levels as unmanned devices are more easily lost and should thereby not handle classified information. Therefore, the security recommendations varied across the subsystems due to their differences in security requirements and available computing power.

This thesis delivered a comprehensive review of the state-of-the-art adversarial attacks and defenses in object detection in the context of unmanned military vehicles – the first of its kind and a valuable resource for the research field and the defense industry. KDA gained valuable information on how AI can safely be deployed in unmanned military systems and how to maintain a cycle of ever-improving AI models throughout the lifetime of the system.

Bibliography

[1] C. Szegedy et al. “Intriguing properties of neural networks”. In: *Intriguing Properties of Neural Networks* (2013).

[2] I.J. Goodfellow, J. Shlens, and C. Szegedy. “Explaining and harnessing adversarial examples”. In: *Explaining and Harnessing Adversarial Examples* (2014).

[3] Y. Wang et al. “Adversarial attacks on Faster R-CNN object detector”. In: *Neurocomputing* 382 (2020), pp. 87–95. DOI: 10.1016/j.neucom.2019.11.051. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85076566920&doi=10.1016%2fj.neucom.2019.11.051&partnerID=40&md5=16e28bfcf7d1b89a2769e2a1968ee67d.

[4] Athanasios Voulodimos et al. “Deep Learning for Computer Vision: A Brief Review”. en. In: *Computational Intelligence and Neuroscience* 2018 (2018), pp. 1–13. ISSN: 1687-5265, 1687-5273. DOI: 10.1155/2018/7068349. URL: https://www.hindawi.com/journals/cin/2018/7068349/ (visited on 03/13/2023).

[5] Junyi Chai et al. “Deep learning in computer vision: A critical review of emerging techniques and application scenarios”. en. In: *Machine Learning with Applications* 6 (Dec. 2021), p. 100134. ISSN: 2666-8270. DOI: 10.1016/j.mlwa.2021.100134. URL: https://www.sciencedirect.com/science/article/pii/S2666827021000670 (visited on 03/13/2023).

[6] Geert Litjens et al. “A survey on deep learning in medical image analysis”. en. In: *Medical Image Analysis* 42 (Dec. 2017), pp. 60–88. ISSN: 1361-8415. DOI: 10.1016/j.media.2017.07.005. URL: https://www.sciencedirect.com/science/article/pii/S1361841517301135 (visited on 03/13/2023).

[7] Abhishek Gupta et al. “Deep learning for object detection and scene perception in self-driving cars: Survey, challenges, and open issues”. en. In: *Array* 10 (July 2021), p. 100057. ISSN: 2590-0056. DOI: 10.1016/j.array.2021.100057. URL: https://www.sciencedirect.com/science/article/pii/S2590005621000059 (visited on 03/13/2023).

[8] Zhenwei Yu, Yonggang Shen, and Chenkai Shen. “A real-time detection approach for bridge cracks based on YOLOv4-FPM”. en. In: *Automation in Construction* 122 (Feb. 2021), p. 103514. ISSN: 0926-5805. DOI: 10.1016/j.autcon.2020.103514. URL: https://www.sciencedirect.com/science/article/pii/S0926580520310943 (visited on 03/13/2023).

[9] Shaoqing Ren et al. “Faster r-cnn: Towards real-time object detection with region proposal networks”. In: *Advances in neural information processing systems* 28 (2015).

[10] Joseph Redmon et al. “You only look once: Unified, real-time object detection”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 779–788.

[11] Wei Liu et al. “SSD: Single Shot MultiBox Detector”. en. In: *Computer Vision – ECCV 2016*. Ed. by Bastian Leibe et al. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2016, pp. 21–37. ISBN: 978-3-319-46448-0. DOI: 10.1007/978-3-319-46448-0_2.

[12] N. Akhtar and A. Mian. “Threat of Adversarial Attacks on Deep Learning in Computer Vision: A Survey”. In: *IEEE Access* 6 (2018), pp. 14410–14430. ISSN: 2169-3536. DOI: 10.1109/ACCESS.2018.2807385.

[13] H. Ren, T. Huang, and H. Yan. “Adversarial examples: attacks and defenses in the physical world”. In: *International Journal of Machine Learning and Cybernetics* 12.11 (2021), pp. 3325–3336. ISSN: 1868-8071. DOI: 10.1007/s13042-020-01242-z.

[14] B. Xu, J. Zhu, and D. Wang. “Adversarial Attacks for Object Detection”. In: Chinese Control Conference, CCC. Vol. 2020-July. 2020, pp. 7281–7287. DOI: 10.23919/CCC50068.2020.9188998. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85091398289&doi=10.23919%2fCCC50068.2020.9188998&partnerID=40&md5=1ec16a71dc6f81fde41374f3985b772b.

[15] B. Xi. “Adversarial machine learning for cybersecurity and computer vision: Current developments and challenges”. In: *Wiley Interdisciplinary Reviews: Computational Statistics* 12.5 (2020). ISSN: 1939-5108. DOI: 10.1002/wics.1511.

[16] N. Akhtar et al. “Advances in Adversarial Attacks and Defenses in Computer Vision: A Survey”. In: *IEEE Access* 9 (2021), pp. 155161–155196. ISSN: 2169-3536. DOI: 10.1109/ACCESS.2021.3127960.

[17] C. Oh, A. Xompero, and A. Cavallaro. “Visual adversarial attacks and defenses”. In: *Advanced Methods and Deep Learning in Computer Vision*. 2021, pp. 511–543. DOI: 10.1016/B978-0-12-822109-9.00024-2. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85108187465&doi=10.1016%2fB978-0-12-822109-9.00024-2&partnerID=40&md5=81126ed98ac8406b3b94441064aff5ac.

[18] Y. Deng et al. “Deep Learning-Based Autonomous Driving Systems: A Survey of Attacks and Defenses”. In: *IEEE Transactions on Industrial Informatics* (2021). DOI: 10.1109/TII.2021.3071405. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85103878746&doi=10.1109%2fTII.2021.3071405&partnerID=40&md5=0f6d3a7341f3a85fb858c5114844ed2f.

[19] A. Amirkhani, M.P. Karimi, and A. Banitalebi-Dehkordi. “A survey on adversarial attacks and defenses for object detection and their applications in autonomous vehicles”. In: *Visual Computer* (2022). DOI: 10.1007/s00371-022-02660-6. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85138057178&doi=10.1007%2fs00371-022-02660-6&partnerID=40&md5=6542734592b302c166e6fa12a5d9394d.

[20] S. Wang et al. “Adversarial Robustness of Deep Sensor Fusion Models”. In: Proceedings - 2022 IEEE/CVF Winter Conference on Applications of Computer Vision, WACV 2022. 2022, pp. 1371–1380. DOI: 10.1109/WACV51458.2022.00144. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85126089794&doi=10.1109%2fWACV51458.2022.00144&partnerID=40&md5=654647e33dbcef156665c48a4115b879.

[21] J.-X. Mi et al. “Adversarial examples based on object detection tasks: A survey”. In: *Neurocomputing* 519 (2023), pp. 114–126. DOI: 10.1016/j.neucom.2022.10.046. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85142908380&doi=10.1016%2fj.neucom.2022.10.046&partnerID=40&md5=b4b86bb2b8f7fd9deb39f710026d55b2.

[22] A. Raja, L. Njilla, and J. Yuan. “Blur the Eyes of UAV: Effective Attacks on UAV-based Infrastructure Inspection”. In: Proceedings - International Conference on Tools with Artificial Intelligence, ICTAI. Vol. 2021-November. ISSN: 1082-3409. 2021, pp. 661–665. ISBN: 978-1-66540-898-1. DOI: 10.1109/ICTAI52525.2021.00105.

[23] Y. Li, G. Xu, and W. Li. “FA: A Fast Method to Attack Real-time Object Detection Systems”. In: 2020 IEEE/CIC International Conference on Communications in China, ICCC 2020. 2020, pp. 1268–1273. DOI: 10.1109/ICCC49849.2020.9238807. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-

85097568584&doi=10.1109%2fICCC49849.2020.9238807&partnerID=40&md5=ccecbcc2a9d9757bd238784393da5d61.

[24] S. Chen et al. “Relevance attack on detectors”. In: *Pattern Recognition* 124 (2022). DOI: 10.1016/j.patcog.2021.108491. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85121770121&doi=10.1016%2fj.patcog.2021.108491&partnerID=40&md5=213c7c9cbed28cb5057e608879ed5ebd.

[25] X. Liu et al. “DPatch: An adversarial patch attack on object detectors”. In: CEUR Workshop Proceedings. Vol. 2301. 2019. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85060640734&partnerID=40&md5=8026a755f91ed54a9e013347d7c3ee14.

[26] H. Huang et al. “RPATTACK: REFINED PATCH ATTACK ON GENERAL OBJECT DETECTORS”. In: Proceedings - IEEE International Conference on Multimedia and Expo. 2021. DOI: 10.1109/ICME51207.2021.9428443. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85121902305&doi=10.1109%2fICME51207.2021.9428443&partnerID=40&md5=4407d235e13f16b366fe4f0c4413d2b9.

[27] Y. Zhang et al. “Camou: Learning a vehicle camouflage for physical adversarial attack on object detectors in the wild”. In: 7th International Conference on Learning Representations, ICLR 2019. 2019. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85083952943&partnerID=40&md5=aa2f289b748be3be55060c4106a50e55.

[28] H. Zhang and J. Wang. “Towards adversarially robust object detection”. In: Proceedings of the IEEE International Conference on Computer Vision. Vol. 2019-October. 2019, pp. 421–430. DOI: 10.1109/ICCV.2019.00051. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85081930250&doi=10.1109%2fICCV.2019.00051&partnerID=40&md5=330c390396412c38fba130e4b6e22979.

[29] C. Xiao et al. “AdvIT: Adversarial frames identifier based on temporal consistency in videos”. In: Proceedings of the IEEE International Conference on Computer Vision. Vol. 2019-October. 2019, pp. 3967–3976. DOI: 10.1109/ICCV.2019.00407. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85081909913&doi=10.1109%2fICCV.2019.00407&partnerID=40&md5=3586174b68e05f81b321f36ff7ff2c39.

