# Table of Contents

# List of Figures

# List of Tables

**Abstract**

This thesis investigates the development and ethical implications of on-device processing person localization systems in public spaces, with a focus on maintaining high standards of privacy. As modern surveillance technologies become increasingly pervasive, there is a pressing need to balance technological advancements with privacy concerns. This thesis demonstrates the feasibility of using on-device processing for object detection on images to significantly enhance data privacy and security while providing actionable insights into visitor behavior in public venues.

The results of this study found that the accuracy of the investigated object detection models were significantly improved when evaluated with deployment-specific images instead of images in generic datasets. The object detection model *YOLOv9* achieved an average precision score of 91.3% on a deployment-specific dataset collected for this thesis project, substantially higher than the 55.6% average precision the same model achieves on the COCO validation set (C.-Y. Wang and Liao, 2024). This highlights the importance of contextually relevant images in the evaluation of model precisions in practice. Furthermore, the research delves into various ethical considerations, discussing the potential risks associated with mass surveillance and the importance of maintaining a balanced approach that respects individual privacy rights while harnessing the benefits of technological innovations.

The principal contributions of this thesis include:

1. The *FIMUS* dataset was gathered, labeled, and released.
2. A person localization system was developed and deployed at the Fisheries and Maritime Museums (FIMUS) aquarium in Esbjerg, Denmark. The resulting data was visualized through heatmaps and bar charts, and its relevance to museum curators and administrators was discussed.
3. Images of varying degrees of quality and relevance to the deployment scenario were utilized to fine-tune object detection models, assessing the impact of data quality and relevance on accuracy of the fine-tuned models.
4. A comprehensive review and discussion of literature regarding privacy, ethics, and the technology.

The practical implementation of the system at the *Fisheries and Maritime Museums* aquarium provides a real-world application of the discussed concepts, showcasing the potential of implementing on-device processing for person localization systems to improve operational efficiencies and visitor experiences without compromising privacy. The thesis also explores future research directions, including enhancing dataset diversity, exploring advanced privacy-preserving techniques, and expanding the technology's application beyond person detection to include features such as fall detection and visitor interaction tracking.

In conclusion, this thesis demonstrates that on-device processing effectively preserves individual privacy, aligning with data protection regulations. It contributes insights into the development and implementation of privacy-preserving person localization systems, addressing both their technical and social dimensions. This work provides a comprehensive understanding of the challenges and opportunities associated with these systems. To further enhance the relevance and applicability of the locational data, collaboration with museum administrators is essential to align the system's capabilities with their specific needs. This approach ensures that technological advancements not only uphold societal well-being but also protect individual privacy, advocating for ongoing innovation guided by rigorous ethical standards and practical engagements.

**Sammendrag**

Denne avhandlingen undersøker utviklingen og de etiske implikasjonene av *on-device* prosesserende personlokaliseringssystemer i offentlige rom, med fokus på opprettholdelse av personvern. Ettersom moderne overvåkningsteknologier stadig blir mer utbredt, er det et presserende behov for å balansere teknologiske fremskritt med personvernhensyn. Denne avhandlingen viser at det er mulig å benytte *on-device* prosessering i objektdetektesjon på bilder til å øke datasikkerhet og personvern, samtidig som den gir nyttig innsikt i besøkendes adferd i kulturelle institusjoner.

Resultatene av studien fant at presisjonen til de undersøkte objektdeteksjonsmodellene forbedres betydelig når de ble evaluert med bilder spesifikt for utplasseringen i stedet for bilder fra generiske datasett. Objektdetekteringsmodellen *YOLOv9* oppnådde en gjennomsnitlig presisjon på 91.3% på datasettet fra prosjektet i denne avhandlingen, hvilket er mye enn 55.6% som er det den samme modellen oppnår på *COCO* valideringssettet (C.-Y. Wang and Liao, 2024). Dette understreker viktigheten av kontekstuelt relevante bilder i vurderingen av modellenes nøyaktighet i praksis. Videre går forskningen inn på ulike etiske betraktninger, og diskuterer potensielle risikoer knyttet til massesurveillance og viktigheten av å opprettholde en balansert tilnærming som respekterer individuelle rettigheter til privatliv samtidig som man utnytter fordelene ved teknologiske innovasjoner.

De viktigste bidragene fra denne avhandlingen inkluderer:

1. Datasettet *FIMUS* ble samlet, merket og utgitt.
2. Et personlokaliseringssystem ble utviklet og implementert i Fiskeri og Søfartsmuseets akvarium i Esbjerg, Danmark. Resulterende data ble visualisert, og dataens relevans for museumsforvaltere og administratorer diskutert.
3. Bilder av ulik kvalitet og relevanse for implementeringsscenarioet ble brukt til å finjustere objektdeteksjonsmodeller, for å vurdere påvirkningsgraden av kvalitet og relevanse av bilder på nøyaktigheten av de finjusterte modellene.
4. En omfattende gjennomgang og diskusjon av litteratur om personvern, etikk og teknologi.

Den praktiske implementeringen av systemet i *Fiskeri- og Søfartsmuseet* gir en virkelig anvendelse av de diskuterte konseptene, og viser potensialet av å implementere *on-device* prosessering i personlokaliseringssystemer til å forbedre operasjonell effektivitet og brukeropplevelser uten å gå på kompromiss med personvern. Avhandlingen utforsker også fremtidige forskningsretninger, inkludert å forbedre mangfoldet i datasett, å utforske avanserte teknikker for personvernbevarelse, og å utvide teknologiens anvendelse utover persondeteksjon til å inkludere funksjoner som deteksjon av fall og sporing av besøkendes interaksjoner.

Den prinsipielle konklusjonen som denne avhandlingen viser, er at *on-device* prosessering effektivt bevarer individets personvern, i tråd med regelverk for databeskyttelse. Den bidrar med innsikter i utviklingen og implementeringen av personlokaliseringssystemer som bevarer personvernet, og tar for seg både tekniske og etiske aspekter ved disse systemene. Avhandlingen gir en omfattende forståelse av utfordringene og mulighetene som er forbundet med disse systemene. For å ytterligere forbedre relevansen og anvendeligheten av lokaliseringsdataen, er samarbeid med museumsadministratorer essensielt for å tilpasse systemets kapabiliteter til deres spesifikke behov. This thesis supports the notion that technological advancements should not only enhance societal well-being but also protect individual privacy, and argues that continuous innovation must be guided by strict ethical standards and guidelines.

# 1 Introduction

Recent advancements in machine learning (ML) have notably enhanced models' capabilities to extract valuable information from images. This is particularly relevant for indoor positioning systems (IPS) with extended capabilities to detect falls or signals for help. Traditionally, various non-visual sensors such as wearable devices, passive infrared sensors (PIR), radio-frequency identification (RFID), and light detection and ranging (LIDAR) have been utilized in IPS. These technologies preserve privacy effectively as they are mostly incapable of identifying personal data. However, for future applications requiring detailed context and capabilities, these sensors fall short where visual sensors excel.

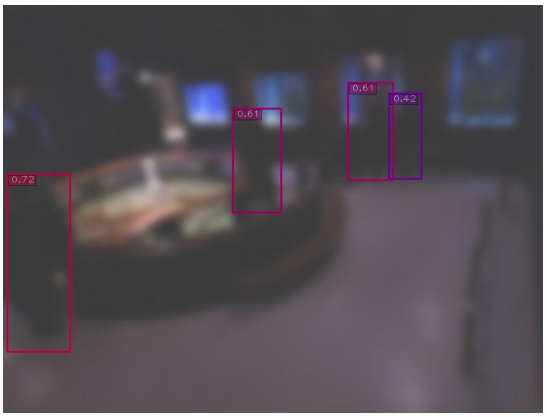
The emergence of automated *on-device* processing marks a significant progression in privacy preservation of visual sensors in IPS. With on-device processing, images no longer need to be analyzed manually or transmitted over the internet, allowing for the creation of more advanced and comprehensive systems. As highlighted by Bibbo et al., integrating indoor positioning systems (IPS) with human activity recognition (HAR) is essential to detect incidents like falls or injuries, especially among elderly individuals living alone (2022).

This thesis explores the development and deployment of a privacy-preservant person locatization system, particularly advocating for on-device processing to mitigate privacy concerns. As part of the on-device processing, sensitive data, such as images, are either deleted or obfuscated immediately after inference, complicating analysis validation but significantly enhancing privacy.

Typically, machine learning models for person detection are pretrained on large, generic datasets and validated on similar datasets, which may not accurately represent specific deployment scenarios. This thesis addresses this gap by collecting a specialized dataset in a typical dark-lit indoor museum/aquarium setting and evaluating various detectors fine-tuned to this environment. A month-long live experiment was conducted at the Fisheries and Maritime Museums (FIMUS) aquarium to assess these systems' real-world applicability. During this experiment, some images were blurred and retained to verify the model's inferences, enhancing our understanding of their practical utility and limitations.

Figure 1a displays an example image from the collected FIMUS dataset, while Figure 1b shows an image from the live experiment that was kept for verification purposes.

(a) Example Image From the FIMUS Dataset.




(b) Example Image From the Live Detections.

Figure 1: Images From The Fisheries and Maritime Museum (FIMUS)

## 1.1 Problem Description

Recent advancements in computer vision and machine learning have transformed the field of person localization in public spaces. Traditionally, surveillance systems rely on centralized processing, where video feeds are sent to a remote server for analysis. This method poses significant privacy risks as it exposes sensitive information to potential interception during transmission. This thesis advocates for a shift towards *on-device* processing, where analytics are performed locally on edge devices, thus enhancing privacy by eliminating the need to transmit raw video data. Another aspect of on-device processing is to algorithmically analyze the video feed by the use of machine learning, instead of humans.

Additionally, the efficacy of object detection models is commonly measured using generic datasets like COCO, which may not accurately represent the specific conditions and image characteristics of a real-world deployment environment. Such discrepancies can result in models performing sub-optimally in actual settings, where variables like lighting, camera angle viewpoint, scale, rotation, and object occlusion play crucial roles. The quality and relevance of the data used to test and train a model for deployment should be as similar to the deployment scenario as possible. However, assessing the impact of dataset relevancy and quality on model accuracy presents substantial challenges, with limited research available on this subject. Consumers are thus left in the dark by the researchers in the field with regards to how much effort should be put into finding the best suited test dataset for model testing and fine-tuning.

## 1.2 Scope

To demonstrate the feasibility and effectiveness of on-device person localization, two devices were deployed at the *Fisheries and Maritime Museum*[1] (FIMUS) aquarium in Esbjerg, Denmark. The indoor, low-light conditions of the aquarium presented specific challenges that the deployment aimed to address. The deployment camera angle viewpoint on the persons in the area was different than most images in the COCO dataset that the models were pre-trained on. For this reason, a dataset consisting of 3,397 images was collected and labeled, providing a foundation for fine-tuning several object detector models to better fit the deployment scenario, and to investigate the impacts of dataset relevancy on the model performances. The dataset was collected using varying and fixed camera configurations, creating a dataset capable of investigation of data quality (different camera configurations) on the model accuracies. Figure 1 showcases an example image from this dataset. During the final month of the project, a device was deployed to anonymously and automatically detect the positions of persons[2] in the room every 2 minutes during opening hours. The data was visualized through heatmaps and bar charts of peak visitation times of the day.

The scope of this thesis is twofold: firstly, the thesis details a demonstation of the implementation of a privacy-preserving person localization system; secondly, the thesis assesses the validity of object detection model performances across general and specific datasets to evaluate the real-world impacts of scientific advancements. This dualistic approach not only highlights the practical application of privacy technologies in object detection but also investigates how different data environments affect model efficacy.

The inclusion of a previously implemented system serves as a practical foundation for this thesis, shortening the theoretical to practical implementation pathway. By utilizing the images from the specific deployment scenario, thus specializing our solution to the problem, the thesis project aims to achieve superior results at the expense of reduced generalizability. However, the general nature of the thesis ensures that the findings and methodologies can be adapted and applied to other problems with similar characteristics.

This thesis is selectively deep on topics such as privacy, privacy preservation in images, and performance metrics of object detectors. These areas are emphasized due to their critical relevance and the necessity of a fundamental understanding of these topics to grasp the project's core objectives.

[1] In original language: Fiskeri og Søfartsmuseet

[2] *People: Where individual persons, or a number of such, are intended, this word should be discarded in favor of persons.(Vizetelly, 1920).*

While other subjects like object detector model architecture, edge devices, challenges in object detection, security, and decision support systems also present interesting avenues of exploration, they are not central to the thesis' primary aims.

This thesis is intended for a diverse reader group, including technical engineers, technical leaders, non-technical social studies scholars, and policymakers engaged in crafting regulations for object detection technologies. This results in a broad scope of topics covered in the thesis, ranging from technical discussions on machine learning models to ethical considerations for the deployment of person localization systems in public spaces. The project of this thesis also spanned several technological disciplines. It required research, development, and effort in edge-device deployment, machine learning, and data science. The scope was limited to manage the workload effectively, see the Limitations section below.

## 1.3 Limitations

**Secure Control of Device** A dataset was built of consenting individuals in an aquarium which was part of a larger museum facility. However, once development was finished and the system was tested, the devices were actively photographing individuals who had *not* given consent to be photographed. Privacy was still preserved by immediately inferencing on and deleting the images. In such an application, it is imperative to not store or upload clear, privacy-intrusive images. Therefore, an existing and already proven secure solution developed by *HallMonitor*, a company specializing in on-device processing solutions based in Esbjerg, was utilized to establish a secure communication channel with the deployed devices. The communication channel was used to extract the analytics data from the devices. This secure system setup, necessary to protect the devices from attackers, is not covered in this thesis due to its proprietary nature.

**Legal Considerations** The discussions and insights in this thesis may apply to global applications, but the legal considerations specifically target the European Union member countries. Further, some of the discussions may be influenced and biased by a heavily european-influenced cultural mindset and thus not be as relevant and applicable to parts outside Europe. More detailed discussions on international privacy laws beyond just the GDPR should have been included if the technology was intended for global application. Additionally, this thesis does not encompass medical device regulations (MDR) neccessary for devices used for medical purposes on humans.

**Fine-tuned Model Development** The project's broad scope resulted in a limited exploration of potential improvements in model fine-tuning. This thesis evaluates the performance of various machine learning models, including models built from the three arhitectures YOLOv3, YOLOv9, and DETR. Two more object detection architectures are also mentioned, but were not (fully) implemented. These are Co-DETR, the current best-performing model on the COCO dataset, and the Faster-RCNN, another popular and good option for object detection. However, the Co-DETR was deemed too complex and resource-intensive for the project's scope to be fully implemented and evaluated, and Faster-RCNN was not prioritized due worse performance than the YOLOv9. The object detectors are discussed in Section 2.7.

**Museum and Aquarium Opening Hours and Visitor Conduct** The project was designed to avoid interference with the normal operations of the aquarium. Consequently, image capture for the dataset was confined to aquarium opening hours, and random visitors were not inquired whether they'd be willing to participate in the project. An early analysis of the visitation patterns revealed the aquarium was busy from the opening at 10:00 until approximately to 15:00, 2 hours before closing time. This meant most images for the dataset were captured in the two hours before closing time where there was least traffic.

**The Task is Object Detection** There are several tasks within the domain of computer vision, each serving distinct purposes and complexities. This project focuses exclusively on simple object detection, which involves locating objects of relevance within an image. Specifically, this thesis addresses single-class object detection with *person* as the sole class of interest. Other tasks in computer vision include person re-identification, image classification, combined image classification and localization, semantic segmentation, and instance segmentation. Re-identification involves recognizing individuals across different images and image classification is the task of classifying the image contents as a whole. The rest of the tasks are illustrated in Figure 2 to display how they differentiate from object detection.