[30] M.P. Karimi, A. Amirkhani, and S.B. Shokouhi. “Robust object detection against adversarial perturbations with gabor filter”. In: 2021 29th Iranian Conference on Electrical Engineering, ICEE 2021. 2021, pp. 187–192. DOI: 10.1109/ICEE52715.2021.9544499. URL: https://www.scopus.com/inward/record.

uri?eid=2-s2.0-85117453087&doi=10.1109%2fICEE52715.2021.9544499&partnerID=40&md5=3f13b7feb4ac11d85bc26d7252f701c5.

[31] M. Naseer et al. “A Self-supervised Approach for Adversarial Robustness”. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. 2020, pp. 259–268. DOI: 10.1109/CVPR42600.2020.00034. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85094862396&doi=10.1109%2fCVPR42600.2020.00034&partnerID=40&md5=527f0f2e0944a399add006d185805920.

[32] L. Aurdal et al. “Adversarial camouflage for naval vessels”. In: Proceedings of SPIE - The International Society for Optical Engineering. Vol. 11169. ISSN: 0277-786X. 2019. ISBN: 978-1-5106-3041-3. DOI: 10.1117/12.2532756.

[33] A.N. Alfimtsev et al. “Hostis humani ET mashinae: Adversarial camouflage generation”. In: *Journal of Advanced Research in Dynamical and Control Systems* 11.2 (2019), pp. 382–392. ISSN: 1943-023X. URL: https://www.researchgate.net/publication/332875737_Hostis_Humani_ET_Mashinae_Adversarial_Camouflage_Generation.

[34] Y. Wang et al. “Dual Attribute Adversarial Camouflage toward camouflaged object detection”. In: *Defence Technology* (2021). DOI: 10.1016/j.dt.2021.12.003. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85121905503&doi=10.1016%2fj.dt.2021.12.003&partnerID=40&md5=b9f7ea3ad2ac889de5de98a331a45b0e.

[35] J. Kim et al. “Camouflaged Adversarial Attack on Object Detector”. In: International Conference on Control, Automation and Systems. Vol. 2021-October. 2021, pp. 613–617. DOI: 10.23919/ICCAS52745.2021.9650004. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85124194916&doi=10.23919%2fICCAS52745.2021.9650004&partnerID=40&md5=9d3bf31ff9a75015eb191e3c65bb67a1.

[36] Jane Webster and Richard T Watson. “Analyzing the past to prepare for the future: Writing a literature review”. In: *MIS Quarterly* 26.2 (June 2002), pp. xiii–xxiii. ISSN: 0276-7783. URL: https://www.jstor.org/stable/4132319#metadata_info_tab_contents.

[37] Micah D. J. Peters et al. “Guidance for conducting systematic scoping reviews”. In: *JBI Evidence Implementation* 13.3 (Sept. 2015), p. 141. ISSN: 2691-3321. DOI: 10.1097/XEB.0000000000000050. URL: https://journals.lww.com/ijebh/Fulltext/2015/09000/Guidance_for_conducting_systematic_scoping_reviews.5.aspx (visited on 12/15/2022).

[38] Barbara Kitchenham. “Procedures for performing systematic reviews”. In: *Keele, UK, Keele University* 33.2004 (2004), pp. 1–26.

[39] Micah D. J. Peters et al. “Best practice guidance and reporting items for the development of scoping review protocols”. In: *JBI Evidence Synthesis* 20.4 (Apr. 2022), pp. 953–968. ISSN: 2689-8381. DOI: 10.11124/JBIES-21-00242. URL: https://journals.lww.com/jbisrir/Fulltext/2022/04000/Best_practice_guidance_and_reporting_items_for_the.3.aspx?context=FeaturedArticles&collectionId=5 (visited on 12/08/2022).

[40] Claes Wohlin. “Guidelines for snowballing in systematic literature studies and a replication in software engineering”. In: *Proceedings of the 18th International Conference on Evaluation and Assessment in Software Engineering*. EASE ’14. New York, NY, USA: Association for Computing Machinery, May 13, 2014, pp. 1–10. ISBN: 978-1-4503-2476-2. DOI: 10.1145/2601248.2601268. URL: https://doi.org/10.1145/2601248.2601268 (visited on 03/17/2022).

[41] C. Xie et al. “Adversarial Examples for Semantic Segmentation and Object Detection”. In: Proceedings of the IEEE International Conference on Computer Vision. Vol. 2017-October. 2017, pp. 1378–1387. DOI: 10.1109/ICCV.2017.153. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85041912770&doi=10.1109%2fICCV.2017.153&partnerID=40&md5=b364ab7034e1ce3864de4b5c696d6155.

[42] D.Y. Yang et al. “Building Towards Invisible Cloak: Robust Physical Adversarial Attack on YOLO Object Detector”. In: 2018 9th IEEE Annual Ubiquitous Computing, Electronics and Mobile Communication Conference, UEMCON 2018. 2018, pp. 368–374. DOI: 10.1109/UEMCON.2018.8796670. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85071522794&doi=10.1109%2fUEMCON.2018.8796670&partnerID=40&md5=8ffc20df53ef998b7913cc76a96b52e9.

[43] Y. Zhao et al. *An Universal Perturbation Generator for Black-Box Attacks Against Object Detectors*. Vol. 11910 LNCS. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). Pages: 72. 2019. 63 pp. DOI: 10.1007/978-3-030-34139-8_7. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85076129873&doi=10.1007%2f978-3-030-34139-8_7&partnerID=40&md5=4b34690167d44f708d992b5c9363d89f.

[44] E.R. Balda, A. Behboodi, and R. Mathar. “Perturbation Analysis of Learning Algorithms: Generation of Adversarial Examples from Classification to Regression”. In: *IEEE Transactions on Signal Processing* 67.23 (2019), pp. 6078–

6091. DOI: 10.1109/TSP.2019.2943232. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85077807462&doi=10.1109%2fTSP.2019.2943232&partnerID=40&md5=01d695bce10cc440e4d47e93d65a2b91.

[45] K.T. Co et al. “Procedural noise adversarial examples for black-box attacks on deep convolutional networks”. In: Proceedings of the ACM Conference on Computer and Communications Security. ISSN: 1543-7221. 2019, pp. 275–289. ISBN: 978-1-4503-6747-9. DOI: 10.1145/3319535.3345660.

[46] Y. Li et al. “Robust adversarial perturbation on deep proposal-based models”. In: British Machine Vision Conference 2018, BMVC 2018. 2019. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85084017828&partnerID=40&md5=f1b267104387c7fb837f2693a61ee2bb.

[47] X. Wei et al. “Transferable adversarial attacks for image and video object detection”. In: IJCAI International Joint Conference on Artificial Intelligence. Vol. 2019-August. 2019, pp. 954–960. DOI: 10.24963/ijcai.2019/134. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85074956312&doi=10.24963%2fijcai.2019%2f134&partnerID=40&md5=ed034978b0bed65494a496ec3ee50e1b.

[48] Y. Wang et al. “An adversarial attack on DNN-based black-box object detectors”. In: *Journal of Network and Computer Applications* 161 (2020). DOI: 10.1016/j.jnca.2020.102634. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85083034800&doi=10.1016%2fj.jnca.2020.102634&partnerID=40&md5=a50efb1d810c59bc4bc8a4eb68716ffc.

[49] S. Khattar and C. Rama Krishna. “Adversarial attack to fool object detector”. In: *Journal of Discrete Mathematical Sciences and Cryptography* 23.2 (2020), pp. 547–562. DOI: 10.1080/09720529.2020.1729504. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85084850471&doi=10.1080%2f09720529.2020.1729504&partnerID=40&md5=2916267641cf3e595d4a3f57830bf454.

[50] K.-H. Chow et al. “Adversarial Objectness Gradient Attacks in Real-time Object Detection Systems”. In: Proceedings - 2020 2nd IEEE International Conference on Trust, Privacy and Security in Intelligent Systems and Applications, TPS-ISA 2020. 2020, pp. 263–272. DOI: 10.1109/TPS-ISA50397.2020.00042. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85100406699&doi=10.1109%2fTPS-ISA50397.2020.00042&partnerID=40&md5=f15ddc37b660b987252ac6ec72308fcd.

[51] K. Yang et al. “Beyond digital domain: Fooling deep learning based recognition system in physical world”. In: AAAI 2020 - 34th AAAI Conference on Artificial Intelligence. 2020, pp. 1088–1095. URL: https://www.scopus.com/

inward / record . uri ? eid = 2 - s2 . 0 - 85105976605 & partnerID = 40 & md5 = ed5396dc1539860e524769a33179ed3d.

[52] H. Zhang, W. Zhou, and H. Li. “Contextual adversarial attacks for object detection”. In: Proceedings - IEEE International Conference on Multimedia and Expo. Vol. 2020-July. 2020. DOI: 10.1109/ICME46284.2020.9102805. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85090384077&doi=10.1109%2fICME46284.2020.9102805&partnerID=40&md5=ee702bfdf8ea42ccb02541a1d04a3b11.

[53] Y. Liu, X. Zhu, and X. Huang. “Efficientwarm restart adversarial attack for object detection”. In: CEUR Workshop Proceedings. Vol. 2881. 2020, pp. 21–23. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85108023974&partnerID=40&md5=a331b115fbb8efcca4e7325cf4d375d6.

[54] Y. Lu et al. “Enhancing cross-task black-box transferability of adversarial examples with dispersion reduction”. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. 2020, pp. 937–946. DOI: 10.1109/CVPR42600.2020.00102. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85094207043&doi=10.1109%2fCVPR42600.2020.00102&partnerID=40&md5=3fcefba3134fedb0024d5be4b14f65a7.

[55] Q. Liao et al. “Fast Local Attack: Generating Local Adversarial Examples for Object Detectors”. In: Proceedings of the International Joint Conference on Neural Networks. 2020. DOI: 10.1109/IJCNN48605.2020.9206811. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85093862586&doi=10.1109%2fIJCNN48605.2020.9206811&partnerID=40&md5=a05f1e358c78d544498f9462510a5741.

[56] Y. Zhang, F. Wang, and W. Ruan. “Fooling object detectors: Adversarial attacks by half-neighbor masks”. In: CEUR Workshop Proceedings. Vol. 2881. 2020, pp. 36–39. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85108028016&partnerID=40&md5=70772ec09504676ece579542c7ba6524.

[57] Z. Chen, J. Liu, and H. Chen. “Generating Adversarial Examples Based on Subarea Noise Texture for Efficient Black-Box Attacks”. In: ACM International Conference Proceeding Series. 2020. DOI: 10.1145/3446132.3446174. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85102967600&doi=10.1145%2f3446132.3446174&partnerID=40&md5=cd564aa71df00da571a7188218bd6902.

[58] Z. Liu et al. “MI-FGSM on Faster R-CNN Object Detector”. In: ACM International Conference Proceeding Series. Vol. PartF168342. 2020, pp. 27–32. DOI: 10.1145/3447450.3447455. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85104180222&doi=10.1145%2f3447450.3447455&partnerID=40&md5=0a17bc4bc2b53986d0d3d337a7e6ead7.

[59] Q. Zhang et al. “Towards cross-task universal perturbation against black-box object detectors in autonomous driving”. In: *Computer Networks* 180 (2020). DOI: 10.1016/j.comnet.2020.107388. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85088224849&doi=10.1016%2fj.comnet.2020.107388&partnerID=40&md5=8c054bed0610afff6950e543c34f7cf9.

[60] S. Huang et al. “Two Improved Methods of Generating Adversarial Examples against Faster R-CNNs for Tram Environment Perception Systems”. In: *Complexity* 2020 (2020). DOI: 10.1155/2020/6814263. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85092644195&doi=10.1155%2f2020%2f6814263&partnerID=40&md5=e75c6ec34c1992c371dfd69955cb7cd4.

[61] X. Wang et al. “Adversarial point cloud perturbations against 3D object detection in autonomous driving systems”. In: *Neurocomputing* 466 (2021), pp. 27–36. DOI: 10.1016/j.neucom.2021.09.027. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85115915801&doi=10.1016%2fj.neucom.2021.09.027&partnerID=40&md5=e9b0adae45d47a66f4910e9b01d13e32.

[62] L.Y.U. Haoran et al. “A CMA-ES-Based Adversarial Attack Against Black-Box Object Detectors”. In: *Chinese Journal of Electronics* 30.3 (2021), pp. 406–412. DOI: 10.1049/cje.2021.03.003. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85110966064&doi=10.1049%2fcje.2021.03.003&partnerID=40&md5=0fe6d5cc634d745f4e1dd373c05acb88.

[63] X. Kuang et al. “A discrete cosine transform-based query efficient attack on black-box object detectors”. In: *Information Sciences* 546 (2021), pp. 596–607. DOI: 10.1016/j.ins.2020.05.089. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85090900676&doi=10.1016%2fj.ins.2020.05.089&partnerID=40&md5=67249d5a6fe88e6dd16e6408dd0f62d2.

[64] S. Huang et al. “An improved ShapeShifter method of generating adversarial examples for physical attacks on stop signs against Faster R-CNNs”. In: *Computers and Security* 104 (2021). DOI: 10.1016/j.cose.2020.102120. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85101097035&doi=10.1016%2fj.cose.2020.102120&partnerID=40&md5=9f99c7d19357a78fb646cd72cdc8c008.

[65] Y. Xiao, C.-M. Pun, and B. Liu. “Fooling deep neural detection networks with adaptive object-oriented adversarial perturbation”. In: *Pattern Recognition* 115 (2021). DOI: 10.1016/j.patcog.2021.107903. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85101503839&doi=10.1016%2fj.patcog.2021.107903&partnerID=40&md5=d6a8626d27b66c597dc07a03d7ad51b8.

[66] M. Yuan and X. Wei. “Generating Adversarial Remote Sensing Images via Pan-Sharpening Technique”. In: AdvM 2021 - Proceedings of the 1st International Workshop on Adversarial Learning for Multimedia, co-located with ACM MM 2021. 2021, pp. 15–20. DOI: 10.1145/3475724.3483602. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85121115685&doi=10.1145%2f3475724.3483602&partnerID=40&md5=26af346e58ec82e0140ef194a5700fd5.

[67] M. Xue et al. “NaturalAE: Natural and robust physical adversarial examples for object detectors”. In: *Journal of Information Security and Applications* 57 (2021). DOI: 10.1016/j.jisa.2020.102694. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85098936506&doi=10.1016%2fj.jisa.2020.102694&partnerID=40&md5=def828d2ba63281565b03823daf95fc6.

[68] S. Liang et al. “Parallel rectangle flip attack: A Query-based Black-box Attack against Object Detection”. In: Proceedings of the IEEE International Conference on Computer Vision. 2021, pp. 7677–7687. DOI: 10.1109/ICCV48922.2021.00760. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85119884470&doi=10.1109%2fICCV48922.2021.00760&partnerID=40&md5=ac57657629dd8c31942a3fcfa6cc7391.

[69] M. Wang, H. Wang, and L. Wang. “PGD-Optimized Patch and Noise Joint Embedded Adversarial Example for Faster RCNN and YOLOv4”. In: 2021 5th International Conference on Automation, Control and Robots, ICACR 2021. 2021, pp. 78–83. DOI: 10.1109/ICACR53472.2021.9605188. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85123466945&doi=10.1109%2fICACR53472.2021.9605188&partnerID=40&md5=2c59ca677bc4ea0e234599e6411356c2.

[70] O. Mohamad Nezami et al. “PICK-OBJECT-ATTACK: Type-specific adversarial attack for object detection”. In: *Computer Vision and Image Understanding* 211 (2021). DOI: 10.1016/j.cviu.2021.103257. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85113699794&doi=10.1016%2fj.cviu.2021.103257&partnerID=40&md5=06a743d339f657f1878fd579a35d0996.

[71] M. Xue et al. “SocialGuard: An adversarial example based privacy-preserving technique for social images”. In: *Journal of Information Security and Applica-*

tions 63 (2021). DOI: 10.1016/j.jisa.2021.102993. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85118694023&doi=10.1016%2fj.jisa.2021.102993&partnerID=40&md5=81eb05493b801c1709731d4643b92a23.

[72] Q. Liao et al. "TRANSFERABLE ADVERSARIAL EXAMPLES FOR ANCHOR FREE OBJECT DETECTION". In: Proceedings - IEEE International Conference on Multimedia and Expo. 2021. DOI: 10.1109/ICME51207.2021.9428301. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85126488328&doi=10.1109%2fICME51207.2021.9428301&partnerID=40&md5=dcf849d8ce1a23a1c1c06acffeea277b.

[73] Y. Li et al. "TransRPN: Towards the Transferable Adversarial Perturbations using Region Proposal Networks and Beyond". In: *Computer Vision and Image Understanding* 213 (2021). ISSN: 1077-3142. DOI: 10.1016/j.cviu.2021.103302.

[74] D. Li, J. Zhang, and K. Huang. "Universal adversarial perturbations against object detection". In: *Pattern Recognition* 110 (2021). DOI: 10.1016/j.patcog.2020.107584. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85089746321&doi=10.1016%2fj.patcog.2020.107584&partnerID=40&md5=cd2ebce7cd467e087fbaf8dc59fb59c8.

[75] S. Liang et al. *A Large-Scale Multiple-objective Method for Black-box Attack Against Object Detection*. Vol. 13664 LNCS. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). Pages: 636. 2022. 619 pp. DOI: 10.1007/978-3-031-19772-7_36. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85142754081&doi=10.1007%2f978-3-031-19772-7_36&partnerID=40&md5=4962559aa9f2c370043a5a7328513566.

[76] M. Yin et al. "ADC: Adversarial attacks against object Detection that evade Context consistency checks". In: Proceedings - 2022 IEEE/CVF Winter Conference on Applications of Computer Vision, WACV 2022. 2022, pp. 2836–2845. DOI: 10.1109/WACV51458.2022.00289. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85126124404&doi=10.1109%2fWACV51458.2022.00289&partnerID=40&md5=36a6cb0db3663e23aea54d8f9f0eb5f3.

[77] J.I. Choi and Q. Tian. "Adversarial Attack and Defense of YOLO Detectors in Autonomous Driving Scenarios". In: IEEE Intelligent Vehicles Symposium, Proceedings. Vol. 2022-June. 2022, pp. 1011–1017. DOI: 10.1109/IV51971.2022.9827222. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85135382030&doi=10.1109%2fIV51971.2022.9827222&partnerID=40&md5=8bd8fb84641716de9734e71f276ad79e.

[78] A. Raja, L. Njilla, and J. Yuan. “Adversarial Attacks and Defenses Toward AI-Assisted UAV Infrastructure Inspection”. In: *IEEE Internet of Things Journal* 9.23 (2022), pp. 23379–23389. ISSN: 2327-4662. DOI: 10.1109/JIOT.2022.3206276.

[79] D. Lang et al. “An Adversarial Attack Method against Specified Objects Based on Instance Segmentation”. In: *Information (Switzerland)* 13.10 (2022). DOI: 10.3390/info13100465. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85140482091&doi=10.3390%2finfo13100465&partnerID=40&md5=a633eda89d53bb89d037a253f45d72bb.

[80] I.A. Elaalami, S.O. Olatunji, and R.M. Zagrouba. “AT-BOD: An Adversarial Attack on Fool DNN-Based Blackbox Object Detection Models”. In: *Applied Sciences (Switzerland)* 12.4 (2022). DOI: 10.3390/app12042003. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85124710020&doi=10.3390%2fapp12042003&partnerID=40&md5=d3c34efc49912cddfda71ea9d1aebd18.

[81] Y. Zhang et al. “Boosting cross-task adversarial attack with random blur”. In: *International Journal of Intelligent Systems* 37.10 (2022), pp. 8139–8154. DOI: 10.1002/int.22932. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85130419585&doi=10.1002%2fint.22932&partnerID=40&md5=09c4f85c8bab8807a6258d0bfa1bc661.