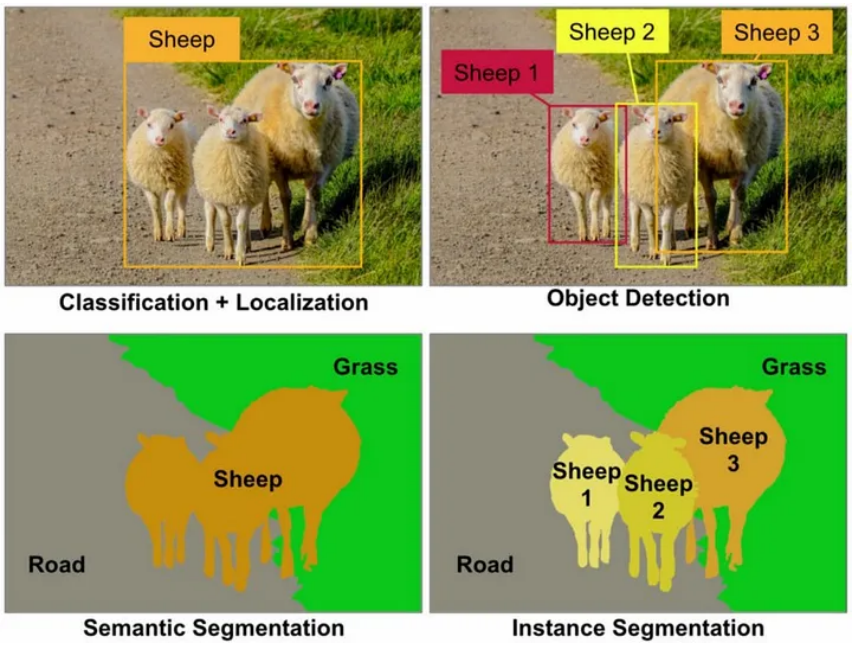



Figure 2: Computer Vision Tasks (Murali, 2021).

The choice of a machine learning model training dataset is reliant on the specific task to be performed. It must contain data suited for that task. For instance, applications tasked with person reidentification require a dataset that includes the identities across images for the persons depicted in the images. An appropriate dataset for such applications, the Person Reidentification in the Wild (PRW), is detailed in Section 3.3.3. A dataset for object detection purposes needs the bounding box and class id for each identified object in the image. In the case of single-class object detection (person localization), only the bouding boxes for each detection for each image would suffice.

**Light Conditions in Aquarium Settings** The environmental conditions within aquarium settings impose specific lighting requirements that significantly influence the deployment and performance of the proposed object detection system. Aquariums typically maintain low-light conditions, presumably to minimize stress for the aquatic animals and improve vision into the tanks for the visitors. In this context, the options for enhancing illumination are often limited, making it impractical to simply increase lighting to facilitate better visual detection technologies. These subdued lighting conditions were an important feature of the thesis project.

**TinyML and Frugal Devices** An initial attempt was made to encompass tinyML and frugal devices in the project. TinyML is when machine learning models are aimed at deployment to heavily resource-constrained environments, e.g. "frugal devices". These are devices where the microcontroller units (MCUs) are accompanied by memory measured in kilobytes, and processor speeds measured in megahertz. Machine learning networks applied to tiny robots are subject to challenges from size, weight, area, and power (SWAP) (Neuman et al., 2022). Many of the

same challenges apply even in applications where the SWAP challenges are not the main concerns. Rajapakse et al. mentions the open challenges and future directions of the next generation tinyML. Catastrophic forgetting, which is when information from previous tasks while learning new ones are forgotten, are a result of the frugal devices' computational resources and memory[3]. The first recommendation for future directions from the authors is to investigate fog computing as a means to offload tasks from the frugal devices. Exploring a system of frugal devices for obtaining sensor data on the edge and "fog devices"[4] for processing of the data would be a time-consuming objective, especially in a real-world scenario where security must be taken seriously. Therefore, tinyML and frugal devices are not elaborated in detail in this thesis, but discussed briefly in C.

**Model Hyperparameter Tuning** The project did not encompass hyperparameter tuning of the models. Hyperparameter tuning is the process of finding the best hyperparameters for a machine learning model. Hyperparameters are the settings of the model that are not learned from the data, such as the learning rate, batch size, and number of epochs. Hyperparameter tuning is a crucial step in the machine learning pipeline, as it can significantly impact the model's performance. However, due to time and resource constraints, hyperparameter tuning was not included in this project. Further comments on hyperparameter tuning are made in Section 3.4.1.

**Person Localization System vs Alternative Notations** In this thesis, the term *person localization system* is employed consistently to describe the technology developed for identifying the positions of individuals. This terminology is preferred as it aligns with the object detection dataset standards where *person* is the designated class name used in object detection model datasets. Furthermore, the choice of "localization" over "positioning" is deliberate to avoid potential confusion; while "positioning" could imply the physical movement or arrangement of persons by the system, "localization" specifically refers to the precise identification of a person's location within a predefined spatial context, accurately reflecting the capabilities and functionality of the implemented system. Additionally, other literature seems to prefer localizatoin over positioning (Gao et al., 2016/11), altough similar systems have also been called indoor *positioning* systems.

## 1.4 Disclaimers

### 1.4.1 Reuse of Previous Work by the same Author

Some of the subsections in the Literature section are heavily based on an unpublished[5] preliminary study for the project of this thesis by the same authors. This preliminary study was written the semester before, and was a theoretical study of how one may achieve "Efficient, accurate, and privacy-preservant object detection in edge devices" (which was its title). This subsections most influenced by the previous work are the subsections 2.5, 2.7, and 2.9.

### 1.4.2 The Use of AI Tools

OpenAI's ChatGPT-4/ChatGPT-4o (OpenAI, 2024) has been prompted with several sections to get suggestions on how to enhance readability and flow. All AI generated suggestions have been verified and major parts rewritten to better fit the intentions of the sections. The author of this thesis takes full responsibility for all of its content. ChatGPT was also used to identify typos in the text, as a final step in the production of this thesis.

[3]Catastrophic forgetting can also be seen in transfer learning, and is why freezing the backbone of a pre-trained model may be a good idea.

[4]Fog devices is an unofficial term for processing-instances located close to the edge.

[5]The preliminary study is not publicly available, since this is not the default practice of NTNU.

### 1.4.3 Privacy of Similar Projects

The author of this thesis is not an expert in privacy. The methods outlined in this thesis are meant to ensure privacy of individuals, but the author cannot guarantee that the methods are foolproof. The author has tried to follow best practices and guidelines from the field and has tried to be transparent about the methods used and the limitations of the methods, but the reader should be aware that following the methods outlined in this thesis may not necessarily be enough to ensure privacy. An investigation into the privacy of similar projects is recommended before deploying a similar system in a real-world setting.

## 1.5 Research Questions

Rather than forming a single hypothesis for this rather widespread project, a set of research questions are formulated to guide the research and to provide a structured approach to the investigation.

**The Research Questions are as follows:**

1. What are some privacy risks associated with traditional person localization systems in public spaces, and how may a system mitigate these privacy concerns?
2. How does the validity of object detection model evaluations change when using data specifically from the intended deployment environment compared to using generic datasets?
3. What are some machine learning architectures suitable for object detection in a real-world deployment scenario?
4. How do the performance metrics of object detection models compare when applied to different quality datasets?

## 1.6 Research Objectives

**The Primary Objectives are to:**

1. Develop a privacy-preserving person localization system using on-device processing to minimize data transmission and enhance data privacy.
2. Demonstrate feasibility and effectiveness of on-device person localization in a practical and realistic setting.
3. Investigate how the quality of the dataset used in model fine-tuning impacts performance.
4. Assess the impact of using real world-specific test data on the evaluations of object detection models, by comparing model evaluations obtained with specific test data to evaluations obtained with more general datasets.

**The Secondary Objectives are to:**

1. Compare the privacy and performance impacts of on-device processing against centralized processing methods.
2. Demonstate how to visualize object detection data by creating visualizations of collected data from a realistic setting.
3. Explore relevant object detection architectures to evaluate their performance in a real-world deployment scenario.

## 1.7 Project Work

**In order to achieve these objectives, the following project work was executed and is detailed in this thesis:**

- Designed and implemented a prototype system that incorporates on-device processing for privacy-preserving person localization.
- Deployed a prototype in an aquarium to gather real-world data and analyze system performance.
- Collected and labeled a dataset of 3394 images to evaluate and fine-tune several object detector machine learning models.
- Conducted comparative studies to evaluate the effects of data specificity and quality on model accuracy.
- Performed a comparative analysis of three object detection architectures under real-world conditions.
- Developed and utilized advanced visualization tools to represent data findings in a way that facilitates clear understanding and decision-making.

This thesis compiles and structures literature on various aspects of person localization systems, bridging multiple domains. It contributes to the field by serving as a collaborative and communicative tool, fostering common understanding among technologists, ethicists, policymakers, and the public.

## 1.8 Structure

The thesis is structured the following way. The descriptions include what each section will cover and why, to increase readability and coherence of the thesis.

**Section 2: Literature Review** The theoretical underpinnings of the project are surveyed and discussed in the literature review. This section contains text from previous work by the same authors (see 1.4.1). The literature review covers a wide range of topics crucial to the development and implementation of on-device processing systems for person localization in environments like museums and aquariums. The following paragraphs provide a summarized overview of the key areas discussed.

Section 2.1 is included to investigate the need for a person localization system in cultural institutions. The review explores alternative, cost-effective methods for visitor behaviour analysis already on the market such as mobile apps and RFID tracking. These technologies provide scalable solutions without the privacy concerns associated with vision-based systems. Museum stakeholders opinions are included in this section. This section presents insights into user perceptions of privacy in smart home environments, which reveal a trade-off between convenience and privacy concerns, indicating similar challenges could arise in public visitor contexts. Section 2.2 about GDPR seek to illustrate how these challenges are addressed in the current EU regulations. It summarizes the requirements for personal data protection and the legal bases for processing such data, which are directly applicable to any person localization system developed. Together with the NIS2 directive, the GDPR promotes the principle of data minimization, pushing for a solution that deletes unnecessary data as soon as the data is redundant.

Section 2.5 provides an in-depth overview of some of the approaches and methods to preserve privacy, specifically in images. The review delineates the privacy advantages of on-device processing over cloud-based systems, emphasizing the importance of local data processing to mitigate privacy risks. Exploration of technologies like federated learning and differential privacy illustrates advanced methods for protecting individual privacy while utilizing data for machine learning.

Section 2.6 presents some various opinions and ethical considerations related to person localization technologies. These perspectives come from a former developer in the field of object detection, an acclaimed writer, and multiple philosophers.

Section 2.7 provides a technological overview of object detector datasets, performance metrics, object detector algorithms, dark-lit environment considerations, and transfer learning. The descriptions for the algorithms are relatively short altough they are important, as the most accurate descriptions are found elsewhere. A subsection of the key differences and similarities between the object detector algorithms of this thesis project are included. Section 2.8 and Section 2.9 brings forth various third party software services and products. These are included due to their relevancy to the project of this thesis and to improve awareness around potential already-existing implementations that may fit a given use case. Specifically, Roboflow and GPT-4 with Vision are evaluated for their utility in building and deploying person localization systems, with considerations on their implications for privacy and data security in real-world applications. Finally, several options for relevant hardware are presented in Section 2.10. This includes microcontrollers and single board computers, specialized components such as graphical-, tensor-, and neural processing units (GPUs, TPUs and NPU), and various sensors capable of relevant functionalities.

**Section 3: Methodology** The Methodology section outlines the detailed technical approaches and methods employed in this project. This section is critical for understanding how the data was collected, processed, and analyzed. It provides a foundation for interpreting the results and findings of this thesis, and a common understanding of the terms.

Section 3.1 provides an overview of the project, including the hardware decisions and the deployment of devices in the aquarium. This section also explains the different machine learning models employed for training and evaluation. Section 3.2 delves into the construction of the *FIMUS* dataset. It covers the camera configurations, image capturing process, and labeling. This section highlights the differences between the inconsistent and consistent partitions of the dataset, detailing the specific settings and processes used to ensure the quality and relevance of the captured images. Section 3.3 briefly describes the external datasets used in the project; COCO, CrowdHuman, Person Reidentification in the Wild, and Football-players. It explains the relevance of each dataset to the project and how they were utilized for model training and evaluation.

Section 3.4 focuses on the training process of the models. The section contains information regarding the licenses of the models, and the hyperparameter tuning process (which was mostly foregone), and the use of Google Colab for cloud training with GPUs, detailing a few pros and cons. Finally, it provides an explanation regarding the lack of validation used in the training of the models, a decision based both on the lack of hyperparameter optimization in the project and issues with google colab service disconnections.

Section 3.6 restates the metrics used to evaluate the performance of the models in this thesis project, and defines them as *COCO AP* and *Vary-Both AP*. Section 3.5 presents the different models that were developed and tested in the project. It includes details about the pre-trained and fine-tuned models, their configurations, and the specific experiments conducted to measure their performance.

Section 2.6 addresses the ethical methodologies that were implemented in the project to ensure preservation of privacy. Section 3.8 discusses the use of heatmaps as a visualization tool to analyze visitor behavior patterns. It details the attempts to create heatmaps using different Python packages and the final implementation using the Supervision module by Roboflow[6].

By following this structure, the Methodology section provides a comprehensive and transparent account of the technical processes involved in the project, ensuring that the results are reproducible and the conclusions are drawn based on a solid and well-documented foundation.

[6]A final heatmap is displayed in the end of this section, and in the results (Section 4.2)

**Section 4: Project Results** This section presents the findings and performance evaluations of the person localization system implemented in the project. It includes detailed experiments designed to address the research questions and objectives outlined in Section 1.5, highlighting their implications and significance within the broader context of the field.

Section 4.1 evaluates the object detection models using various dataset partitions and presents the results of different experimental setups. It includes a detailed analysis of average precisions and recalls, and explores the effects of fine-tuning, different test-set compositions, and input image sizes on model accuracy and inference latency.

Section 4.2 discusses the visualization of data through heatmaps and peak hours analysis, providing insights into visitor behavior patterns and engagement levels. Subsections illustrate how heatmaps and bar charts are used to analyze and visualize the data, highlighting visitor engagement patterns and identifying peak visitation times.

By presenting these results, this section provides a comprehensive overview of feasibility and effectiveness of a person localization system. It also serves as a demonstration of the practical implications and potential applications of the person localization system in a real-world setting.

**Section 5: Reflections & Overall Discussions** This section provides a comprehensive discussion of the broader implications of the research findings, reflecting on the results and their significance for the development of similar systems. It evaluates the viability of the approach used in this thesis to address the presented problem and restates the research questions from the introduction, providing concrete answers to each.

Section 5.1 discusses the privacy implications of image deletion versus obfuscation post-analysis. It critiques existing studies on privacy preservation and argues for the need to consider demographic and cultural differences in privacy perceptions.

Section 5.2 examines the use of third-party services and products, highlighting potential drawbacks such as loss of control, data privacy concerns, cost efficiency, performance optimization, and scalability issues. It emphasizes the importance of balancing the convenience of third-party solutions with the need for customization, security, and long-term cost management.