[82] X. Kang et al. “Crafting universal adversarial perturbations with output vectors”. In: *Neurocomputing* 501 (2022), pp. 294–305. DOI: 10.1016/j.neucom.2022.06.005. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85133761133&doi=10.1016%2fj.neucom.2022.06.005&partnerID=40&md5=517b06c1c54515dd02c9ccac818bd6e3.

[83] D. Wang et al. “Daedalus: Breaking Nonmaximum Suppression in Object Detection via Adversarial Examples”. In: *IEEE Transactions on Cybernetics* 52.8 (2022), pp. 7427–7440. DOI: 10.1109/TCYB.2020.3041481. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85099195781&doi=10.1109%2fTCYB.2020.3041481&partnerID=40&md5=de3e1c102bf6664463027c98da384251.

[84] Z. Lu et al. “An Enhanced Image Patch Tensor Decompostion for Infrared Small Target Detection”. In: *Remote Sensing* 14.23 (2022). ISSN: 2072-4292. DOI: 10.3390/rs14236044.

[85] K. Chan and B.H.C. Cheng. *EvoAttack: An Evolutionary Search-Based Adversarial Attack for Object Detection Models*. Vol. 13711 LNCS. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). Pages: 97. 2022. 83 pp. DOI:

10.1007/978-3-031-21251-2_6. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85142712384&doi=10.1007%2f978-3-031-21251-2_6&partnerID=40&md5=a76fac50f3f40f243db2469cf4c1f6be.

[86] H. Zanddizari, B. Zeinali, and J.M. Chang. "Generating Black-Box Adversarial Examples in Sparse Domain". In: *IEEE Transactions on Emerging Topics in Computational Intelligence* 6.4 (2022), pp. 795–804. DOI: 10.1109/TETCI.2021.3122467. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85118679531&doi=10.1109%2fTETCI.2021.3122467&partnerID=40&md5=3b949e638916f48ce077af25ddf7c52f.

[87] Y. Wang et al. "Improving the Imperceptibility of Adversarial Examples Based on Weakly Perceptual Perturbation in Key Regions". In: *Security and Communication Networks* 2022 (2022). DOI: 10.1155/2022/8288855. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85145971921&doi=10.1155%2f2022%2f8288855&partnerID=40&md5=7155f88c2fc0d38e3e3d4487dda16317.

[88] A.K. Akan, E. Akbas, and F.T.Y. Vural. "Just noticeable difference for machine perception and generation of regularized adversarial images with minimal perturbation". In: *Signal, Image and Video Processing* 16.6 (2022), pp. 1595–1606. DOI: 10.1007/s11760-021-02114-x. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85123487267&doi=10.1007%2fs11760-021-02114-x&partnerID=40&md5=3f40899248c2bbb8736b420ee8d8e4bb.

[89] X. Li et al. "Playing Against Deep-Neural-Network-Based Object Detectors: A Novel Bidirectional Adversarial Attack Approach". In: *IEEE Transactions on Artificial Intelligence* 3.1 (2022), pp. 20–28. DOI: 10.1109/TAI.2021.3107807. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85132954656&doi=10.1109%2fTAI.2021.3107807&partnerID=40&md5=3ba1e437f019c1ef0aee0583c3847360.

[90] R. Hu and T. Rui. "Stealth Attacks: A Natural and Robust Physical World Attack Against Object Detectors". In: 2022 2nd International Conference on Computer Science, Electronic Information Engineering and Intelligent Control Technology, CEI 2022. 2022, pp. 184–187. DOI: 10.1109/CEI57409.2022.9950141. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85143355020&doi=10.1109%2fCEI57409.2022.9950141&partnerID=40&md5=f445f466e2119086abee7bd29d714cc3.

[91] Z. Cai et al. "Zero-Query Transfer Attacks on Context-Aware Object Detectors". In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Vol. 2022-June. 2022, pp. 15004–15014. DOI: 10.

1109/CVPR52688.2022.01460. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85140365390&doi=10.1109%2fCVPR52688.2022.01460&partnerID=40&md5=cf31e53922a4cb296cdba7fa3040b629.

[92] J. Ye et al. “Adversarial Attack Algorithm for Object Detection Based on Improved Differential Evolution”. In: Proceedings of SPIE - The International Society for Optical Engineering. Vol. 12350. 2022. DOI: 10.1117/12.2653117. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85141878201&doi=10.1117%2f12.2653117&partnerID=40&md5=f520d236cea7a607bc7e458197b74309.

[93] K. Eykholt et al. “Physical adversarial examples for object detectors”. In: 12th USENIX Workshop on Offensive Technologies, WOOT 2018, co-located with USENIX Security 2018. 2018. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85084164612&partnerID=40&md5=0bbb45b90f37d6eddcafdaf834e8dc37.

[94] D.R. Chambers and H.A. Garza. “Physically realizable adversarial examples for convolutional object detection algorithms”. In: Proceedings of SPIE - The International Society for Optical Engineering. Vol. 10988. 2019. DOI: 10.1117/12.2520166. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85072556525&doi=10.1117%2f12.2520166&partnerID=40&md5=7b6e776aba082b24cdcd7302a925f2f1.

[95] S. Thys, W.V. Ranst, and T. Goedeme. “Fooling automated surveillance cameras: Adversarial patches to attack person detection”. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops. Vol. 2019-June. 2019, pp. 49–55. DOI: 10.1109/CVPRW.2019.00012.

[96] Y. Zhao et al. “Seeing isn’t believing: Towards more robust adversarial attack against real world object detectors”. In: Proceedings of the ACM Conference on Computer and Communications Security. 2019, pp. 1989–2004. DOI: 10.1145/3319535.3354259. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85075943736&doi=10.1145%2f3319535.3354259&partnerID=40&md5=653e1c1513fd093b416fe7345b188b97.

[97] A. Adhikari et al. “Adversarial patch camouflage against aerial detection”. In: Proceedings of SPIE - The International Society for Optical Engineering. Vol. 11543. 2020. DOI: 10.1117/12.2575907. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85096361483&doi=10.1117%2f12.2575907&partnerID=40&md5=ec2b5e1c3f9232b5afab8f373df788e2.

[98] Y. Li et al. “Exploring the vulnerability of single shot module in object detectors via imperceptible background patches”. In: 30th British Machine Vision Conference 2019, BMVC 2019. 2020.

[99] H. Li and Y. Zhao. “Fool object detectors with l0-norm patch attack”. In: CEUR Workshop Proceedings. Vol. 2881. 2020, pp. 17–20. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85108028233&partnerID=40&md5=40caa564a3a0df6e354b7927b42a32da.

[100] Y. Huang et al. *New Threats Against Object Detector with Non-local Block*. Vol. 12365 LNCS. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). Pages: 497. 2020. 481 pp. DOI: 10.1007/978-3-030-58565-5_29. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85097367619&doi=10.1007%2f978-3-030-58565-5_29&partnerID=40&md5=e3536f295e952d0f300838299858836f.

[101] Y. Zhao, H. Yan, and X. Wei. “Object hider: Adversarial patch attack against object detectors”. In: CEUR Workshop Proceedings. Vol. 2881. 2020, pp. 28–31. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85108026159&partnerID=40&md5=f4deff8f92c4610aae202b24f3039bcf.

[102] J. Bao. “Sparse adversarial attack to object detection”. In: CEUR Workshop Proceedings. Vol. 2881. 2020, pp. 32–35. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85108028449&partnerID=40&md5=59e7b253ef7fb060ce13fddfab82c91d.

[103] Z. Shi et al. “Adversarial attacks on object detectors with limited perturbations”. In: ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings. Vol. 2021-June. 2021, pp. 1375–1379. DOI: 10.1109/ICASSP39728.2021.9414125. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85115193239&doi=10.1109%2fICASSP39728.2021.9414125&partnerID=40&md5=4fe4cb71c023985afd5e930417e08b87.

[104] D. Lang et al. “Attention-Guided Digital Adversarial Patches on Visual Detection”. In: *Security and Communication Networks* 2021 (2021). ISSN: 1939-0114. DOI: 10.1155/2021/6637936.

[105] Y. Kim et al. “Extended spatially localized perturbation gan (Eslp-gan) for robust adversarial camouflage patches†”. In: *Sensors* 21.16 (2021). DOI: 10.3390/s21165323. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85111943960&doi=10.3390%2fs21165323&partnerID=40&md5=e1f579f960a86dcf540eee0442693e0a.

[106] J. Tan et al. “Legitimate Adversarial Patches: Evading Human Eyes and Detection Models in the Physical World”. In: MM 2021 - Proceedings of the 29th ACM International Conference on Multimedia. 2021, pp. 5307–5315. DOI: 10.1145/3474085.3475653. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85119356726&doi=10.1145%2f3474085.3475653&partnerID=40&md5=8cb9a0ff1deab87a3b59cd9c9170514a.

[107] Y.-C.-T. Hu et al. “Naturalistic Physical Adversarial Patch for Object Detectors”. In: Proceedings of the IEEE International Conference on Computer Vision. 2021, pp. 7828–7837. DOI: 10.1109/ICCV48922.2021.00775. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85124396883&doi=10.1109%2fICCV48922.2021.00775&partnerID=40&md5=f395882c953a2d4e3b75bd937ff47325.

[108] M. Lu et al. “Scale-adaptive adversarial patch attack for remote sensing image aircraft detection”. In: *Remote Sensing* 13.20 (2021). DOI: 10.3390/rs13204078. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85117358790&doi=10.3390%2frs13204078&partnerID=40&md5=2235d10a02872db1213d3d446363da3d.

[109] A. Zolfi et al. “The translucent patch: A physical and universal attack on object detectors”. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. 2021, pp. 15227–15236. DOI: 10.1109/CVPR46437.2021.01498. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85113700665&doi=10.1109%2fCVPR46437.2021.01498&partnerID=40&md5=a22c474a03ecd6b4a24da24a70e0ec44.

[110] Y. Wang et al. “Towards a physical-world adversarial patch for blinding object detection models”. In: *Information Sciences* 556 (2021), pp. 459–471. DOI: 10.1016/j.ins.2020.08.087. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85095585099&doi=10.1016%2fj.ins.2020.08.087&partnerID=40&md5=86ab263fbb7431affb2c5f33c0f064cc.

[111] Y. Zhang et al. “Adversarial Patch Attack on Multi-Scale Object Detection for UAV Remote Sensing Images”. In: *Remote Sensing* 14.21 (2022). DOI: 10.3390/rs14215298. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85141885578&doi=10.3390%2frs14215298&partnerID=40&md5=a314497319a7da53a34c75496d110018.

[112] F. Dong et al. “An Asterisk-shaped Patch Attack for Object Detection”. In: Proceedings - 2022 7th IEEE International Conference on Data Science in Cyberspace, DSC 2022. 2022, pp. 126–133. DOI: 10.1109/DSC55868.2022.00024. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-

85141365557&doi=10.1109%2fDSC55868.2022.00024&partnerID=40&md5=b035880f70fba089d85d264603356ec8.

[113] J. Lian et al. “Benchmarking Adversarial Patch Against Aerial Detection”. In: *IEEE Transactions on Geoscience and Remote Sensing* 60 (2022). DOI: 10.1109/TGRS.2022.3225306. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85144041372&doi=10.1109%2fTGRS.2022.3225306&partnerID=40&md5=5250c285d4a1f66d8a07cf3f246e095d.

[114] S. Sun et al. “CAMA: Class activation mapping disruptive attack for deep neural networks”. In: *Neurocomputing* 500 (2022), pp. 989–1002. DOI: 10.1016/j.neucom.2022.05.065. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85132407662&doi=10.1016%2fj.neucom.2022.05.065&partnerID=40&md5=60698f1b43630af9f32585f659a99121.

[115] Z. Cai et al. “Context-Aware Transfer Attacks for Object Detection”. In: Proceedings of the 36th AAAI Conference on Artificial Intelligence, AAAI 2022. Vol. 36. 2022, pp. 149–157. ISBN: 978-1-57735-876-3.

[116] G. Lovisotto et al. “Give Me Your Attention: Dot-Product Attention Considered Harmful for Adversarial Patch Robustness”. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Vol. 2022-June. 2022, pp. 15213–15222. DOI: 10.1109/CVPR52688.2022.01480. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85138249844&doi=10.1109%2fCVPR52688.2022.01480&partnerID=40&md5=3807602df4f55ac14ab7e9bf7929e637.

[117] H. Zhang and X. Ma. “Misleading attention and classification: An adversarial attack to fool object detection models in the real world”. In: *Computers and Security* 122 (2022). DOI: 10.1016/j.cose.2022.102876. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85136507510&doi=10.1016%2fj.cose.2022.102876&partnerID=40&md5=d2ba09e31f2ec075d95520cf956f7d16.

[118] Y. Jia et al. “Physical Adversarial Attack on a Robotic Arm”. In: *IEEE Robotics and Automation Letters* 7.4 (2022), pp. 9334–9341. DOI: 10.1109/LRA.2022.3189783. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85134247073&doi=10.1109%2fLRA.2022.3189783&partnerID=40&md5=2e04e400e638b8be7e40c6eec9d9a243.

[119] A. Toheed et al. “Physical Adversarial Attack Scheme on Object Detectors using 3D Adversarial Object”. In: 2022 2nd International Conference on Digital Futures and Transformative Technologies, ICoDT2 2022. 2022. DOI: 10.1109/ICoDT255437.2022.9787422. URL: https://www.scopus.com/inward/record.

uri?eid=2-s2.0-85133173940&doi=10.1109%2fICoDT255437.2022.9787422&partnerID=40&md5=aa77057bb3c4b036dc13c2c51e264899.

[120] A. Du et al. “Physical Adversarial Attacks on an Aerial Imagery Object Detector”. In: Proceedings - 2022 IEEE/CVF Winter Conference on Applications of Computer Vision, WACV 2022. 2022, pp. 3798–3808. DOI: 10.1109/WACV51458.2022.00385. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85121146128&doi=10.1109%2fWACV51458.2022.00385&partnerID=40&md5=7936a9e3bca69a10f594834e4098cbcc.

[121] X. Lei et al. “Using Frequency Attention to Make Adversarial Patch Powerful Against Person Detector”. In: *IEEE Access* (2022), pp. 1–1. DOI: 10.1109/ACCESS.2022.3215762. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85140747335&doi=10.1109%2fACCESS.2022.3215762&partnerID=40&md5=2d3b57ebe5cf9c97ce5eec8341f6ec51.

[122] B. Xi et al. “Bio-inspired adversarial attack against deep neural networks”. In: CEUR Workshop Proceedings. Vol. 2560. 2020, pp. 1–5. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85081554901&partnerID=40&md5=b76b88376b6f610976c2a1a757ef08ea.

[123] S. Hu et al. “CCA: Exploring the possibility of contextual camouflage attack on object detection”. In: Proceedings - International Conference on Pattern Recognition. ISSN: 1051-4651. 2020, pp. 7647–7654. ISBN: 978-1-72818-808-9. DOI: 10.1109/ICPR48806.2021.9413194.

[124] L. Huang et al. “Universal physical camouflage attacks on object detectors”. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. 2020, pp. 717–726. DOI: 10.1109/CVPR42600.2020.00080. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85094323833&doi=10.1109%2fCVPR42600.2020.00080&partnerID=40&md5=20a6b113e37a7233d9d4b62c508a2f90.

[125] X. Deng et al. “Adversarial examples with transferred camouflage style for object detection”. In: Journal of Physics: Conference Series. Vol. 1738. Issue: 1. 2021. DOI: 10.1088/1742-6596/1738/1/012130. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85101937963&doi=10.1088%2f1742-6596%2f1738%2f1%2f012130&partnerID=40&md5=60751f9ec8dab8a82bc514274013670a.

[126] Jiakai Wang et al. “Dual Attention Suppression Attack: Generate Adversarial Camouflage in Physical World”. In: *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, TN, USA: IEEE, June 2021, pp. 8561–8570. ISBN: 978-1-66544-509-2. DOI: 10.1109/

CVPR46437.2021.00846. URL: https://ieeexplore.ieee.org/document/9577412/ (visited on 02/06/2023).

[127] N. Suryanto et al. “DTA: Physical Camouflage Attacks using Differentiable Transformation Network”. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Vol. 2022-June. 2022, pp. 15284–15293. DOI: 10.1109/CVPR52688.2022.01487. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85143062887&doi=10.1109%2fCVPR52688.2022.01487&partnerID=40&md5=2f609f71eb2c236675d67cf4e52fcd7b.

[128] D. Wang et al. “FCA: Learning a 3D Full-Coverage Vehicle Camouflage for Multi-View Physical Adversarial Attack”. In: Proceedings of the 36th AAAI Conference on Artificial Intelligence, AAAI 2022. Vol. 36. 2022, pp. 2414–2422. ISBN: 978-1-57735-876-3.

[129] Y. Duan et al. “Learning Coated Adversarial Camouflages for Object Detectors”. In: IJCAI International Joint Conference on Artificial Intelligence. 2022, pp. 891–897. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85137904389&partnerID=40&md5=da1a60df484b9e45dbd6a87775fd17b8.

[130] Z. Hu et al. “Adversarial Texture for Fooling Person Detectors in the Physical World”. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Vol. 2022-June. ISSN: 1063-6919. 2022, pp. 13297–13306. ISBN: 978-1-66546-946-3. DOI: 10.1109/CVPR52688.2022.01295.

[131] Y. Li et al. *InvisibiliTee: Angle-Agnostic Cloaking from Person-Tracking Systems with a Tee*. Vol. 13531 LNCS. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). Pages: 175. 2022. 162 pp. DOI: 10.1007/978-3-031-15934-3_14. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85138007308&doi=10.1007%2f978-3-031-15934-3_14&partnerID=40&md5=424a5c8cfaa484d86693cc2e679d2f53.

[132] C. Xiao et al. “MeshAdv: Adversarial meshes for visual recognition”. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Vol. 2019-June. 2019, pp. 6891–6900. DOI: 10.1109/CVPR.2019.00706. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85078760755&doi=10.1109%2fCVPR.2019.00706&partnerID=40&md5=d22376b1aacdcaf0d974ecdf1538b7e4.

[133] Y. Huang, A.W.-K. Kong, and K.-Y. Lam. *Attacking Object Detectors Without Changing the Target Object*. Vol. 11672 LNAI. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). Pages: 15. 2019. 3 pp. DOI: 10.1007/978-3-030-29894-4_1. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85072865237&doi=10.1007%2f978-3-030-29894-4_1&partnerID=40&md5=676b781381dfc8936d412af5d770b507.

[134] S.-T. Chen et al. *ShapeShifter: Robust physical adversarial attack on faster R-CNN object detector*. Vol. 11051 LNAI. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). Pages: 68. 2019. 52 pp. DOI: 10.1007/978-3-030-10925-7_4. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85061149928&doi=10.1007%2f978-3-030-10925-7_4&partnerID=40&md5=df18bda1b970af64d699d4ab40fea5eb.

[135] Y. Huang, A.W.K. Kong, and K.-Y. Lam. "Adversarial signboard against object detector". In: 30th British Machine Vision Conference 2019, BMVC 2019. 2020. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85087328601&partnerID=40&md5=5782bba2beed13ae43f80cc52c15f68d.

[136] Y. Chen et al. "Scaling Camouflage: Content Disguising Attack Against Computer Vision Applications". In: *IEEE Transactions on Dependable and Secure Computing* (2020). DOI: 10.1109/TDSC.2020.2971601. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85079468593&doi=10.1109%2fTDSC.2020.2971601&partnerID=40&md5=40a8b3345be4f8012f3d2e4f3a0b8227.

[137] A. Hamdi, M. Muller, and B. Ghanem. "SADA: Semantic adversarial diagnostic attacks for autonomous applications". In: AAAI 2020 - 34th AAAI Conference on Artificial Intelligence. 2020, pp. 10901–10908. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85093124057&partnerID=40&md5=7c7284091f8bc891c435a9e8a9b94be1.