Section 5.4 explores the ethical considerations of developing person localization systems. It discusses the potential risks of mass public control and the importance of ethical frameworks such as utilitarianism and deontological ethics in evaluating the impacts of these technologies. The section underscores the need for rigorous ethical assessments and informed regulation to ensure the responsible deployment of person localization technologies.

Section 5.5 reflects on the hardware and software choices made in the project, considering their adequacy and potential areas for improvement. It discusses the benefits and limitations of using Python for on-device processing and suggests considerations for optimizing device efficiency with lower-level programming languages.

Section 5.6 summarizes the results of the project, comparing COCO AP and Vary-Both AP metrics and discussing their implications for model selection. It highlights the importance of balancing precision and recall in evaluating object detection models and the need for comprehensive performance metrics.

Section 5.7 provides concise answers to the research questions posed at the beginning of the thesis, summarizing the findings and their implications. It addresses privacy risks, the validity of model evaluations using specific datasets, suitable machine learning architectures, and the impact of dataset quality on performance metrics.

Section 5.8 situates the project within a broader context by discussing its implications for public environment deployment, privacy preservation, and the applicability of object detection models in various settings. It emphasizes the practical utility of the system for crowd management and operational efficiency, and highlights the essential ethical considerations for deploying such technologies in public spaces. The project demonstrates the feasibility and effectiveness of on-device person detection in realistic settings, providing valuable insights into model performance and adaptability, while underscoring the importance of thoughtful and comprehensive ethical standards in technology development.

**Section 6: Future Work** This section outlines a comprehensive research agenda to advance on-device person detection systems, building on the findings of this thesis. It addresses remaining questions and proposes several future research directions. These include evaluating fine-tuned models on generic datasets to assess their generalizability and robustness across diverse environments, expanding the FIMUS dataset to create a more diversified person localization dataset, and exploring advanced privacy-preserving techniques such as differential privacy and homomorphic encryption.

Further, it proposes expanding practical applications of on-device detection technologies to areas such as smart home automation, healthcare monitoring, and personalized user experiences in public venues. Enhancing data visualization tools by integrating additional variables for heatmap generation and creating zones to visualize visitor flow and exhibition popularity are also recommended.

These propositions aim to guide future research, ensuring the continued development and ethical deployment of on-device person detection systems.

**Section 7: Conclusions** This section summarizes the key insights and findings of the thesis, concluding the research on the viability and ethical implications of on-device person detection systems. It encapsulates the significance of privacy preservation, the importance of relevant datasets for model evaluation, and the ethical considerations essential for responsible technology deployment. The conclusion also reflects on the practical implementation, future research directions, and policy recommendations, providing a comprehensive closure to the thesis.

**Appendix A: Code Snippets** Appendix A provides guidance on the sequence for setting camera configurations to achieve consistent image quality for the dataset, detailed in Table 2. It emphasizes the importance of setting the shutter speed first and allowing the automatic gain control to stabilize before disabling automatic settings and manually adjusting the white balance gains.

**Appendix B: Camera Settings Explanation** This is a quick and easy-access overview for understanding the camera module settings that were used in this project.

**Appendix C: TinyML and Frugal Devices** This appendix spans 4 paragraphs which landed outside the scope of the project, but is included for its relevancy for similar systems with lower complexity hardware.

**Formatting Guidelines** This document adheres to APA formatting guidelines for quotations, by indenting block quotations longer than 40 words and enclosing shorter quotations within the text in quotation marks. Page numbers are excluded due to the efficiency of digital search tools. Title case capitalization is applied to section headings (also in line with APA style).

# 2 Literature Review

The advent of modern object detection has enabled more sophisticated and automated methods for understanding visitor engagement and flow in cultural institutions. This literature review aims to explore the current state of research on the various topics of this thesis. This includes visitor behaviour analysis and stakeholders perspectives, individual privacy, object detection, the influence of fine-tuning models on specialized data, and existing third-party services for deploying similar person localization systems.

## 2.1 Visitor Behaviour Analysis and Stakeholder Perspectives

Some of the traditional methods of analyzing visitor behaviour include surveys, manual counting, and direct observation. Today, more technology-driven and practical applications may be used to gain insights in visitor engagement and experience in a museum or aquarium setting. In this subsection, we look at an alternative to computer vision tracking systems. Afterwards, a study on the perceived value of visitor tracking to museum stakeholders is brought foreward.

### 2.1.1 Questions Asked by Visitors in a Mobile App

Pérez Cortés et al. had visitors ask questions in a mobile app while moving through the museum 2023. Visitor movement through the museum was inferred from the data by leveraging question keyword content, knowledge of exhibit layout, and question timestamps. This removed the need for more costly, vision-based applications for detecting and tracking visitor movement. This study illustrates one way of conducting affordable, dependable and scalable visitor analysis without the need for costly devices.

### 2.1.2 Perceived Value to Museum Stakeholders

Lanir et al. explored an alternative approach to museum visitor behaviour analysis, and its perceived value to museum curators, administrators and department heads (2017). Wearable RFID trackers[7] were given to the visitors, and beacons were positioned at positions deemed important by the museum curators. The beacons would then communicate the positions of the visitors to the system. This allowed for the collection of data on key metrics like exhibit popularity, average visit duration, and common visitor paths. The authors noted that technology-based visitor behaviour analysis was generally well-received by museum curators, offering valuable data that could enhance the visitor experience.

The study of Lanir et al. further discussed the divergent views between the curators and the administrators on the utility of visitor behaviour analysis systems (2017). Administrators and department heads generally viewed these systems favorably, citing the financial justification for expensive exhibitions: *We really need to know if this expenditure was worthwhile* (Lanir et al., 2017). On the contrary, museum curators expressed skepticism. One curator remarked:

> A temporary exhibition won't change after you deploy it, and understanding how it is used by the public would not help me in my next exhibition, since they are each very different. My main reason to analyze behaviour would be to satisfy my curiosity. (Lanir et al., 2017)

This contrast underscores the varied perspectives within museums regarding the value and implications of behaviour analysis technologies.

[7] The requirement for visitors to wear RFID trackers represents a significant drawback as it may be perceived as intrusive (although completely privacy preservant).

### 2.1.3 User Perceptions of Smart Home Privacy

In a 2018 study, researchers conducted semi-structured interviews of 11 smart home owners were conducted to figure out user perceptions of smart home IoT privacy and obsoleteness as there are frequent upgrades and new products on the market (S. Zheng et al., 2018). Responses regarding the privacy demands versus an increase in convenience of the IoT devices were concerning:

> I think it's more likely that a lot of these things will become obsolete... If that's what happens then I have to buy another device. It still might be worth it for the convenience. (Participant 10)

> [The security concern] is always kind of in the back of my mind because of all that IoT stuff that always goes on, and everyone says how easily hackable they are. But I think my peace of mind that I get from having them outweighs my worry of what could be potentially taken advantage of. (Participant 6)

These responses indicate that the convenience and connectedness of the devices surpass the desire to preserve privacy[8]. This is a promising finding for the development of visual systems in museums and aquariums, as it suggests that the benefits of the system may outweigh the privacy concerns of the visitors. It also illustrates the need for regulations to prevent solutions from being developed that are too privacy-invasive, as the end-users will not prioritize privacy over convenience.

## 2.2 The General Data Protection Regulation (GDPR)

The general data protection regulation (GDPR) is a single set of regulations to guarantee privacy and protection of personal data. A quick review of the GDPR should be on the agenda of anyone affiliated with systems not inherently preservant of privacy[9].

The GDPR entered into applicability in the EU on 25th of May 2018 has two major impacts. 1) It leaves individuals with more control over their data, and 2) it facilitates a level playing field for all companies; there is now a single set of data protection rules for all companies operating in the European Economic Area (EEA)[10]. The most relevant sections of the GDPR to this thesis are the regulations regarding personal data.

## 2.3 Personal Data

Personal data is any form of information that can be connected to an identifiable data subject. The following definition was given by the european parliament in 2016:

> The term 'personal data' means any information relating to an identifed or identifable natural person ('data subject'); an identifable natural person is one who can be identifed, directly or indirectly, in particular by reference to an identifer such as a name, an identifcation number, location data, an online identifer or to one or more factors specifc to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person. (The European Parliament, 2016)

Privacy, which is closely knit to personal data, is more discussed in Section 2.5. First, we take a look at some approaches to managing personal data.

[8] Connectedness of IoT is still an issue since there is no standard communication protocols. Matter is a Zigbee-based wireless protocol that operate as an application layer on top of Thread, Ethernet or Wifi. It allows IoT devices to communicate, but it is not yet widely adopted. This may be due to the high costs, as the Connectivity Standard Alliance demands a yearly membership for the rights to use th CSA's trademarks. Additionally, the devices must be certified for security, adding extra costs as many companies must buy this certification service. Another great improvement in terms of connectivity that does not limit privacy is the EU regulation that all smaller devices sold after 2024 must have USB-C connection. This also applies to laptops in 2026.

[9] More specifically, *informational* privacy. This term is introduced in 2.5

[10] The EEA consists of all EU countries plus Iceland, Liechtenstein and Norway.

**Approaches to Managing Personal Data** Various methodologies can be adopted to manage personal data within a system.

One approach involves transforming information so that it no longer qualifies as personal data. This can be achieved through techniques such as differential privacy, which ensures that the processed data cannot be traced back to an individual. See Section 2.5.3 for a detailed explanation of differential privacy.

A second approach for managing personal data involves establishing lawful grounds for the processing of personal data. This necessitates adherence to legal frameworks that justify the use of personal data under specified conditions, thereby ensuring compliance with data protection regulations. Processing of personal data is permissible under the GDPR only when it satisfies at least one of the following legal bases:

**Legal Bases for Processing Personal Data**

- The data subject has given explicit consent.
- It is necessary for the performance of a contract to which the data subject is a party.
- It is necessary for compliance with a legal obligation to which the controller[11] is subject.
- It is necessary to protect the vital interests of the data subject or of another natural person.
- It is necessary for the performance of a task carried out in the public interest or in the exercise of official authority vested in the controller.
- It is necessary for the purposes of the legitimate interests pursued by the controller or by a third party, except where such interests are overridden by the interests or fundamental rights and freedoms of the data subject which require protection of personal data.

Additionally, the controller is responsible for compliance with the 3 requirements summarized below and should be able to demonstrate this compliance at any given time.

- Security Documentation In the event of a breach of personal data, the controller must document that proper precautions were made to secure the data. One of these precautions is to deleted data that is no longer needed. This rule to delete no-longer-needed data is often overlooked and violated by companies (Sandtrø, 2022).
- Data breaches Breaches of personal data must be reported within 72 hours. Companies failing to do so are economically sanctioned, but even worse, it damages the reputation of the company. In such cases it is common to uncover more failures (Sandtrø, 2022). This is often that the company has failed to make, or failed to document, the efforts they have made to sufficiently protect the data (the first requirement).
- Rights of the data subject The data subject has the right to be informed about how their personal data is handled. This is commonly achieved through the company's privacy declaration, which must be comprehensive and regularly updated. Additionally, companies are encouraged to proactively communicate this information to clients, for instance, via email. According to privacy experts (Sandtrø, 2022), adopting such practices is an effective way of building and maintaining trust with customers.

There are multiple other approaches for managing personal data that are more specific to the management of *images*. The ones discussed in this thesis are primarily concerned with removing the individual information of persons from the images. These methods are discussed in Section 2.5.

[11] The controller refers to the party controlling the data

## 2.4 The Network & Information Security 2 (NIS2) Directive

The NIS2 Directive (European Parliament and Council of the European Union, 2022) is a more recent EU regulation that came into force in January 2023. Unlike the GDPR, which broadly addresses the protection of personal data, NIS2 is specifically targeted toward technology. As an update to the EU's cybersecurity framework, NIS2 focuses on strengthening the security of network and information systems. It emphasizes the critical need for robust security measures in systems that process personal data to prevent unauthorized access and data leaks.

Both NIS2 and GDPR highlight the principle of data minimization, which mandates that object detection systems process only the necessary amount of personal data for their intended function. This practice not only bolsters security but also supports privacy by minimizing potential data exposure. Adhering to these principles is vital for maintaining user trust and ensuring compliance with EU regulations, particularly when deploying object detection technologies in environments where data sensitivity is paramount.

## 2.5 Preservation of Individual Privacy in Images

Building on the previously introduced regulations in Section 2.2, the contents of this section aims to provice a deeper insight into the methods of preserving the individual privacy of individuals *in images*. It is heavily influenced by previous work of the same author. See the disclaimers in Section 1.4 for more details.

The first definition of privacy was given by Brandeis and Warren in 1890 as *the right to be let alone*. A more comprehensive definition of privacy that is more relevant to the modern age of digitalization and the topics of this thesis is the following:

Privacy as informational self-determination:

> Privacy is the claim of individuals, groups and institutions to determine for themselves, when, how and to what extent information about them is communicated to others. (Westin, 1967)

There are multiple dimensions to privacy. The definition of Westin covers informational privacy, which is most relevant to this thesis. This definition includes groups and institutions, however, in most legal systems, privacy is defined as a basic human right that only applies to natural persons[12]. The term *individual privacy*[13] encapsulates the individual focus of privacy as opposed to the broader interpretations of privacy that might apply to groups, organizations, or institutions. Other dimensions of privacy include spatial privacy, territorial privacy, and bodily privacy (Fischer-Hbner and Berthold, 2017). These will not be further discuss in this thesis.

Preservation of individual privacy refers to maintaining the personal space and confidentiality of individuals, ensuring that their private lives and personal integrity are not invaded or exposed (without consent).

Protection of personal data is important due to the regulations, preserving the individual privacy is also essential to maintain trust with customers, and to avoid (*delay*[14] for brief discussion on the ethicality of person localization systems development.) the onset of a dystopian society...

[12] A natural person (also sometimes referred to as a physical person) is a title used to identify an individual human being. This is different from a legal person, which can be an individual or a company (Termly, 2023).
[13] Individual privacy is sometimes referred to as personal privacy.
[14] See Section 5.4

### 2.5.1 Privacy in Images

Protection of personal data in general is very similar to protection of personal data in images. Protection deals with the management and security of personal information—data that can identify an individual, such as names, addresses, and biometrics. This protection is primarily about the correct handling, processing, storage, and destruction of personal data to prevent unauthorized access, misuse, or breaches.

There are multiple methods, both pre- and post-processing, for preserving individual privacy in images. One example of a pre-processing privacy preservation method is to hide the facial regions optically during capture, which was done in a study on fall detection by X. Wang et al. (X. Wang et al.).

Post-processing methods include various techniques to obscure identifiable information after the data has been captured. These range from simple blurring and pixelation to more sophisticated approaches such as k-anonymity (Sweeney, 2002) and differential privacy. Six of the simple, easy-to-implement methods are shown in Figure 3, demonstrating practical implementations.

(a) Blurred entire image of Hong Kong street to protect privacy of citizens.

(b) Blurred face of individual by a sea town in Cinque Terre.

(c) Masked faces.

(d) Pixelated faces.

(e) Unconventional method: replace faces. May be done as effectively as the other approaches, but is likely to be seen as an unprofesional approach.

(f) Deleted image. This is the most effective and secure, but removes the possibility of verifying results and is unsuitable for most vision-based applications.

Figure 3: Six Methods of Individual Privacy Preservation in Images

K-anonymity was claimed to be a mathematically proven method for anonymyzation of personal data, but has been critizised by its successor, the l-diversity criterion, for not being robust in the events where attackers have background data (Machanavajjhala et al., 2007). Differential privacy is a statistical disclosure control algorithm to process individual data from a group to produce close-to-real outcomes without disclosing the personal data of individuals (Y. Huang et al., 2023). Differential privacy is explained and illustrated in 2.5.3.