[138] G. Lovisotto et al. "SLAP: Improving physical adversarial examples with short-lived adversarial perturbations". In: Proceedings of the 30th USENIX Security Symposium. 2021, pp. 1865–1882. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85114521134&partnerID=40&md5=21a9d3c01a95753e8a221b245c64e09e.

[139] A.S. Rakin et al. "Deep-Dup: An adversarial weight duplication attack framework to crush deep neural network in multi-tenant FPGA". In: Proceedings of the 30th USENIX Security Symposium. 2021, pp. 1919–1936. URL: https://www.scopus.

com / inward / record . uri ? eid = 2 - s2 . 0 - 85114456844 & partnerID = 40 & md5 = 92559b66de20c9bb14014c0bc0f9e029.

[140] K.-H. Chow and L. Liu. “Perception Poisoning Attacks in Federated Learning”. In: Proceedings - 2021 3rd IEEE International Conference on Trust, Privacy and Security in Intelligent Systems and Applications, TPS-ISA 2021. 2021, pp. 146–155. ISBN: 978-1-66541-623-8. DOI: 10.1109/TPSISA52974.2021.00017.

[141] X. Ji et al. “Poltergeist: Acoustic adversarial machine learning against cameras and computer vision”. In: Proceedings - IEEE Symposium on Security and Privacy. Vol. 2021-May. 2021, pp. 160–175. DOI: 10 . 1109 / SP40001 . 2021 . 00091. URL: https : / / www . scopus . com / inward / record . uri ? eid = 2 - s2 . 0 - 85115109028&doi=10.1109%2fSP40001.2021.00091&partnerID=40&md5=ed947f2c0b3c64aa71544cca97a0cfdd.

[142] Y. Xu et al. *Universal Physical Adversarial Attack via Background Image*. Vol. 13285 LNCS. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). Pages: 14. 2022. 3 pp. DOI: 10 . 1007 / 978 - 3 - 031 - 16815 - 4 _ 1. URL: https : / / www . scopus . com / inward / record . uri ? eid = 2 - s2 . 0 - 85140469633 & doi = 10 . 1007 % 2f978 - 3 - 031 - 16815 - 4 _ 1 & partnerID = 40 & md5 = 41879ec3b8fb0c72bd4c05333045ea66.

[143] X. Wei, Y. Guo, and J. Yu. “Adversarial Sticker: A Stealthy Attack Method in the Physical World”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2022), pp. 1–1. ISSN: 0162-8828. DOI: 10 . 1109 / TPAMI . 2022 . 3176760.

[144] Z. Hu and Z. Zhong. *Towards Practical Robustness Improvement for Object Detection in Safety-Critical Scenarios*. Vol. 1271 CCIS. Communications in Computer and Information Science. Pages: 83. 2020. 66 pp. DOI: 10.1007/978-3-030-59621-7_4. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85096585937&doi=10.1007%2f978-3-030-59621-7_4&partnerID=40&md5=b0579d68918f7887f241fc39ee58b96f.

[145] A. Saha et al. “Role of spatial context in adversarial robustness for object detection”. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops. Vol. 2020-June. 2020, pp. 3403–3412. DOI: 10.1109/CVPRW50498.2020.00400. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85090166975&doi=10.1109%2fCVPRW50498.2020.00400&partnerID=40&md5=0a5255d2d5d3b38b93e59c1d5240b7cd.

[146] X. Chen et al. “Robust and accurate object detection via adversarial learning”. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. 2021, pp. 16617–16626. DOI: 10.1109/CVPR46437.2021.01635. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85104627189&doi=10.1109%2fCVPR46437.2021.01635&partnerID=40&md5=48d09c971e566f62ce2fc1da284df207.

[147] O. Poursaeed et al. “Robustness and Generalization via Generative Adversarial Training”. In: Proceedings of the IEEE International Conference on Computer Vision. 2021, pp. 15691–15700. DOI: 10.1109/ICCV48922.2021.01542. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85126983966&doi=10.1109%2fICCV48922.2021.01542&partnerID=40&md5=12bbe8f053219d229579ed450672c8fc.

[148] P.-C. Chen, B.-H. Kung, and J.-C. Chen. “Class-Aware Robust Adversarial Training for Object Detection”. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. 2021, pp. 10415–10424. DOI: 10.1109/CVPR46437.2021.01028. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85117310488&doi=10.1109%2fCVPR46437.2021.01028&partnerID=40&md5=40c744d57645abe65c84037c3e4b75a9.

[149] W. Xu, H. Huang, and S. Pan. “USING FEATURE ALIGNMENT CAN IMPROVE CLEAN AVERAGE PRECISION AND ADVERSARIAL ROBUSTNESS IN OBJECT DETECTION”. In: Proceedings - International Conference on Image Processing, ICIP. Vol. 2021-September. 2021, pp. 2184–2188. DOI: 10.1109/ICIP42928.2021.9506689. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85125586498&doi=10.1109%2fICIP42928.2021.9506689&partnerID=40&md5=3109f9ae47e351f574c83a9a3066a805.

[150] R. Canady et al. “Adversarially Robust Edge-Based Object Detection for Assuredly Autonomous Systems”. In: Proceeding - 2022 IEEE International Conference on Assured Autonomy, ICAA 2022. 2022, pp. 97–106. DOI: 10.1109/ICAA52185.2022.00021. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85130042789&doi=10.1109%2fICAA52185.2022.00021&partnerID=40&md5=73e5df1e04e7cd7cf39e8c33966b21ea.

[151] W. Xu et al. “ROBUST AND ACCURATE OBJECT DETECTION VIA SELF-KNOWLEDGE DISTILLATION”. In: Proceedings - International Conference on Image Processing, ICIP. ISSN: 1522-4880. 2022, pp. 91–95. ISBN: 978-1-66549-620-9. DOI: 10.1109/ICIP46576.2022.9898031.

[152] Y. Sui et al. “Towards Robust Detection and Segmentation Using Vertical and Horizontal Adversarial Training”. In: Proceedings of the International

Joint Conference on Neural Networks. Vol. 2022-July. 2022. DOI: 10.1109/IJCNN55064.2022.9892759. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85140744257&doi=10.1109%2fIJCNN55064.2022.9892759&partnerID=40&md5=7e97f066c5ae62b267aedbdbf28e990e.

[153] W. Xue et al. “A Cascade Defense Method for Multidomain Adversarial Attacks under Remote Sensing Detection”. In: *Remote Sensing* 14.15 (2022). DOI: 10.3390/rs14153559. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85137042404&doi=10.3390%2frs14153559&partnerID=40&md5=1266fbd5093257b406406f071a270738.

[154] J. Cheng and V. Hu. “Defending Convolutional Neural Network-Based Object Detectors against Adversarial Attacks”. In: 2020 9th IEEE Integrated STEM Education Conference, ISEC 2020. Vol. 2020-January. 2020. DOI: 10.1109/ISEC49744.2020.9397815. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85104876174&doi=10.1109%2fISEC49744.2020.9397815&partnerID=40&md5=f67f7acaf7f826fd895fad059c004588.

[155] P.-Y. Chiang et al. “Detection as regression: Certified object detection by median smoothing”. In: Advances in Neural Information Processing Systems. Vol. 2020-December. 2020. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85108414178&partnerID=40&md5=8fe396655e82c42a687f79b68718771d.

[156] G. Zhou et al. “Information distribution based defense against physical attacks on object detection”. In: 2020 IEEE International Conference on Multimedia and Expo Workshops, ICMEW 2020. 2020. DOI: 10.1109/ICMEW46912.2020.9105983. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85091755486&doi=10.1109%2fICMEW46912.2020.9105983&partnerID=40&md5=abe1d44ffaf123aaee4a9cec2fa97d78.

[157] J. Bao et al. *Improving Adversarial Robustness of Detector via Objectness Regularization*. Vol. 13022 LNCS. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). Pages: 262. 2021. 252 pp. DOI: 10.1007/978-3-030-88013-2_21. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85118948780&doi=10.1007%2f978-3-030-88013-2_21&partnerID=40&md5=9e1f2760177ac6083851bf2e0931b12d.

[158] Y. Yu et al. “Defending Person Detection Against Adversarial Patch Attack by Using Universal Defensive Frame”. In: *IEEE Transactions on Image Processing* 31 (2022), pp. 6976–6990. ISSN: 1057-7149. DOI: 10.1109/TIP.2022.3217375.

[159] X. Wang et al. “Using bilateral filtering and autoencoder to defend against adversarial attacks for object detection”. In: *Journal of Electronic Imaging* 31.4 (2022). DOI: 10.1117/1.JEI.31.4.043040. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85142268875&doi=10.1117%2f1.JEI.31.4.043040&partnerID=40&md5=1d7bf4cc3b16ce122c11c86147dd55b8.

[160] Y. Zhang, B. Dong, and F. Heide. *All You Need Is RAW: Defending Against Adversarial Attacks with Camera Image Pipelines*. Vol. 13679 LNCS. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). Pages: 343. 2022. 323 pp. DOI: 10.1007/978-3-031-19800-7_19. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85142681050&doi=10.1007%2f978-3-031-19800-7_19&partnerID=40&md5=804840d2c97fab8b7526e33122a619b8.

[161] F. Alamri, S. Kalkan, and N. Pugeault. “Transformer-encoder detector module: Using context to improve robustness to adversarial attacks on object detection”. In: Proceedings - International Conference on Pattern Recognition. 2020, pp. 9577–9584. DOI: 10.1109/ICPR48806.2021.9413344. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85108389917&doi=10.1109%2fICPR48806.2021.9413344&partnerID=40&md5=4d1174ac494f9f7b3b78537aeee1c792.

[162] K.-H. Chow and L. Liu. “Robust Object Detection Fusion against Deception”. In: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2021, pp. 2703–2713. DOI: 10.1145/3447548.3467121. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85114914389&doi=10.1145%2f3447548.3467121&partnerID=40&md5=be60fef9c69fc7ec3ae6e3a09461a8c6.