There has been considerable research focused on preserving privacy within the realm of machine learning (Ravi et al., 2023). A fundamental principle shared across various use cases is that deleting data serves as the most definitive means of ensuring privacy, assuming such measures are practicable. When only non-personal data is retained, the application achieves unequivocal security concerning privacy.

### 2.5.2 Federated Learning

In many systems relying on machine learning, being able to utilize locally stored personal data may augment the system to perform better for the situation it was created for. However, sharing this personal data with a centralized model may not be possible due to the legal bases for processing personal data (see sec:legal-bases-processing-personal-data).

Federated learning, also known as collaborative learning, is a decentralized approach to training machine learning models. It does not require exchange of data from client devices to global servers. The concept of federated learning is seen in Figure 4.

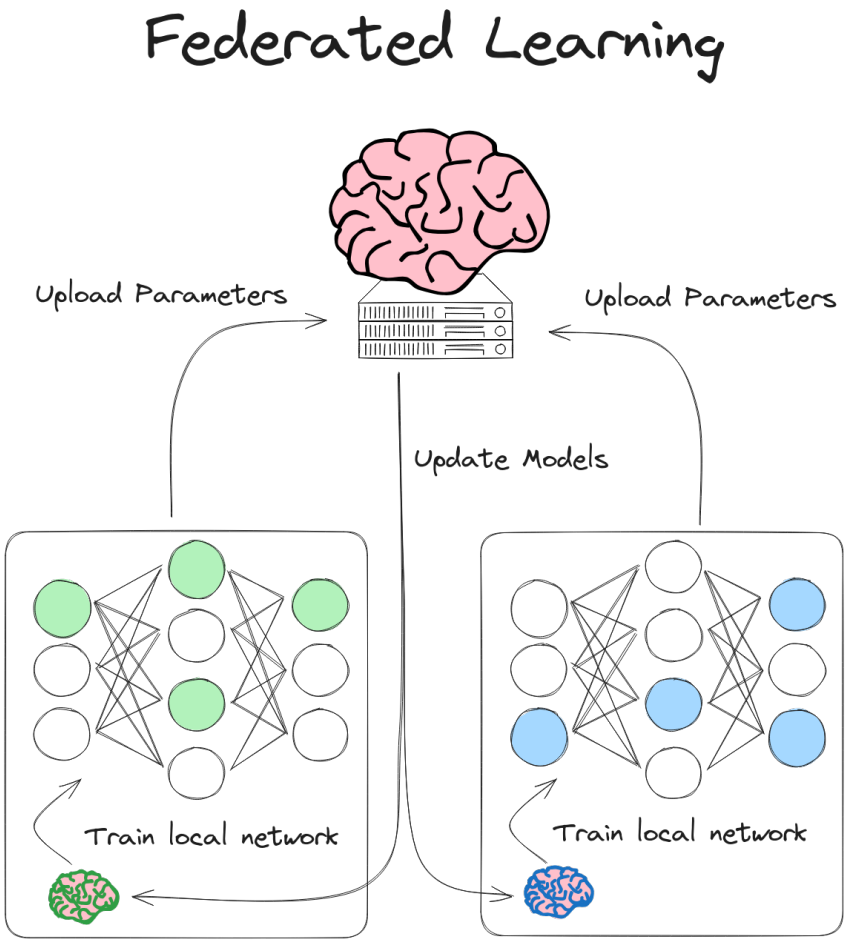



Figure 4: The Federated Learning Concept

Federated learning is described in the article of Antunes et al. (2022):

> A central entity manages the learning process and distributes the training algorithm to each participating data holder. Each participant generates a local model trained with their private data and shares the resulting parameters with the central entity. Finally, the central entity employs an aggregation algorithm to combine the parameters of all local models into a single global model.

In summary, FL enables the training of ML models locally (at the location of the data) and only shares the resulting model, which is not reverse-engineerable, with the requesting party. Therefore, FL avoids the need to share the private datasets and sensitive data to others, preventing exposition to entities conducting studies and enabling data usage for broader purposes (Gu et al., 2021).

The FL process is reliant on having ground truth data on the edge for training the models correctly, but obtaining the ground truth for edge device models operating on *visual data* is difficult. The way this may be achieved, is by having a powerful edge device perform the inferences with a computationally expensive but accurate model, and using the inference results of this model as the ground truth for training a separate, possibly faster and less computationally expensive model to replace the other at a later stage. Otherwise, one could also perform the training under conditions where the ground truth is known, for example by manually inputting the number of people in an area, then having the model learn to arrive at the same count based on the camera input.

Improvement of machine learning models devices in the healthcare industry present challenges due to the sensitive nature of medical data from patients. Centralized training of machine learning models may violate laws such as the GDPR, because of the way data is being collected and used unbeknownst to the data subject (Antunes et al., 2022). To tackle these issues, Antunes et al. (2022) proposes the usage of FL[15] to tackle these issues.

Furthermore it should be noted that FL is a method to deal with the existential nature of data in edge computing devices, best described as *isolated islands*, and to use the data on edge devices before it is deleted or obscured, to improve the intelligence of the devices in privacy preservant and protective way. An important measure to take in the development of FL models is to ensure that the models are not reverse-engineerable, as the models may contain personal data. This is done by adding noise to the data, which makes it impossible to backtrack the data to the individuals. This may be done by a method such as differential privacy, which is discussed in Section 2.5.3.

### 2.5.3 Differential Privacy

Regulations regarding personal data also applies to the events where pieces of information are aggregated to identify a person. The concept of differential privacy is to make data of individuals privacy-preservant through describing them as a group. Data from the group of people may be used, but without the possibility of backtracking the information to certain individuals. See Figure 5.

In more technical terms: Differential privacy is a statistical disclosure control algorithm to process individual data from a group to produce close-to-real outcomes without disclosing the personal data of individuals (Y. Huang et al., 2023). This means that the data is processed in a way that the results are close to the real results, but the data is not disclosed. This is done by adding noise to the data, which makes it impossible to backtrack the data to the individuals.

Differential privacy is particularly pertinent in the context of federated learning. In this approach, client devices add controlled noise to their model updates—or weights—before sending them to a centralized server. This noise addition prevents the server from being able to infer individual-specific information from the model updates. The degree of noise is regulated by a privacy parameter, often referred to as a privacy budget. This strategy allows the central server to aggregate these noisy updates from all participating nodes to update the global model. Contrary to the original statement, the noise is not removed but rather managed in such a way that the aggregated model maintains utility while protecting individual privacy (Sharma, 2023).

[15] Specifically, the FL method described in the works of Yang et al.(2019)

Note that differential privacy is a definition, not an algorithm (Dwork et al., 2011). In other words, we can have many different algorithms that satisfy the privacy demands for a given use case. For example, Dwork et al. mentions the Laplace mechanism (outlined in the same authors works from 2006) as an optimal mechanism for answering "tally" type questions differentially privately (2011). For more advanced situations, other algorithms, such as the method outlined by Blum et al. (2011), are more suitable (Dwork et al., 2011).

The big tech giants like Apple, Google and Microsoft employ differential privacy in their data collection and analysis to ensure the privacy of their users. Differential privacy is a method to ensure that the data is not personal, and thus not subject to the GDPR.

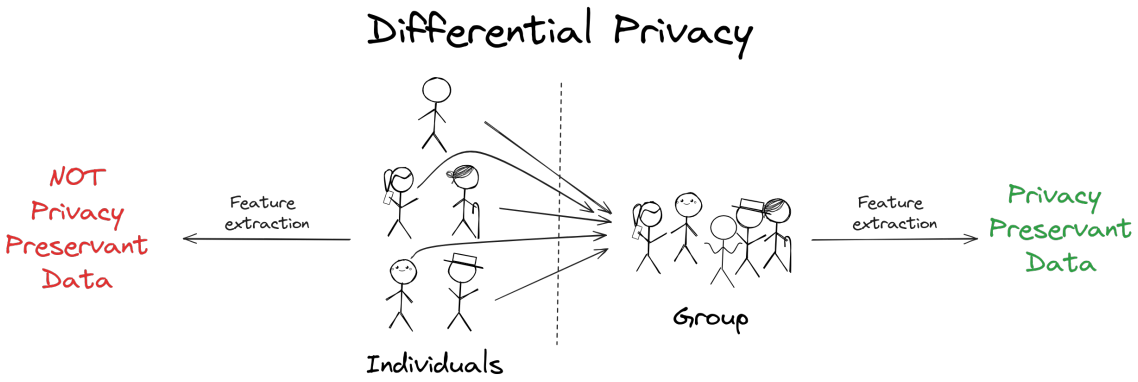



Figure 5: The Differential Privacy Concept

### 2.5.4 On-Device Processing

According to Z. Huang et al. (2022), there are four methods for running tasks on resource-constrained edge computing devices. This is relevant in applications where user's concerns for privacy increases if data is directly transmitted to a server. These methods are seen in Table 1, and explained discussed in the following paragraphs.

| Method | Advantages | Disadvantages |
|---|---|---|
| Data encryption | Privacy protection<br>Fast calculation | Much bandwidth |
| Traditional ML | Little resource consumption | Relying on the Internet<br>Poor robustness |
| Task sharing | Reducing stress on a single device | Much bandwidth<br>Large latency |
| Deep learning | Privacy protection<br>High robustness | High resource consumption |

Table 1: Comparison of Methods for Running Tasks on Resource-Constrained Edge Computing Devices (Z. Huang et al., 2022)

**Data Encryption** The first method, data encryption, would be one way of transmitting images in a more secure way. This should be done in a lossless way to maintain the image quality to preserve the accuracy of the detectors. Doing so is not trivial, and is a research field on its own. A few methods that may function well, e.g. blurring only the faces, are discussed in Section 2.5.7.

**Traditional Machine Learning** The second method for running tasks on resource-constrained devices is to implement the less computationally expensive traditional machine learning methods. Unfortuneately, this is a unprefferable solution in many situations due to their lack of accuracy compared to the deep learning models. This is displayed in Figure 6). Here, we see the mean average precisions (mAPs) have skyrocketed since the introduction of neural networks as backbones for object detection algorithms. Note that the best performing algorithm in 2015, the year after COCO was introduced, scored a mAP50-95 of 35.9. This value has nearly doubled with the development of YOLOv4, and has been further improved with the more advanced transformer technologies thereafter.

Traditional machine learning algorithms may, however, be a good option for devices with low computing power and memory resources as they are generally low-demanding and may be easier to understand[16]. The traditional methods are most applicable to use cases with clear, deterministic logic. Traditional machine learning methods were the most prominent prior to 2014, while deep learning based detection models have been the completely dominant approach to image recognition tasks. To achieve similar accuracies to those of the deep learning models but with the low computational demands of traditional machine learning, one might consider to investigate the field of tinyML, which was scoped out of this thesis 1.3. Some considerations are, however, added in appendix C due to their relevancy to resource-constrained edge computing devices, a topic considered bordering to the thesis.

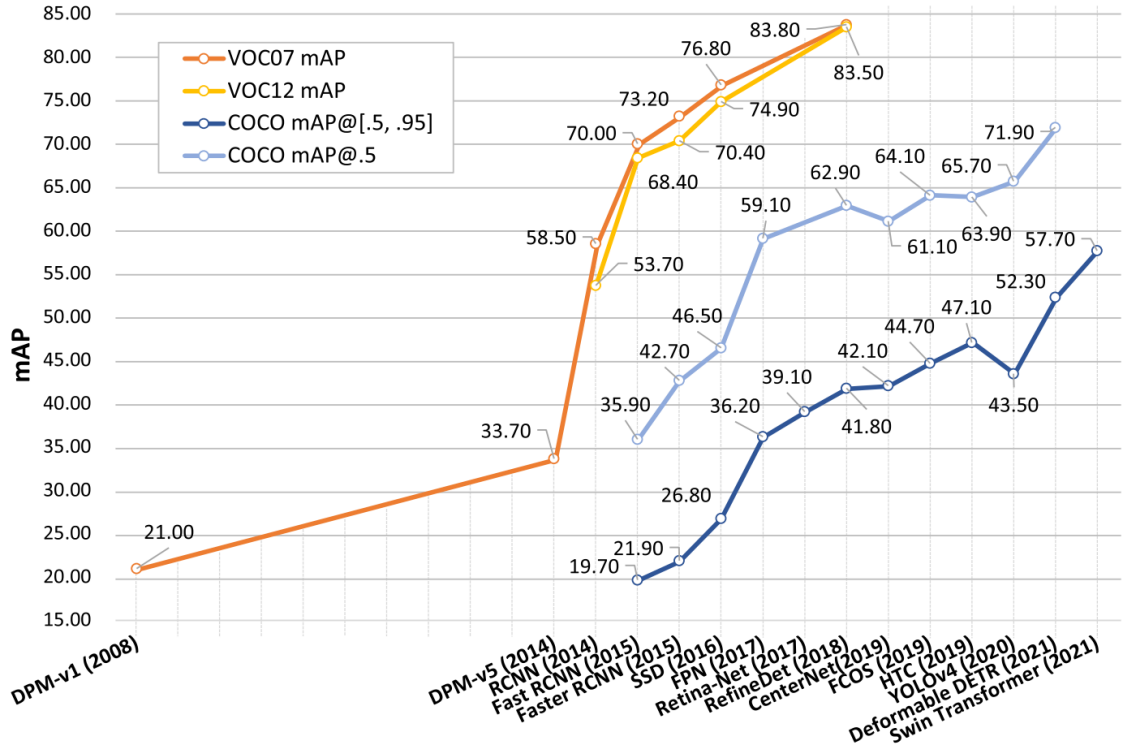



Figure 6: Accuracy Improvement of Object Detection on VOC07, VOC12, and MS-COCO datasets (Zou et al., 2023)

**Task Sharing** The third method, sharing the workload over multiple devices, is not an uncommon practice in technology. See for example Eufy's solution with a home-base device in Section 2.9.1), where the camera devices take images and send them to a more powerful computer for processing. This approach reserves privacy, because the images are never sent outside of the local network, and it can be implemented with simple TCP/IP[17] communication.

[16]Neural networks struggle with explainability and are often referred to as black boxes

[17]Transmission Control Protocol/Internet Protocol is a set of standardized rules that allow devices to communicate with each other on a network.

Task sharing results in low latency, fast networks, but introduces (1) the need of having a central hub, (2) extra work of setting up the transmission protocols, (3) another source of error and (4) the need to encrypt/decrypt images prior/post transmission to ensure security. However, due to periodically scarcity in the availaility of specialized hardware such as GPUs, this approach could be useful since a single GPU per facility may achieve a higher throughput than processing large data on the CPUs of multiple edge devices.

**Deep Learning** As opposed to traditional machine learning, This section outlines some methods to retain the privacy of individuals by using different sensors or implementing neural network on the edge devices, often referred to as on-device processing or edge computation. Which term of on-device processing and edge computation is used may be dependent on which aspect of the concept the author chooses to emphasize; the actual process that is happening on a device, or the architectural decision of making the computation on the edge.

### 2.5.5 Depth Cameras

A widely used approach within the domain of anonymous fall detection, is to use of RGB depth cameras to capture depth information (X. Wang et al., 2020). As only depth information is captured, the data remains completely anonymous from the start.

### 2.5.6 Deletion of Images

In an investigation of an already existing internet of things (IoT) system for wildlife monitoring: 'Where's The Bear', relying on motion-triggered cameras, three challenges of visual systems in such applications were discussed Elias et al., 2017. The drawbacks were (1) the transmission of enormous numbers (sometimes millions) of images over low-bandwidth networks, which tend to happen in automatically (motion-) triggered applications, (2) motion sensors triggered by weather conditions or by animals that were not of interest, and (3) redundancy of images taken of the same individual animal. While the 2nd and 3rd drawbacks are not applicable to this project, the 1st is.

Elias et al. proposed a solution to this challenge: edge computing. Edge computing, also referred to as on-device processing, encapsulates similar concepts but emphasizes slightly different aspects of the computing approach. While *on-device processing* specifically indicates that the computational tasks are carried out directly on the device itself, *edge computing* underscores that these tasks are performed close to the data sources, i.e., at the *edge* of the network.

The deployment of visual systems in public spaces presents challenges related to privacy, not only because of the immediate access to private data, but also due to the recent breakthroughs in object detection allowing the extraction of sensitive information from visual data. The altogether only completely safe way to ensure complete and total privacy of data, is to not have the data at all.

Edge computing and on-device processing allows for the image is obscured or deleted right after analysis without ever leaving the edge device. In this way, only the anonymous analysis results are communicated online. This would mean that the personal data (1) exists *just* while the analysis is running, (2) is never sent online, and (3) is thus a lot less vulnerable to attacks. The perpetrator's device would need to be physically connected to the device and the attack would need to happen in real time. In those cases, the perpetrator could quite likely just as well take the photo himself. This is an approach to achieve low-latency, high bandwidth, high availability, low cost communications and fast response to/from the sensors.

The images would in some cases benefit in multiple ways from being obscured instead of deleted. This approach is discussed in the following paragraph.

### 2.5.7 Obfuscation

Another way to remove the privacy concern is by obscuring the images after analysis in such a way that individuals may never be identified.

Obfuscation is the action of making something obscure, which means to conceal or make unclear. To obscure an image is often used interchangeably with "to blur", but they are not the same. To blur means to make something indistinct or hazy, and is a specific method of obscuring an image. Other methods for obscuring an image is masking or pixelate the faces of individuals. These methods are illustrated in Figure 3.

**Blurring the Faces** In a 2019 study, faces were detected with a thermal-detecting camera and then photos were captured with an RGB camera, blurring the area the face was detected by the thermal camera (Ma et al.). This approach is privacy preservant as long as all faces are blurred, but may fail if the algorithm does not detect all faces. In those cases, however, most humans would likely also struggle to identify a person based on the face. On the contrary, in many cases, blurring the entire image would compress the image, making it faster and easier to transfer, and be the faster option than having to detect all faces in an image.

**Perceptions of Privacy Enhancement Methods** A questionnaire study of 328 students indicated that blurred images were not considered by the students to provide satisfactory privacy protection (Edgcomb and Vahid, 2012). Participants were given 18 randomly ordered videos, and were asked to rate the privacy on a Likert[18] scale from 1-5. The obfuscation methods, or privacy enhancements as they called them, and the results are displayed in Figure 7. The results show that blurred images were only considered privacy preservant for 23 percent of participants. Regardless, an important notion is that the images of this survey are from within a private home, posing higher demands and expectations with regards to privacy than what is typically done in a more public space.

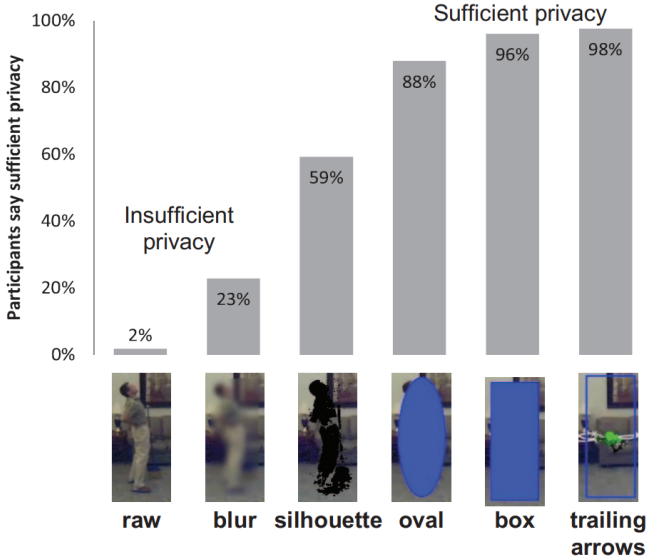



Figure 7: Privacy Enhancements Methods in the Study of Edgcomb and Vahid (2012)

[18]Likert scale: A scale of odd options, where the participant may answer a neutral middle-option and distribution should be equally distributed in both directions thereafter. An often used questionnaire scale in psychology research.

## 2.6 Ethical Considerations in the Development of Person Positioning Technologies

As we advance the capabilities of technologies such as YOLOv9 for person localization, it becomes imperative to consider the ethical implications of our developments. The narrative of George Orwell's dystopian novel *1984* serves as a reminder of the potential societal consequences of intelligent, extensive and automated surveillance. Orwell's portrayal of a society where history is constantly rewritten and individual privacy is obliterated highlights the dangerous path we might tread if these technologies are misused by those in control of political power.

"The Party seeks power entirely for its own sake. We are not interested in the good of others; we are interested solely in power, pure power." (Orwell, 1949). This may remind of some politicians, i.e. american presidents, who may decide to deploy person localization devices to keep population under control while tightening the grip on the population...

### 2.6.1 Joseph Redmon Quit Computer Vision Development

Joseph Redmon, the creator of the initial versions of YOLO, decided to cease his work on the project due to its military applications. This illustrates a profound ethical stance. Redmon's choice underscores the responsibility of developers in considering the broader impacts of their work. The resignation marks a critical point in the discourse on the moral responsibilities of researchers and developers in the field of artificial intelligence and machine learning. The discussion of how to responsibly regulate and develop AI applications is still ongoing, and the decisions made by individuals like Redmon are crucial in shaping the future of the field.

Joseph Redmon's Twitter Posts:

> I stopped doing CV research because I saw the impact my work was having. I loved the work but the military applications and privacy concerns eventually became impossible to ignore. (Redmon, 2020a).
> But basically all facial recognition work would not get published if we took Broader Impacts sections seriously. There is almost no upside and enormous downside risk. (Redmon, 2020a).
> [...] I bought in to the myth that science is apolitical and research is objectively moral and good no matter what the subject is. (Redmon, 2020a).
> If you worked in a knife factory and a guy came in and thanked you for making knives because he killed many people with those knives and then he showed you a video of himself killing people with a knife you made how would you feel then about working in your knife factory? (Redmon, 2020b).

### 2.6.2 Philosophical Perspectives

Immanuel Kant's deontological ethics emphasizes the importance of adhering to moral rules or duties. According to Kantian philosophy, actions are morally right if they are in accordance with a moral rule or principle, regardless of the consequences. A Kantian ethics viewpoint slightly adjusted to the development of localization technologies might suggest that developers have a duty to create technology which uphold principles such as privacy, no matter where their technology is applied.

Utilitarianism, a consequentialist theory primarily developed by Jeremy Bentham and John Stuart Mill, posits that the rightness or wrongness of actions depends on their outcomes, specifically their contribution to overall happiness or utility. In the context of localization technologies, utilitarianism would require a careful assessment of the potential benefits and harms. While such technologies can enhance safety, efficiency, and convenience, they also pose significant risks to privacy and individual freedoms. Developers must strive to maximize the overall good while minimizing potential harms, ensuring that the societal benefits outweigh the risks and negative consequences.

Modern philosophers, computer scientists and artificial intelligence researchers like (respectively) Sam Harris, Stuart Russell, and Eliezer Yudowsky also discuss the implications of AI. They are, however, discussing topics such as the long-term risks of artificial general intelligence, the control problem, and the value alignment problem. These are not so relevant for a person localization system without capabilites of decisionmaking. They are all voicing a concern, however, for the lack of regulations for the rapid-growing field of AI, where politicians oblivious to the nature of AI is making the regulations.

### 2.6.3 Practical Ethical Framework for Development

In developing technologies capable of tracking and analyzing human behaviour, transparency and accountability are paramount. Developers must ensure that the design, implementation, and deployment processes are transparent, allowing for public scrutiny and informed consent. Clear guidelines and regulations should be established to hold developers and users accountable for the ethical use of localization technologies. This includes regular audits, impact assessments, and the involvement of diverse stakeholders in decision-making processes.

Privacy safeguards are critical in mitigating the ethical risks associated with localization technologies. Robust data protection measures must be implemented to secure personal information from unauthorized access and misuse. Techniques such as anonymization, encryption, and differential privacy can help protect individual privacy while allowing for the beneficial use of data. Legal frameworks like the General Data Protection Regulation (GDPR) in the European Union set important precedents for protecting personal data and ensuring privacy rights.

Continuous monitoring and ethical auditing are essential to ensure that localization technologies are used responsibly. Regular assessments should be conducted to evaluate the ethical implications of these technologies, identifying and addressing potential risks and unintended consequences. This involves establishing independent oversight bodies and ethical review boards to provide ongoing guidance and recommendations for ethical practices in the development and deployment of localization technologies.

The AI community must remember the lessons from pioneers like Redmon. We must strive to develop technologies that do not compromise ethical standards for convenience or profitability. Historical examples and dystopian stories of technological misuse and ethical failures should inform current practices, guiding the development of localization technologies in a manner that prioritizes ethical considerations and societal well-being.

The development of localization technologies presents complex ethical challenges that require us to be vigilant and proactive. By embedding ethical considerations into the fabric of our technological innovations, we can avoid the dystopian futures forewarned by Orwell and ensure that these tools serve to support and enhance human society, rather than diminish it.

### 2.6.4 Object Detection Performance Benchmark Datasets

There are multiple benchmark datasets for machine learning applications. The area of facial emotion recognition alone has at least five benchmark datasets (Saurav et al., 2022). For the task of object detection, the Common Objects in Context (COCO) dataset (Lin et al., 2014) has been widely used since its introduction in 2014, with its 330 000 annotated images.

Another well-known, widely adopted dataset for classification, object detection and segmentation is the PASCAL Visual Object Classes (VOC) (Everingham et al., 2010). The PASCAL VOC websites include several challenges, i.e. VOC2005 through VOC2012, for researchers to benchmark their detectors. Even though the challenges have completed, one can still evaluate new methods on their datasets.

A third dataset is the CrowdHuman dataset. This may be the most relevant for a detector aiming to detect persons, as it consists of 24 370 images with in total 400 000 human (person) instances in diverse occlusions and variations.

For any use case implementation however, it is vital to have a dataset that is relevant to the problem at hand. For a detector aiming to detect persons in a dark-lit museum, the most relevant dataset would be one with images from dark-lit museums.

In real-world applications there are licenses for using datasets for training a model. Testing and benchmarking a solution against a certain dataset is typically free to do, but the datasets are often under a license which forbids commercial use.

### 2.6.5 Object Detection Performance Benchmark Metrics

Machine learning can be seen as a gamified[19] version of statistics and software engineering. Object detection is a subset of machine learning. Modifications and new advances in object detection methods may be instantly evaluated by running inference on benchmark datasets and compare them to the other state of the art (SOTA) models.

Partly due to the aforementioned gamified nature of machine learning models, which metrics are deemed important may have a significant impact on the development of the models. There are competitions on the data science platform Kaggle, where data and machine learning specialists may compete for the best scores. The developers of the best-performing models are awarded prize money in many of the competitions. The target variables for the competitions are what drives development. According to Zou et al., the developments primarily pursue two main goals: enhancing prediction accuracy and increasing computational efficiency (2023. Additionally, the evaluation of object detectors extends to more, harder-to-measure, abilities. This can be their ability to transfer their capabilities to new domains, such as learning to detect a new category it has not previously been trained for. There's not yet been a focus on energy efficiency, which needs to happen soon, should development continue for AI in the current pace (Luccioni, 2023).

The most used measurement of performance for an object detector model is the *mean Average Precision* (mAP) for varying values of *IoU thresholds* (Zou et al., 2023). The average precision is the average when taking the average of precision values under various recalls. The mean is when this is averaged for all the object classes in the dataset. The IoU represents how well the predicted box fits to the ground truth. The average precision may be calculated fixing the IoU threshold, fixing the confidence threshold, or varying them both. More on this later.

First the thesis provides an overview to understand the concepts of true positives, false positives, false negatives, the confusion matrix, precision and recall. These are easiest to explain if the task is image classification and not object detection. For 2.6.5 and 2.6.5, we will use the example of image classification, but the concepts are the same for object detection, with the difference that the bounding box positioning is also taken into account.

**Understanding TP, FN, and FP, and the Confusion Matrix** For a machine learning model dealing with a regression problem[20], the metrics usually used to evaluate its performance is the number of true positives, false negatives and false positives.

These may be defined as follows:

1. True Positive (TP): The number of instances correctly identified by the model as positive. For instance, if your model is tasked with identifying people in images, a true positive would be an instance where the model correctly identifies a person.

2. False Negative (FN): The number of instances where the model incorrectly identifies a positive instance as negative. Using the same example, this would be a situation where the model fails to identify a person who is actually in the image.

[19] Gamification is the practice of applying typical elements of game play (e.g. point scoring, competition with others, rules of play) to an activity, typically as an online marketing technique, to encourage engagement with a product or service (Dictionary, 2023).

[20] Object detection is also a regression problem, as the model is simply relating the independent variable input image pixels to a dependent variable output of the bounding boxes and classes.

3. False Positive (FP): The number of instances where the model incorrectly identifies a negative instance as positive. This could occur if the model identifies a person in an image where there is no person.

The confusion matrix is a table used to illustrate these numbers. An example of a confusion matrix is shown in Figure 8.

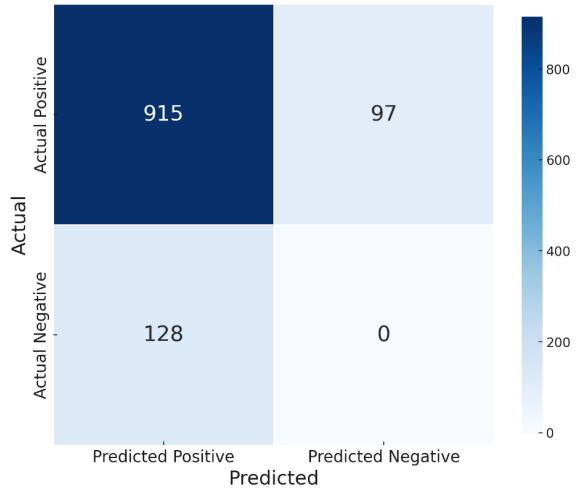



Figure 8: Confusion Matrix

The confusion matrix in Figure 8 displays that the model has detected 915 people correctly, failed to detect 128 people, and incorrectly detected 97 people where there were none. For classification tasks, it is common to have the table show which class the model has detected, and which class the object actually is. For single class object detection, the confusion matrix is sufficient as-is.

Further the TPs, FNs and FPs are used to calculate the precision, recall and F1 score of a machine learning model.