[163] A. Amirkhani and M.P. Karimi. “Adversarial defenses for object detectors based on Gabor convolutional layers”. In: *Visual Computer* 38.6 (2022), pp. 1929–1944. DOI: 10.1007/s00371-021-02256-6. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85111147521&doi=10.1007%2fs00371-021-02256-6&partnerID=40&md5=9b9bad956a5e949587aa6ed2f282711a.

[164] Z. Dong, P. Wei, and L. Lin. *Adversarially-Aware Robust Object Detector*. Vol. 13669 LNCS. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). Pages: 313. 2022. 297 pp. DOI: 10.1007/978-3-031-20077-9_18. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85142752951&doi=10.1007%2f978-3-031-20077-9_18&partnerID=40&md5=57e51ea6ac637f11e279b56440c0dfae.

[165] R. Searle and P. Gururaj. “Establishing security and trust for object detection and classification with confidential AI”. In: Proceedings of SPIE - The International Society for Optical Engineering. Vol. 12113. ISSN: 0277-786X. 2022. ISBN: 978-1-5106-5102-9. DOI: 10.1117/12.2618303.

[166] C. Kyrkou et al. “Towards artificial-intelligence-based cybersecurity for robustifying automated driving systems against camera sensor attacks”. In: Proceedings of IEEE Computer Society Annual Symposium on VLSI, ISVLSI. Vol. 2020-July. 2020, pp. 476–481. DOI: 10.1109/ISVLSI49217.2020.00-11. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85090425314&doi=10.1109%2fISVLSI49217.2020.00-11&partnerID=40&md5=33c1546f5e72fedf309b0e5b262522ae.

[167] S. Li et al. *Connecting the Dots: Detecting Adversarial Perturbations Using Context Inconsistency*. Vol. 12368 LNCS. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). Pages: 413. 2020. 396 pp. DOI: 10.1007/978-3-030-58592-1_24. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85097396278&doi=10.1007%2f978-3-030-58592-1_24&partnerID=40&md5=ede203495366edd2d2e7a37dd1d31415.

[168] W. Chai, Y. Lu, and S. Velipasalar. “WEIGHTED AVERAGE PRECISION: ADVERSARIAL EXAMPLE DETECTION FOR VISUAL PERCEPTION OF AUTONOMOUS VEHICLES”. In: Proceedings - International Conference on Image Processing, ICIP. Vol. 2021-September. 2021, pp. 804–808. DOI: 10.1109/ICIP42928.2021.9506613. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85125583091&doi=10.1109%2fICIP42928.2021.9506613&partnerID=40&md5=3e912687014ecc3a1f358de73e09943d.

[169] M. Yin et al. “Exploiting Multi-Object Relationships for Detecting Adversarial Attacks in Complex Scenes”. In: Proceedings of the IEEE International Conference on Computer Vision. 2021, pp. 7838–7847. DOI: 10.1109/ICCV48922.2021.00776. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85114153206&doi=10.1109%2fICCV48922.2021.00776&partnerID=40&md5=074638483e14ae9a5fad8e427f92dc0e.

[170] C. Xiang and P. Mittal. “DetectorGuard: Provably Securing Object Detectors against Localized Patch Hiding Attacks”. In: Proceedings of the ACM Conference on Computer and Communications Security. 2021, pp. 3177–3196. DOI: 10.1145/3460120.3484757. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85114264581&doi=10.1145%2f3460120.3484757&partnerID=40&md5=3342cbf6489d54a030fa7409dbd32fce.

[171] P.-H. Chiang, C.-S. Chan, and S.-H. Wu. “Adversarial Pixel Masking: A Defense against Physical Attacks for Pre-trained Object Detectors”. In: MM 2021 - Proceedings of the 29th ACM International Conference on Multimedia. 2021, pp. 1856–1865. DOI: 10.1145/3474085.3475338. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85119347536&doi=10.1145%2f3474085.3475338&partnerID=40&md5=69cc763d1e4284e29adb7eaebe9ccc79.

[172] J. Liu et al. “Segment and Complete: Defending Object Detectors against Adversarial Patch Attacks with Robust Patch Detection”. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Vol. 2022-June. 2022, pp. 14953–14962. DOI: 10.1109/CVPR52688.2022.01455. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85139542265&doi=10.1109%2fCVPR52688.2022.01455&partnerID=40&md5=d66252db79dc652b6b92428c9be50773.

[173] Z. Xu et al. “LanCeX: A Versatile and Lightweight Defense Method against Condensed Adversarial Attacks in Image and Audio Recognition”. In: *ACM Transactions on Embedded Computing Systems* 22.1 (2022). DOI: 10.1145/3555375. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85146423456&doi=10.1145%2f3555375&partnerID=40&md5=7f7854f04bbfc4d779e63e15694832c7.

[174] Kaiming He et al. *Mask R-CNN*. Jan. 24, 2018. arXiv: 1703.06870[cs]. URL: http://arxiv.org/abs/1703.06870 (visited on 04/27/2023).

[175] Tsung-Yi Lin et al. *Focal Loss for Dense Object Detection*. version: 2. Feb. 7, 2018. DOI: 10.48550/arXiv.1708.02002. arXiv: 1708.02002[cs]. URL: http://arxiv.org/abs/1708.02002 (visited on 04/27/2023).

[176] Sikkerhetsloven. *Lov om nasjonal sikkerhet*. LOV-2018-06-01-24. Lovdata, Jan. 1, 2019. URL: https://lovdata.no/dokument/NL/lov/2018-06-01-24 (visited on 05/03/2023).

[177] A. Braunegg et al. *APRICOT: A Dataset of Physical Adversarial Attacks on Object Detection*. Vol. 12366 LNCS. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). Pages: 50. 2020. 35 pp. DOI: 10.1007/978-3-030-58589-1_3. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85097436278&doi=10.1007%2f978-3-030-58589-1_3&partnerID=40&md5=43ae9a655d4cfae740f3ccbf9d9b330b.

[178] J. Zhang et al. “Enhance Domain-Invariant Transferability of Adversarial Examples via Distance Metric Attack”. In: *Mathematics* 10.8 (2022). DOI: 10.3390/math10081249. URL: https://www.scopus.com/inward/record.uri?eid=2-

s2.0-85128781732&doi=10.3390%2fmath10081249&partnerID=40&md5=aeadfe38fb2670f266e08e706d8dc6d7.

[179] Jean-Marie Henckaerts et al., eds. *Customary international humanitarian law*. Cambridge ; New York: Cambridge University Press, 2005. 2 pp. ISBN: 0-521-80888-X.

[180] Z. Zhong, Z. Hu, and X. Chen. "Quantifying DNN Model Robustness to the Real-World Threats". In: Proceedings - 50th Annual IEEE/IFIP International Conference on Dependable Systems and Networks, DSN 2020. 2020, pp. 150–157. DOI: 10.1109/DSN48063.2020.00033. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85090409436&doi=10.1109%2fDSN48063.2020.00033&partnerID=40&md5=7d7e7d611368083371fd3f38303ba628.

[181] A. Mahmoud et al. "PyTorchFI: A Runtime Perturbation Tool for DNNs". In: Proceedings - 50th Annual IEEE/IFIP International Conference on Dependable Systems and Networks, DSN-W 2020. 2020, pp. 25–31. DOI: 10.1109/DSN-W50199.2020.00014. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85092635000&doi=10.1109%2fDSN-W50199.2020.00014&partnerID=40&md5=874bf9910d364ddb1fc9d20717758bcd.

[182] T. Du et al. "DetectS ec: Evaluating the robustness of object detection models to adversarial attacks". In: *International Journal of Intelligent Systems* 37.9 (2022), pp. 6463–6492. DOI: 10.1002/int.22851. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85124389749&doi=10.1002%2fint.22851&partnerID=40&md5=5386cf9d6a72c3a628ae4f3bd8815a2a.

[183] S. Vellaichamy et al. "DetectorDetective: Investigating the Effects of Adversarial Examples on Object Detectors". In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Vol. 2022-June. 2022, pp. 21452–21459. DOI: 10.1109/CVPR52688.2022.02082. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85141810563&doi=10.1109%2fCVPR52688.2022.02082&partnerID=40&md5=11f48780b3b2cae77cc2b84e4d4767cf.

Appendix A

Tools to Aid Future Research on Adversarial Attacks and Defenses

This appendix presents tools to aid future research on adversarial attacks and defenses in object detection, shown in Table A.1. These papers were found during the systematic literature review on adversarial attacks and defenses in object detection, described in Section 4.1 starting on Page 23. These tools got placed in the appendix rather than the results as they are neither adversarial attack- nor defense methods. However, these papers were considered relevant as they introduce tools that can be useful for industry use cases and for future research on adversarial attacks- and defenses.

The datasets can be used for, among other things, adversarial training purposes and attack generation. They can also be used for research towards the understanding of adversarial attacks, where visual tools also can play an important role. Benchmarks and robustness evaluation frameworks are especially useful for evaluating object detectors' robustness against adversarial attacks - useful for both researchers and practitioners in the industry.

Article	Year	Tool	Target Attack Method
APRICOT [177]	2020	Dataset	Physical Adversarial Patch
AOCO [24]			Adversarial Example
Zhong et al. [180]		Benchmark	
PyTorchFI [181]		Robustness Evaluation	
DetectSec [182]	2022		
DetectorDetective [183]		Visual Tool	

Table A.1: **Tools for Adversarial Attacks and Defenses in Object Detection.** AE: Adversarial Example, AP: Adversarial Patch.

Appendix B

Other Data Types and Multi-Sensor Object Detection

Based on the inclusion and exclusion criteria, this literature review focused solely on images and videos for adversarial attacks and defenses in object detection. However, incorporating other data types like LiDAR point clouds and infrared imagery could prove advantageous for unmanned military vehicles. Furthermore, integrating data from multiple sensors can significantly enhance the system's overall performance in multi-sensor object detection. If this is to be implemented in the future, preliminary research should be done to assess the threat of adversarial attacks.