**Understanding Precision and Recall** For a balanced metric of precision and recall we also have the F1 Score, combining the two in a single value. Here's a breakdown of each:

Precision: Measures the accuracy of positive predictions. It is the ratio of correctly predicted positive observations to the total predicted positives. High precision relates to a low rate of false positives. For object detecion of persons, precision would be how accurate the model is when it claims to detect a person.

$$\text{Precision} = \frac{TP}{TP + FP} \tag{1}$$

Recall (Sensitivity or True Positive Rate): Measures the ability of the model to find all the relevant cases within a dataset. It is the ratio of correctly predicted positive observations to the all observations in actual class. High recall relates to a low rate of false negatives. For object detection of persons, recall would tell us how many of the actual persons in the image the model was able to detect. We must note that since precision does not take false negatives into account, using recall as a performance metric may be vital in the situations where detecting all the persons may be important.

$$\text{Recall} = \frac{TP}{TP + FN} \tag{2}$$

F1 Score: The weighted average of Precision and Recall. This score takes both false positives and false negatives into account. It is particularly useful when the class distribution is uneven. F1 Score is best if there is some sort of balance between Precision and Recall in the system.

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{3}$$

The precision-recall curve is commonly used to assess the performance of a general machine learning model (see Figure 9 for an example). The precision-recall curve is a graph that shows the trade-off between precision and recall for different thresholds for confidence in the object class. As you allow your model to be more uncertain in its inferences[21], the number of hallucinations will also increase and thus the precision drops. The area under this curve is the average precision (AP) of the model.

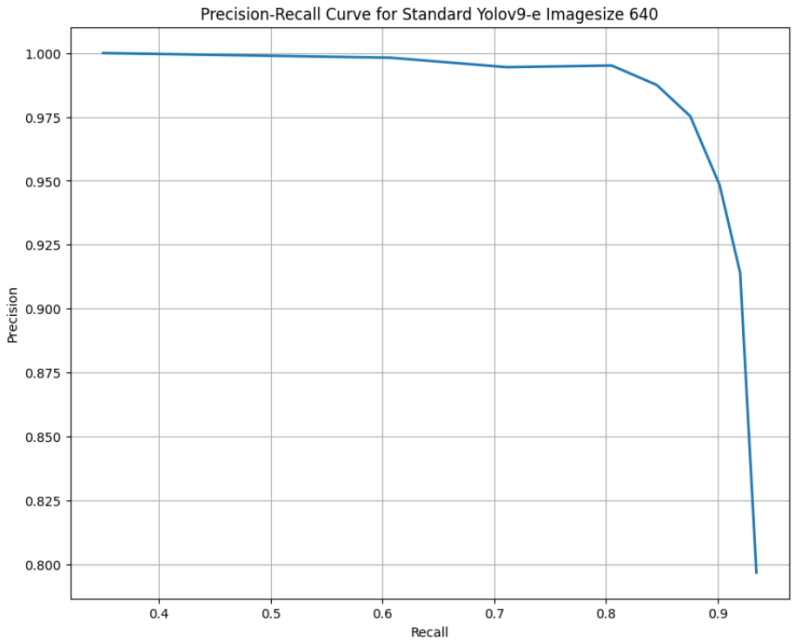



Figure 9: Precision-Recall Curve

The area under the Precision-Recall curve is the average precision (AP) of the model. This can be expressed as follows:

$$\text{AP} = \sum_{n} (R_n - R_{n-1}) P_n \tag{4}$$

where $R_n$ is the recall at the $n$ n-th threshold, $R_{n-1}$ is the recall at the previous threshold, and $P_n$ is the precision at the $n$ n-th threshold.

Alternatively, AP can be represented as an integral:

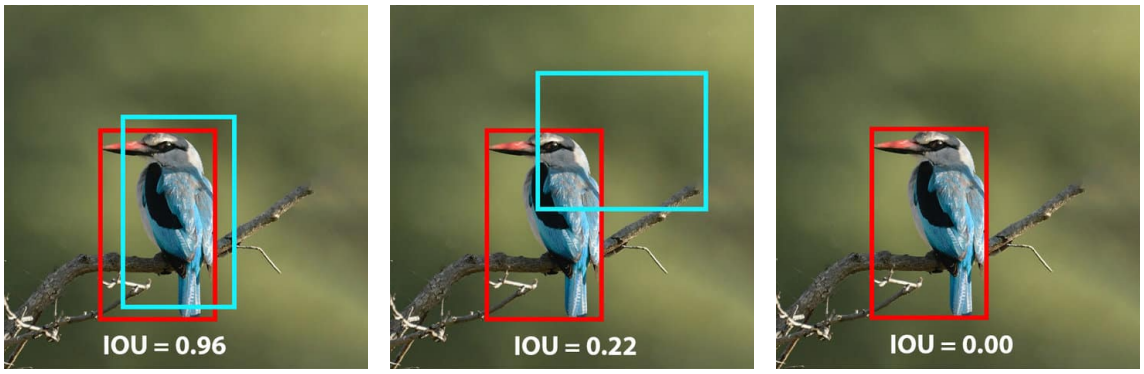
$$\text{AP} = \int_0^1 P(R) dR \tag{5}$$

where $P(R)$ is the precision as a function of recall $R$.

**Understanding the IoU Metric** Accuracy in object detection refers to both detecting the object *and* its location accurately. Combining both in one metric would simplify benchmarking. The precision, recall and f1-score all neglect the positioning precision of bounding boxes.

For assessing localization accuracy, the Intersection over Union (IoU) is calculated. This compares the predicted bounding box and the ground truth bounding box in a way so boxes need to fit as closely to the ground truth bounding box as possible to get the best score (which is 1.0). See Figure 10.

[21] For object detection, there are at least three ways of allowing the model to be more uncertain. Fixing the confidence threshold and vary the IoU, or fixing the IoU and vary the confidence threshold, or by averaging over both thresholds.




(a) True Positive (b) False Positive (c) False Negative

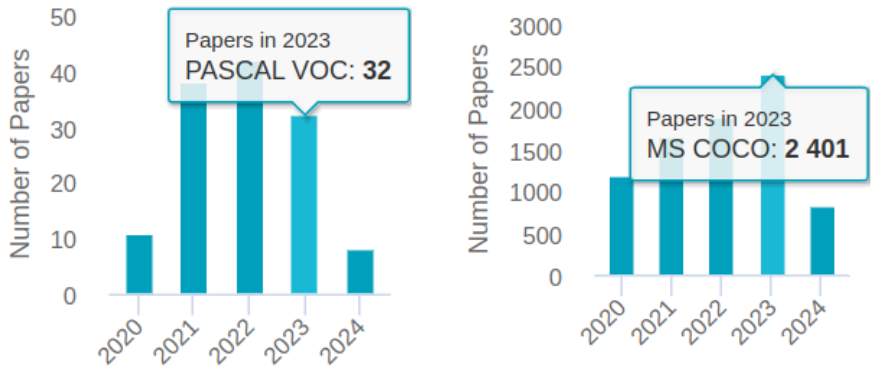
Figure 10: Intersection over Union (IoU) (OpenCV, 2022). The red boxes are the ground truth, and the blue are the model predictions.

A high IoU values equates to having a good fit with the ground truth bounding box. If the IoU value is over a certain threshold, we define the detection to be a true positive (see 10a). If the predicted bounding box has little overlap, we identify this as a false positive (see 10b). This may also be called a hallucination. If we don't have detections for a ground truth bounding box (see 10c), we have a false negative. A fourth case is where the predicted bounding box fully overlaps with the ground truth, but covers a larger area. In this case, we have a low IoU due to the high area of union, and thus a false positive.

The equation for calculating the IoU of a predicted bounding box and a ground truth bounding box is as follows:

$$\text{Intersect over Union} = \frac{\text{Area of Overlap}}{\text{Area of Union}} \tag{6}$$

**Conclusion of Performance Metrics** The Pascal Visual Object Challenge (VOC) was a standard way of measuring performance. Here, the IoU value was fixed (typically at 0.5), while the confidence in detections was averaged over multiple confidence thresholds. The fixed IoU threshold is typically set at 0.5 or higher. Which value is best dependends on the accuracy demands of the sceneario, and is why retaining the ability to adjust the threshold is a good idea when implementing an object detector. Today, the VOC AP metric is seldom used (see 11). Following the introduction of MS-COCO datasets in 2014, researchers started to pay more attention to the accuracy of object localization instead of using a fixed IoU threshold (Zou et al., 2023).




(a) Pascal VOC: 32 (2023) (b) COCO: 2401 (2023)

Figure 11: PASCAL VOC and COCO Approximate Number of Usages (META AI, 2024)

The single-most important metric for object detection in the COCO dataset challenge was the mean Average Precision (mAP) over 10 IoU thresholds of .50:.05:.95. This rewards the detectors with more accurate localization (Lin et al., 2014).

A third method of measuring AP would be to vary both the confidence threshold and the IoU threshold. This would consider both the classification *and* localization accuracy, but the calculation is also computationally heavier. In the Project Results, we discuss whether this computational expense makes a difference in the evaluation of the object detectors in this project on the FIMUS dataset.

## 2.7 Object Detection Algorithms

This subsection includes a brief summarization of the evolution of object detection, including the transition from traditional methods to morw modern methods such as the YOLO series and vision transformers.

The evolution of object detection can be divided into two major historical phases: before and after 2014, as illustrated in Figure 12. Prior to 2014, traditional object detection methods, such as the Viola-Jones detectors (Viola and Jones, 2001), Histogram of Oriented Gradients (HOG), and Deformable Part-Based Models (DPMs) were prevalent[22]. During this era, *mixture models* were developed to improve detection granularity by recognizing the different parts of the same object, such as the doors and windows of a car.

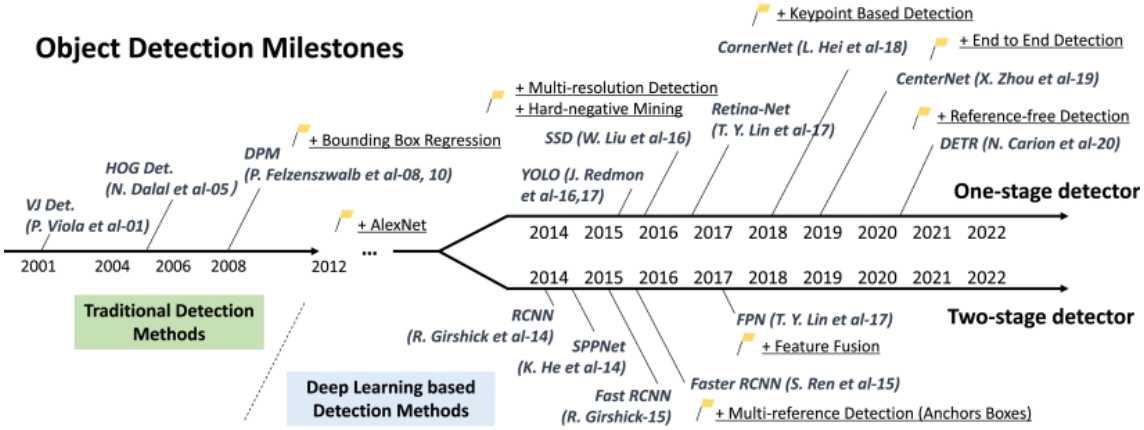



Figure 12: Object Detection Milestones (Zou et al., 2023)

Despite these advancements, it was not until the introduction of Region-based Convolutional Neural Networks (R-CNN) in 2014 that the accuracy of object detection systems began to improve significantly. This paradigm shift marked a substantial advancement in the field, leveraging deep learning techniques to enhance detection performance dramatically (Zou et al., 2023). The period following 2014 has seen rapid progress, introducing sophisticated object detectors like You Only Look Once (YOLO) and Detection Transformers (DETR).

[22]These are just some honorable mentiones of some of the most sucessful and widely adopted models of the time (Li et al., 2012)

### 2.7.1 You Only Look Once (YOLO)

The YOLO (You Only Look Once) object detection algorithm is renowned for its efficiency in real-time object detection. YOLO divides the image into a grid and predicts bounding boxes and class probabilities for each grid cell. Unlike traditional object detection algorithms that require multiple passes through a network, YOLO processes images in a single pass. This approach significantly enhances detection speed. The YOLO algorithm was the first one-stage (single pass) detector, and led the way for the development of other various popular networks such as the RetinaNet and DETR.

**YOLOv3** YOLOv3 marked a significant advancement in the YOLO series by integrating multi-scale predictions and a deeper feature extractor. A deeper feature extractor refers to the use of more layers in the convolutional neural network (CNN), which allows the model to capture more complex features from the images. These improvements improved speed and accuracy from the first versions of YOLO.

**YOLOv9** YOLOv9 represents what was the latest and most advanced version in the YOLO series at the beginning of this thesis project. It features numerous optimizations for faster training and increased accuracy, especially in challenging conditions such as low light and occlusions. YOLOv9's architecture is streamlined to reduce computational overhead, enabling it to perform well even on less powerful devices. This version also benefits from enhanced post-processing techniques that refine the accuracy of its predictions.

**YOLOv10** YOLOv10, built with ultralytics and RT-DETR, is the current latest addition to the series. The commit message "add yolov10" was made on 23rd of May, signalling the first date of the release.

### 2.7.2 Detection Transformers (DETR)

The Detection Transformer (DETR) is an innovative machine learning method introduced by the Facebook (Meta) Research team. DETR leverages a transformer encoder-decoder architecture, similar to those used in natural language processing models. This architecture enables the model to handle complex object detection tasks by processing global information within images, rather than relying solely on local features. See Figure 13 for an illustration of the architecture.

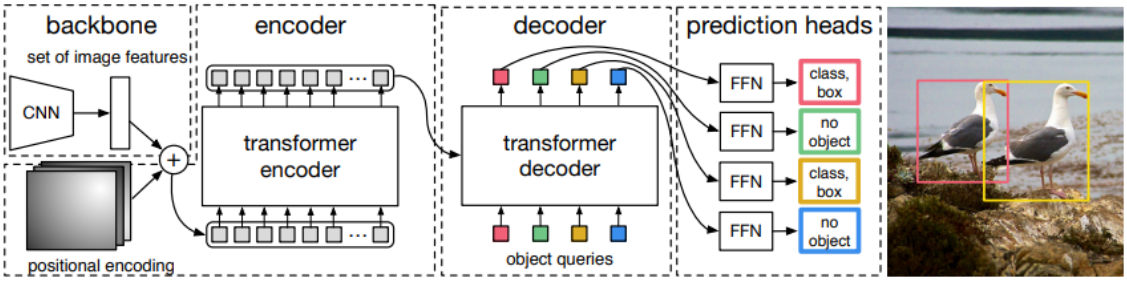



Figure 13: The Architecture of the DETR Model (Carion et al., 2020)

In the paper where DETR was first introduced, Carion et al. demonstrated that DETR outperforms several competitive baselines on tasks such as panoptic segmentation[23]. They achieved these results by integrating a simple segmentation head trained on top of a pre-trained DETR.

[23]This is a challenging pixel-level segmentation task where an image is divided into meaningful regions.

The conclusions of their study highlight the potential of DETR:

> We presented DETR, a new design for object detection systems based on transformers and bipartite matching loss for direct set prediction. The approach achieves comparable results to an optimized Faster R-CNN baseline on the challenging COCO dataset. DETR is straightforward to implement and has a flexible architecture that is easily extensible to panoptic segmentation, with competitive results. In addition, it achieves significantly better performance on large objects than Faster R-CNN, likely thanks to the processing of global information performed by the self-attention. This new design for detectors also comes with new challenges, particularly regarding training, optimization, and performances on small objects. Current detectors required several years of improvements to cope with similar issues, and we expect future work to successfully address them for DETR. (Carion et al., 2020)

The conclusions encapsulates the core innovations and findings of the DETR model. However, the authors also acknowledge the challenges that come with this new approach. Training and optimizing DETR is complex, especially when dealing with small objects. The initial performance issues on training and optimalization present challenges in terms of being able to fine-tune the models to a specialized dataset. Despite these challenges, the authors express optimism that DETR can overcome these initial hurdles, much like how earlier detectors evolved over time.

### 2.7.3 Comparison of YOLO and DETR

While DETR offers a groundbreaking approach by utilizing transformers for object detection, its complexity and the need for domain-specific training models can limit its adaptability and scalability. In contrast, the YOLO series, particularly YOLOv9, provides a more robust solution for real-time applications. YOLO's ability to quickly process images, coupled with continual improvements in both speed and accuracy, makes it a more practical choice for diverse and dynamic environments.

The introduction of YOLOv9 highlights a notable weakness in the DETR series. C.-Y. Wang and Liao pointed out:

*However, since it is extremely difficult for DETR series object detector to be applied to new domains without a corresponding domain pre-trained model, the most widely used real-time object detector at present is still YOLO series.* (2024)

This assessment underscores the flexibility and widespread adoption of the YOLO architecture in various operational contexts. In contrast, DETR's specialized and computationally intensive requirements make it less versatile for broader applications without significant adjustments and domain-specific training. Thus, while DETR presents a novel and highly effective approach, YOLO remains the preferred choice for real-time, adaptable object detection tasks.

### 2.7.4 Dark-Lit Environments

Being able to detect and locate people in dark-lit environments have been previously attempted, usually for security concerns in public spaces. Park et al. developed a system for detecting people in dark-lit environments using a convolutional neural network (2020). They modified the three input channels which usually take RGB to take as input instead (i) the original infrared image, (ii) a difference image from the previous frame, and (iii) a background subtraction mask. Their dataset is vastly different from the setting for this thesis, as the individuals in their photos were far away from the cameras. However, they found that their system was able to detect people in dark-lit environments with an accuracy of 90%. This is a promising result, as it shows that it is possible to detect people in dark-lit environments using infrared imaging and CNNs. They used FLIR cameras, which make images from heat. Doing inferences on pure infrared images may be harder because the infrared radiation may be less prevalent than the heat a FLIR camera may capture. The FLIR cameras are expensive, and thus not considered viable for this project.

The article "YOLO in the Dark - Domain Adaptation Method for Merging Multiple Models" by Sasagawa and Nagahara presents a novel approach for improving object detection in low-light conditions (2020).

The key facets and findings were:

- The proposed method merges pre-trained models from different domains using glue layers and a generative model, enabling adaptation to new tasks without additional datasets.
- The YOLO-in-the-Dark model combines the "Learning-to-See-in-the-Dark" model with the YOLO model, enhancing object detection capabilities in low-light images.
- By creating latent features from existing datasets, the generative model trains glue layers efficiently, reducing the need for new data and computational resources.
- The YOLO-in-the-Dark model effectively detects objects in raw short-exposure low-light images with fewer computing resources than traditional methods.

These findings highlight the model's potential for efficient, high-performance object detection in challenging lighting conditions, making it a valuable tool for person localization applications in real-world low-lit environments.

### 2.7.5 Transfer Learning and the Effectiveness of Fine-tuning

Transfer learning is the process of transferring knowledge from a source domain to a different but related target domain. In practice, this means having a pre-trained model fine-tune on a dataset that is specialized for the task at hand. Extending a model's capabilities to learn to correctly identify a new object class or improving the detection accuracy are typical examples of transfer learning use cases. The following research on transfer learning for object detection models demonstrates accuracy gains when fine-tuning pre-trained models compared to training from scratch.

Wei et al. introduced Feature Corrective Transfer Learning (FCTL) in their study (2024). This approach enhances object detection in non-ideal visual conditions by incorporating a feature similarity loss during training. The Non-Ideal Image Transfer Faster R-CNN (NITF-RCNN) model, developed using this method, showed improved detection accuracy in challenging environments by aligning feature maps between ideal and non-ideal images.

Another study used a generative model to create synthetic training data, which was then used to pre-train an object detector (Paiano et al., 2023). This pre-trained detector was subsequently fine-tuned on a limited real dataset. This method, applied to detect cars in urban environments and fish in underwater settings, resulted in improved detection accuracy compared to using real data alone. The key advantage was leveraging the large synthetic dataset to enhance the detector's initial training phase before fine-tuning it on the actual data, yielding better performance in both domains. An illustration of the process is depicted in figure 14.

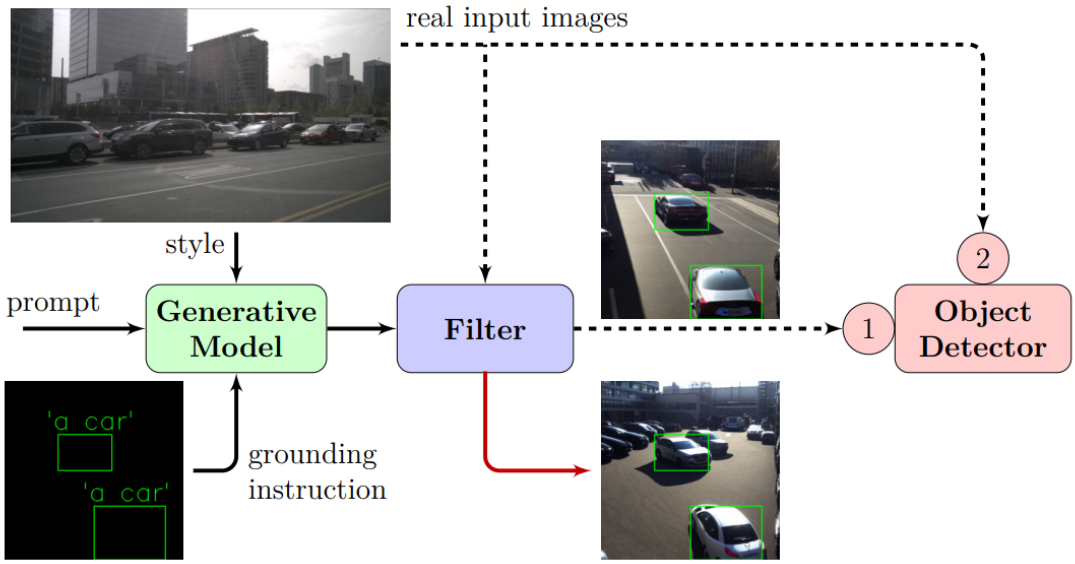



Figure 14: Transfer Learning for Object Detection With Generative Models (Paiano et al., 2023)

Description for Figure 14 as given by the authors:

> We employ a L2I pretrained model to generate images for transfer learning to an object detector. We can filter out suboptimal generated images based on benchmark metrics. For instance, the image along the red arrow is discarded because the generative model has depicted many cars outside the bounding boxes designated in the grounding instruction. With the remaining generated images, we pretrain the object detector, followed by a fine-tuning on the real dataset. Dashed lines indicate the data used for training the models. (Paiano et al., 2023)

Generating artificial training data is not uncommon practice for model fine-tuning. Elias et al. discussed some of the challenges with visual systems in an investigation of an existing internet of things (IoT) system for wildlife monitoring *Where's The Bear*. The system suffered from three major drawbacks; (1) the transmission of enormous numbers (sometimes millions) of images over low-bandwidth networks, which tend to happen in automatically (motion-) triggered applications, (2) motion sensors triggered by weather conditions or by animals that were not of interest, and (3) redundancy of images taken of the same individual animal.

Elias et al. implemented solutions to the challenges. The first solution was to utilize edge computing to avoid sending every image online. This is an approach to achieve low-latency, high bandwidth, high availability, low cost communications and fast response to/from the sensors. In this thesis we make the distinction between edge computing and on-device processing. Edge computing is when the processing happens *close to* the edge, but not necessarily on the device itself. For *Where's the Bear's* solution, the data was transmitted to a device close to the sensors for processing, greatly reducing the distance for the transmissions.

The next solution of Elias et al., was to train their models using artificially produced, specialized training data. Since images of bears are uncommon and they needed a high number of images of bears, they took empty images from the wildlife camera's and placed them in images to improve the model. See figure 15.

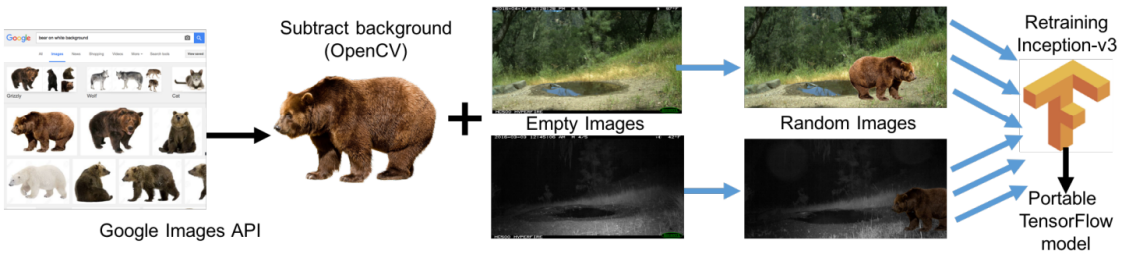



Figure 15: Where's the Bear's Artificially Placed Bears (Elias et al., 2017)

The results of the study was, however, that a model trained on real images from one of the cameras performed slightly better than the one trained on artificially produced images. This motivates the decision to not put artificially placed humans in images for a person localization system.

The results of these aforementioned studies show that accuracies of machine learning models may be improved by fine-tuning on specialized data. Next, the thesis outlines some third-party services and products, similar to the Where's the Bear system, that were included in this thesis for their possible usefullness to the project.

## 2.8 Third-Party Services

Roboflow is a platform designed to simplify and enhance the process of building and deploying machine learning models, particularly in the domain of computer vision. The platform offers comprehensive tools for data management, model training, and deployment, making it highly valuable for applications requiring precise object detection, including the localization and detection of persons.

### 2.8.1 Roboflow

Roboflow's ecosystem comprises several key components that streamline the development of computer vision models:

- **Data Management**: Roboflow provides tools for annotating, organizing, and augmenting image data. These features facilitate the creation of high-quality datasets that are essential for training accurate models. Datasets created with these tools are then stored and hosted on the Roboflow server, open for other people to use.
- **Pre-Trained Models**: The platform offers a wide range of pre-trained models optimized for various tasks. Users can leverage these models to accelerate the development process, especially when combined with transfer learning techniques to adapt these models to specific tasks. This also means that any model you create yourself will be available for your potential industry competitors.
- **Model Training and AutoML**: For users without deep technical expertise in model architecture, Roboflow's AutoML capabilities offer an automated way to generate models tailored to their unique datasets. This enables a quick and easy-to-grasp way of implementing machine learning for a use case.
- **Deployment**: Roboflow enables seamless model deployment via APIs, allowing models to be integrated into applications effortlessly. This API-driven approach supports both cloud-based and local deployments, ensuring flexibility according to user needs with regards to inference speed due to network latency and data privacy and security.

The platform's ability to manage and process data through a user-friendly interface allows for rapid iteration and experimentation, reducing the time from concept to deployment.

#### Use Case: Detection of Persons

Roboflow excels in scenarios requiring the detection of specific objects within varied environments, such as detecting persons in crowded or complex scenes. The platform supports the deployment of models capable of identifying and localizing persons with high accuracy, which is crucial for applications in security, retail analytics, and urban planning.

One application would be using Roboflow to train models on the CrowdHuman dataset, without the need to download the dataset or train the model locally. Models may be fine-tuned for scenarios such as monitoring museum traffic. The Roboflow website contains multiple guides for how such applications may be implemented.

### 2.8.2 OpenAIs Generative Pretrained Transformer 4 (GPT-4) with Vision

The capabilities of large language models (LLMs) have expanded beyond text to include various tasks, such as visual processing. GPT-4 with Vision is an example of a multimodal model (LMM). Many products already integrate OpenAI's ChatGPT, and adding visual capabilities can enhance these applications. LLMs with vision can semantically understand scenes, like predicting a riot in a bar street in England or recognizing a fish feeding event at an aquarium, and gauging crowd reactions[24]. This could allow automated applications to provide insights without human analysis, making them faster, more accurate, and scalable compared to current surveillance systems.

However, the generative nature of GPT models poses a challenge: their performance can vary daily. While many experiments show promising results, they are static and may not reflect consistent performance over time. Consequently, models may perform well in tests but fail post-deployment.

To address this, a website has been dedicated to measure how the GPT-4 with Vision[25] performs across a range of experiments. The website is made by the team at Roboflow, but let's other users submit their experiments for daily checkups through git pull requests. Out of 13 of the experiments currently posted, 5 have failed every day the last 7 days, and 2 have failed at least once in the last 7 days. One of the experiments, counting fruits in a bowl, is alternating every day between success and failure. This proves the point that generative models may still be considered too unreliable for many applications.

Further, in May 2024, OpenAI introduced its newest edition of the renounced ChatGPT series; the ChatGPT-4o. *o* is for omnimodal, and refers to its ability to perform in a multitude of modalities, including vision. This model was tested for the task of object detection, but rendered unsatisfactory performances[26]. A review of ChatGPT-4o, including a more in-depth description of the experiment on object detection, is found here. The experiment is displayed in Figure 16. ChatGPT-4o, misplaced two bounding boxes when prompted to detect the dog in the image.

[24]There has been considerable research focused on detecting the mood of people. This requires high resolution images of good quality. One model would detect people or faces, and another would get cut-outs of those faces to detect the mood of each individual.

[25]Previously called GPT-4V https://platform.openai.com/docs/guides/vision.

[26]The other GPT models *Gemini*, *GPT-4 with Vision*, and *Claude 3 Opus* were previously tested for the identical task, and failed.

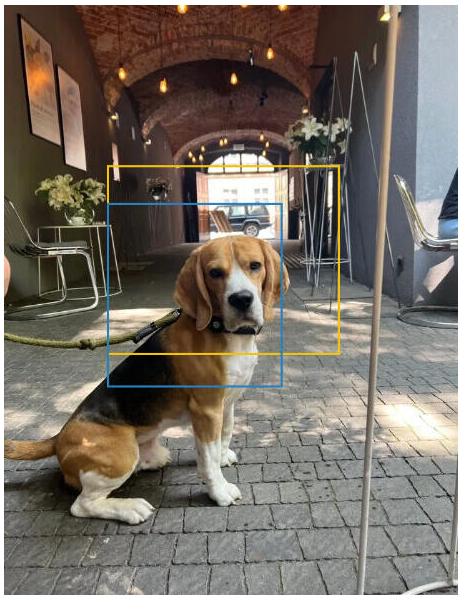


Figure 16: ChatGPT-4o Object Detection Experiment (Leo Ueno, 2024)

The two preceding sections highlights the strengths and capabilities of third-party services such as Roboflow and GPT4-V, but there are some more downsides not yet mentioned that need to be evaluated before moving forward with a third-party option. Further discussion of third-party services are found in Section 5.2.

## 2.9 Third-Party Products

Intelligent visual edge image/video devices seem to have an endless number of applications. In the following section we will take a look at some actors and their products to get an overview of the market of edge AI cameras.

The target of products such as the 'EufyCam 3' and the 'Aqara FP2 mmWave' is the private smart home sector. The i-PRO's WV-S71300-F3 has many of the same functionalities but targets enterprises instead. These products are presented in the following paragraphs and displayed in Figure 17.