To avoid losing valuable information, papers on object detection using other data types and multi-sensor object detection were collected separately for future research. The following list contains the papers found during the systematic literature search for this thesis:

[1] A. Lehner et al., '3D-VField: Adversarial Augmentation of Point Clouds for Domain Generalization in 3D Object Detection', presented at the Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2022, pp. 17274–17283. doi: 10.1109/CVPR52688.2022.01678.

[2] M. Abdelfattah, K. Yuan, Z. J. Wang, and R. Ward, 'Adversarial Attacks on Camera-LiDAR Models for 3D Car Detection', presented at the IEEE International Conference on Intelligent Robots and Systems, 2021, pp. 2189–2194. doi: 10.1109/IROS51168.2021.9636638.

[3] I. Sobh, A. Hamed, V. R. Kumar, and S. Yogamani, 'Adversarial Attacks onMulti-task Visual Perception for Autonomous Driving', Journal of Imaging Science and Technology, vol. 65, no. 6, 2021, doi: 10.2352/J.ImagingSci.Technol.2021.65.6.060408.

[4] X. Wang, M. Cai, F. Sohel, N. Sang, and Z. Chang, 'Adversarial point cloud perturbations against 3D object detection in autonomous driving systems', Neurocomputing, vol. 466, pp. 27–36, 2021, doi: 10.1016/j.neucom.2021.09.027.

[5] M. Cai, N. Sang, X. Wang, and J. Zhang, 'Adversarial point cloud perturbations to attack deep object detection models', presented at the Proceedings - 2020 IEEE 22nd International Conference on High Performance Computing and Communications, IEEE 18th International Conference on Smart City and IEEE 6th International Conference on Data Science and Systems, HPCC-SmartCity-DSS 2020, 2020, pp. 1042–1049. doi: 10.1109/HPCC-SmartCity-DSS50907.2020.00140.

[6] S. Wang, T. Wu, A. Chakrabarti, and Y. Vorobeychik, 'Adversarial Robustness of Deep Sensor Fusion Models', presented at the Proceedings - 2022 IEEE/CVF Winter Conference on Applications of Computer Vision, WACV 2022, 2022, pp. 1371–1380. doi: 10.1109/WACV51458.2022.00144.

[7] Y. Cao et al., 'Adversarial sensor attack on LiDAR-based perception in autonomous driving', presented at the Proceedings of the ACM Conference on Computer and Communications Security, 2019, pp. 2267–2281. doi: 10.1145/3319535.3339815.

[8] . Bayer, D. Münch, and M. Arens, 'APMD: Adversarial Pixel Masking Derivative for Multispectral Object Detectors', presented at the Proceedings of SPIE - The International Society for Optical Engineering, 2022. doi: 10.1117/12.2637977.

[9] C. Chen and T. Huang, 'Camdar-adv: Generating adversarial patches on 3D object', International Journal of Intelligent Systems, vol. 36, no. 3, pp. 1441–1453, 2021, doi: 10.1002/int.22349.

[10] Y. Zhu, C. Miao, T. Zheng, F. Hajiaghajani, L. Su, and C. Qiao, 'Can We Use Arbitrary Objects to Attack LiDAR Perception in Autonomous Driving?', presented at the Proceedings of the ACM Conference on Computer and Communications Security, 2021, pp. 1945–1960. doi: 10.1145/3460120.3485377.

[11] Q. Sun, A. A. Rao, X. Yao, B. Yu, and S. Hu, 'Counteracting Adversarial Attacks in Autonomous Driving', presented at the IEEE/ACM International Conference on Computer-Aided Design, Digest of Technical Papers, ICCAD, 2020. doi: 10.1145/3400302.3415758.

[12] F. Spasiano, G. Gennaro, and S. Scardapane, 'Evaluating Adversarial Attacks and Defences in Infrared Deep Learning Monitoring Systems', presented at the

Proceedings of the International Joint Conference on Neural Networks, 2022. doi: 10.1109/IJCNN55064.2022.9891997.

[13] J. Zhang, Y. Lou, J. Wang, K. Wu, K. Lu, and X. Jia, 'Evaluating Adversarial Attacks on Driving Safety in Vision-Based Autonomous Vehicles', IEEE Internet of Things Journal, vol. 9, no. 5, pp. 3443–3456, 2022, doi: 10.1109/JIOT.2021.3099164.

[14] X. Zhu, X. Li, J. Li, Z. Wang, and X. Hu, 'Fooling Thermal Infrared Pedestrian Detectors in Real World Using Small Bulbs', presented at the 35th AAAI Conference on Artificial Intelligence, AAAI 2021, 2021, pp. 3616–3624.

[15] H. Wang, H. Shen, B. Zhang, Y. Wen, and D. Meng, Generating Adversarial Point Clouds on Multi-modal Fusion Based 3D Object Detection Model, vol. 12918 LNCS. in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 12918 LNCS. 2021, p. 203. doi: 10.1007/978-3-030-86890-1_11.

[16] X. Zhu, Z. Hu, S. Huang, J. Li, and X. Hu, 'Infrared Invisible Clothing: Hiding from Infrared Detectors at Multiple Angles in Real World', presented at the Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2022, pp. 13307–13316. doi: 10.1109/CVPR52688.2022.01296.

[17] Y. Cao et al., 'Invisible for both Camera and LiDAR: Security of multi-sensor fusion based perception in autonomous driving under physical-world attacks', presented at the Proceedings - IEEE Symposium on Security and Privacy, 2021, pp. 176–194. doi: 10.1109/SP40001.2021.00076.

[18] T. Kim, H. J. Lee, and Y. M. Ro, 'MAP: MULTISPECTRAL ADVERSARIAL PATCH TO ATTACK PERSON DETECTION', presented at the ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings, 2022, pp. 4853–4857. doi: 10.1109/ICASSP43922.2022.9747896.

[19] Z. Xiong, H. Xu, W. Li, and Z. Cai, 'Multi-Source Adversarial Sample Attack on Autonomous Vehicles', IEEE Transactions on Vehicular Technology, vol. 70, no. 3, pp. 2822–2835, 2021, doi: 10.1109/TVT.2021.3061065.

[20] B. Liu, Y. Guo, J. Jiang, J. Tang, and W. Deng, 'Multi-view Correlation based Black-box Adversarial Attack for 3D Object Detection', presented at the Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2021, pp. 1036–1044. doi: 10.1145/3447548.3467432.

[21] K. Yun, T. Lu, A. Huyen, P. Hammer, and P. Wang, 'Neurosymbolic hybrid approach to driver collision warning', presented at the Proceedings of SPIE - The International Society for Optical Engineering, 2022. doi: 10.1117/12.2620209.

[22] Z. Cheng et al., Physical Attack on Monocular Depth Estimation with Optimal Adversarial Patches, vol. 13698 LNCS. in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 13698 LNCS. 2022, p. 532. doi: 10.1007/978-3-031-19839-7_30.

[23] J. Tu et al., 'Physically realizable adversarial examples for lidar object detection', presented at the Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2020, pp. 13713–13722. doi: 10.1109/CVPR42600.2020.01373.

[24] W. Jiang, Z. He, J. Zhan, W. Pan, and D. Adhikari, 'Research progress and challenges on application-driven adversarial examples: A survey', ACM Transactions on Cyber-Physical Systems, vol. 5, no. 4, 2021, doi: 10.1145/3470493.

[25] K. Yang, T. Tsai, H. Yu, M. Panoff, T.-Y. Ho, and Y. Jin, 'Robust Roadside Physical Adversarial Attack against Deep Learning in Lidar Perception Modules', presented at the ASIA CCS 2021 - Proceedings of the 2021 ACM Asia Conference on Computer and Communications Security, 2021, pp. 349–362. doi: 10.1145/3433210.3453106.

[26] W. Park, N. Liu, Q. A. Chen, and Z. M. Mao, 'SENSOR ADVERSARIAL TRAITS: ANALYZING ROBUSTNESS OF 3D OBJECT DETECTION SENSOR FUSION MODELS', presented at the Proceedings - International Conference on Image Processing, ICIP, 2021, pp. 484–488. doi: 10.1109/ICIP42928.2021.9506183.

[27] C. You, Z. Hau, and S. Demetriou, 'Temporal Consistency Checks to Detect LiDAR Spoofing Attacks on Autonomous Vehicle Perception', presented at the MAISP 2021 - Proceedings of the 2021 1st Workshop on Security and Privacy for Mobile AI, 2021, pp. 13–18. doi: 10.1145/3469261.3469406.

[28] Y. Zhang et al., 'Towards Backdoor Attacks against LiDAR Object Detection in Autonomous Driving', presented at the SenSys 2022 - Proceedings of the 20th ACM Conference on Embedded Networked Sensor Systems, 2022, pp. 533–547. doi: 10.1145/3560905.3568539.

[29] J. Sun, Y. Cao, Q. A. Chen, and Z. Morley Mao, 'Towards robust LiDAR-based perception in autonomous driving: General black-box adversarial sensor

attack and countermeasures', presented at the Proceedings of the 29th USENIX Security Symposium, 2020, pp. 877–894.

[30] M. Abdelfattah, K. Yuan, Z. Jane Wang, and R. Ward, 'TOWARDS UNIVERSAL PHYSICAL ATTACKS ON CASCADED CAMERA-LIDAR 3D OBJECT DETECTION MODELS', presented at the Proceedings - International Conference on Image Processing, ICIP, 2021, pp. 3592–3596. doi: 10.1109/ICIP42928.2021.9506016.

[31] H. Kim and C. Lee, 'Upcycling adversarial attacks for infrared object detection', Neurocomputing, vol. 482, pp. 1–13, 2022, doi: 10.1016/j.neucom.2022.01.090.

[32] R. Dollahite, K. Wang, K. Li, Y. Zhang, and Z. Zhang, 'Verifying Adversarial Robustness of 3D Object Detectors for Autonomous Vehicles', presented at the 2022 IEEE MIT Undergraduate Research Technology Conference, URTC 2022, 2022. doi: 10.1109/URTC56832.2022.10002253.