### 2.9.1 EufyCam 3

The EufyCam 3 is a battery driven camera with solar a panel, only requiring two hours of sunlight to become fully charged. The EufyCam has functionality for face recognition. This is a self-learning AI which improves with time, up to an accuracy of 99.9% Eufy, 2022. To do this, the cameras communicate to an edge computing 'home base' to perform the machine learning tasks, and to save images to a hard drive. You may see an image of the EufyCam product in Figure 17a.

The EufyCam deals with low light situations in two different ways: a motion activated spotlight, turning on to film in the dark with a self-provided source of light, and a black and white night vision using six infrared LEDs to capture the video. There's also functionality to set activity areas, and to detect animals or cars.

Eufy's product does not assure privacy by deleting or obscuring images, but rather keeps them on the local area network (LAN). The user's privacy is preserved through storing the videos and images on a private in-house 1TB device, communicated from the camera devices via 2.4GHz WiFi.

The EufyCam 3 product is an edge computing device, and the Prasidh Chhadbria, director of Harvard Undergraduate AI, highlight three main advantages of the edge computing approach

(2022). (1) Time saved, as the edge devices do not have to constantly send a lot of data to the cloud; (2) Save on cost in terms of cheaper local storage, rather than more expensive cloud storage; (3) Privacy. The data is not sent to another server, but exists on the device itself so you have more control over where your data is going.

### 2.9.2 Aqara Presence Sensor FP2

The Aqara FP2 utilizes multiple passive infrared (PIR) sensors rather than a RGB camera to make detections. The FP2 may detect falls, and can localize up to 5 persons in an area, but a device may only do one of the functionalities at a time. This is due to the fact that fall detection requires the device to be mounted in the ceiling, while presence detections are only accurate when the device is mounted on the wall.

However, the devices has more functionalities. With the ability to set rules for separate zones in the area, one may toggle lights only where a person is located, for example over a workbench. Making this application part of a visual system would possibly facilitate for more applications such as being able to automatically label specific items in the area. One could combine the use of zones with a visual computing module where it would only compute and analyze data, not only when something in the frame is moving (which is typical for wildlife detection cameras (2.5.6)), but when certain rules are triggered, such that a person has been located in a certain zone for a specified amount of time.

### 2.9.3 i-PRO

Another big actor is i-PRO, providing AI network cameras to the market with edge computed people counting, face detection, and people attribute search (See their product: product wv-s71300-f3). The applications of the cameras they offer are often video monitoring and security features. i-PRO has informational web pages about surveillance policies and security. Most, if not all, of their cameras are NDAA compliant as well, which is a requirement to use them on american federal ground with regards to who produces the hardware of the system. Trusted manufacturers is a requirement for products capable of breaching privacy. See i-PRO's website for more information.

i-PRO had a big project where over a 100 cameras were installed in a arts museum in Monaco, where their cameras AI VMD that will give intrusion alerts when movement is detected in areas that should not be accessed, and virtual line crossing, giving alerts when people have crossed a digitally set line of an image. They also had AI scene change detection, detecting any changes in the image in a fixed part of the scenery. Also, AI people detection was used so to generate details about the visitors, so that guards that were interested in specific individuals had to opportunity to track specific individuals. These applications also generated statistics so the costumer had an overview of knowing in real time how many visitors were in the museum and even in the separate rooms or in front of each gallery. The cameras used in this solution were all fish-eye models, illustrating how fish-eye lenses may be the way to go for inside-application areas.

(a) EufyCam 3 (b) Aquara FP2 (c) i-PRO WV-S71300-F3

Figure 17: Smart Cameras From Aquara, Eufy and i-PRO

### 2.9.4 Viso

Viso.ai offers products for nearly every use case from abandoned luggage at airports, real time video stream weapon detections, detection of stopped vehicles, to parking space information. Their no-code platform (Viso Suite) enables a fast pipeline for developing new applications out of existing software. Viso.ai also has a lot of great articles on their web page regarding visual computing topics (see for example Boesch, 2023). The Viso Suite is marketed as a way to *Automate manual work, reduce development costs, solve scaleability, privacy and security end-to-end, accelerating every step of the enterprise computer vision development life cycle.* (This thesis is not sponsored).

### 2.9.5 VMukti

Not only does VMukti have some of the longest and most confusing product names on the market (*Real-time Edge AI based Smart Cloud Camera*), but also some of the biggest fishes in their pond of costumers. This pond includes Google, Amazon Web Services, and Microsoft[27]. One of their products, the Real-time Edge AI based Smart Cloud Camera, provides the user with a live stream of video from the camera. This may create privacy issues should the wrong user get access to the video stream, and it is likely demanding more power and network bandwidth than what it would take to only communicate the results of an analysis. VMukti's other product, the *Edge AI Based 5MP PTZ ANPR Bullet Camera VM-72BPTZ5AIVE* is listed with cutting edge technologies, including *local data processing, filtered data transfer to the cloud, and faster decision-making.* However, its hard to figure out from their website what data is processed locally, and what their decision-making is faster than.

VMukti delivers solutions for surveillance of vehicles, school buses, healthcare, shopping malls, smart cities, warehouses, campuses, examinations, premises, elections and banking. For outside monitoring, VMukti offer cameras that may connect through the mobile network, for monitoring outside remote locations.

## 2.10 Hardware

Both single board computers and microcontrollers are viable options for devices capable of on-device processing. In this section, we present some considerations regarding microcontrollers, followed by a brief presentation of three single board computers; the Raspberry Pi 3 model A+, the NVIDIA Jetson Nano, and the Radxa/OKdo ROCK 4 SE

### 2.10.1 Microcontrollers

A microcontroller differs from a single-board computer in that it lacks the general-purpose user interface and memory management functionality that a more general-purpose computer would have. A microcontroller is basically just a chip, and are generally supposed to run one task repeatedly. Some options for microcontrollers are the ESP32 or the MAX78002. The latter is optimized for convolutional neural net operations, which would speed up the inference latency. This would most likely be a great option for a person localization system.

A microcontroller must be programmed. eCos, short for "Embedded Configurable Operating System," presents a compelling alternative for microcontroller applications. Unlike traditional, general-purpose operating systems like Ubuntu or Debian designed for desktop or server environments, eCos is specifically tailored for resource-constrained systems. The lightweight and configurable nature makes it well-suited for microcontrollers with limited processing power and memory. eCos provides a modular architecture, allowing developers to selectively include only the components necessary for their particular application, minimizing the overall footprint. This level of customization enables efficient utilization of hardware resources and ensures optimal performance in environments where efficiency is paramount.

[27] Also Azure, which is owned by Microsoft

Additionally, eCos is open-source, fostering a collaborative development community (still alive today blog post), enabling developers to adapt the operating system to the unique requirements of their microcontroller-based projects. Overall, eCos stands out as a viable and adaptable choice for systems where an operating system tailored to the specific demands of microcontroller applications is required.

Another option for an operating system is MicroPython. In an article on realpython.com/, the author includes multiple compatible hardware, for example the ESP8266 and **ESP32**. MicroPython has its advantages in simplicity, ease of use, and powerful capabilities.

### 2.10.2 Single-Board Computers

A simpler-to-implement system would use a single-board computer with a general purpose, widely adopted operating system instead. There are multiple options to pick from. The best option depends on the use case sceneario and budget. We present three options for single-board computers able and suitable for person localization systems.

**Raspberry Pi** The Raspberry Pi company has provided single-board computers since 2012 that are beginner-friendly and affordable. They are considered great options for general computing devices, and are popular for novices to get an introduction to edge device programming. The SBCs are also a viable option for applications that do not demand a lot of computational power. Raspberry Pi comes in several models and series, with Raspberry Pi 5 being the newest addition.

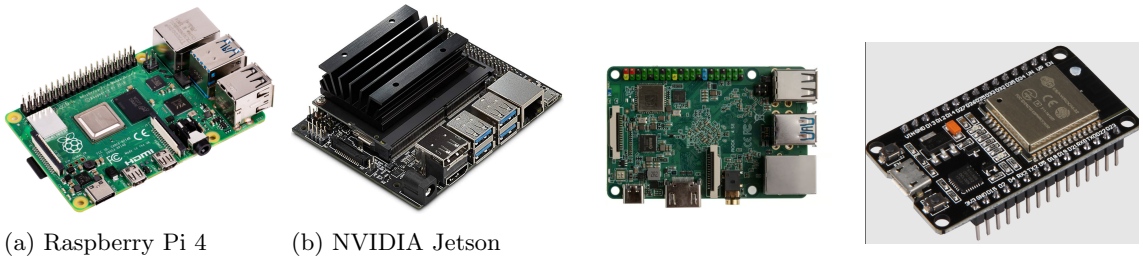
**Jetson Nano** Measuring in GFLOPs (an indirect measure for AI performance), the Jetson Nano board has a nearly 22x performance advantage over the Raspberry Pi model 3. This is reflected in the pricing, as the Jetson Nano is considerably more expensive (149USD on Amazon.com compared to Raspberry Pi at 35USD for single purchases (14 of December, 2023)). The Jetson is developed by NVIDIA, the leading company when it comes to AI and deep learning. NVIDIA is the company behind CUDA, a software toolkit for utilizing GPUs for accelerated processing.

The Jetson Nano Developer Kit has two camera serial interfaces (CSIs), enabling stereo vision. Additionally, it has a 128-core CUDA GPU, making it the computationally strongest out of this selection, and the most fit for running AI applications. However, Jetson Nanos have been hard to source in the past, and are a bit more expensive than the two other SBCs.

**ROCK 4 SE** Another consideration is Radxa and OKdo's[28] ROCK 4 SE. It is one out of four models in the ROCK 4 series. Entering the market as an alternative to the Raspberry Pi, it includes all the necessary components and interfaces for advanced video and imaging tasks, such as object detection. It also has better performance than the aforementioned single board computers. It does not have any AI-enhancing hardware, however, only a strong CPU which is not what is usually prioritized for AI-specific hardware. The ROCK 4 SE supports USB type C 12V.

There's also the upgraded ROCK 5 model A, featuring both a GPU and a 6TOPS (Tera operations per second) NPU. TOPS is not a good measurement on its own due to the fact of neural networks consisting of more than only the operations which the NPU is specialized to perform. Therefore, having computational strong hardware to back up a high throughput otherwise as well is important for an NPU to be fully utilized; if not, it will only wait for the CPU to finish its mid-inference operations (aiMotive Team, 2021).

[28]They seem to have a partnership, making them both the developer and distributors of the ROCK SBCs.



(a) Raspberry Pi 4 model B

(b) NVIDIA Jetson Nano

(c) Radxa ROCK 4 SE

(d) ESP32

Figure 18: Single board computers and a microcontroller

### 2.10.3 Specialized Hardware

Utilizing specialized hardware in edge devices is one way of accelerating AI tasks. Many tasks, however, do not require more computationally strong hardware than a CPU. There are options, however, for those that want to do real-time object detection, or train models on the edge device itself[29]. The following paragraph mentions and discuss some differences between GPUs, TPUs, and NPUs.

**Graphics Processing Units** The GPUs are mostly everything you need for an already trained machine learning model. They were developed to handle massive amounts of parallel processes. Until NVIDIA created CUDA, giving developers more direct access to the GPU, they were almost exclusively used for gaming. In most applications, which only run inferences, this is sufficiently fast to do most tasks.

GPUs may also come with enhancements to better prepare them for neural network operations. NVIDIA Volta and Turing GPUs feature Tensor Cores, which are built to accelerate AI applications, giving researchers up to 3x the training speed (Gupta, 2019).

**Tensor Processing Units** Specifically developed to accelerate machine learning training and inferences, the TPUs are much faster than a standard GPU. The TPU is and application-specific integrated circuit developed by Google for Google's TensorFlow software. It is optimized for tensors, which are extensively used in deep learning. A 0-dimensional (D) matrix is a scalar, a 1D matrix is known as a vector, and a 2D matrix is what most would call a matrix. 3D would be somewhat of a cube of numbers, while a tensor is a matrix of a higher dimension (rank) than 3. Visualizing a multidimensional space is hard because we live in, or are only able to experience/conceptualize, a 3D reality. In deep learning, the layers in the neural network are vectors, where each vector component are functions, and thus tensors. Without truly understanding how neural networks function, it is typically the general conception that these tensors represent various features. This is why adding hidden layers to the NN may increase the amount of features a network is capable of encapsulating. During backpropagation in training of a neural network, these tensor values are transformed. When one tensor is transformed, the following connected tensor in the network needs to do the corresponding operation. These transformations are not trivial, as they are dependent on the weights and biases of the network, and thus some complex operations need to be performed. These tensor operations are what TPUs are optimized to perform, both in terms of speed and energy efficiency in neural network training and inference.

TPUs are less accurate than GPUs, which reflects the differences in usage intents. Exact numbers are less important for neural networks,where 8-bits may suffice, than in gaming, where 32-bits is often preferred.

[29]Training models locally on the edge could be a plausible use case in an system implementing federated learning. Federated learning is described and discussed in Section 2.5.2

**Neural Processing Units** NPUs are specialized hardware accelerators designed for executing artificial neural network tasks efficiently and with high throughputs. Although not perfect substitutes for GPUs, they offer high performance for low power consumption, making them suitable for edge devices. With the inflated GPU prices, NPUs may be a good option for accelerating machine learning tasks.

### 2.10.4 Sensors

When selecting sensors for a person localization system, it is important to balance the need for specialization against the desire for future development flexibility. A general solution retains the ability to later specialize and optimize the system.

**Passive Infrared Sensors** Passive Infrared (PIR) sensors are widely used for presence detection. These sensors detect changes in infrared radiation within their field of view, emitted by warm objects like humans. The PIR sensor consists of a pyroelectric sensor that generates voltage when detecting temperature changes, aided by a Fresnel lens that focuses the infrared radiation onto the sensor. The generation of a voltage is not enough to determine the position of objects in the environment, but the sensor has some other advantages. As passive devices, PIR sensors do not emit radiation, making them energy-efficient and non-intrusive. The sensor could be effective in detecting precense to determine whether or not to run inference on an image. However, using PIR sensors for this purpose might biased results, potentially increasing the detection counts near the edges of the frame. PIR sensors are typically used for automated lights. A PIR sensor is diplayed in Figure 19.

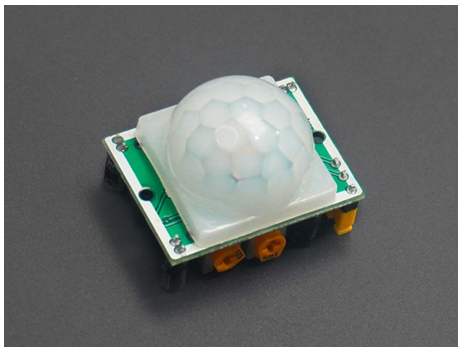


Figure 19: Passive Infrared Sensor (PIR)

**Night Vision Cameras** Night vision cameras are useful in low-light environments where traditional cameras falter. Most night vision cameras use passive infrared sensors to detect IR radiation naturally emitted by objects without actively illuminating them. However, some systems do include an IR illuminator that emits light to enhance the camera's ability to see in total darkness. Infrared imaging plays a crucial role in night-time pedestrian detection, forest fire detection, and maritime rescue. The capability to see in the low light conditions is particularly useful in settings like museums and aquariums, which often feature dim lighting to preserve exhibits or create atmospheric conditions.

There are some challenges, however. The main issue with using IR images for machine learning is that they provide different information compared to visible-light images, which might affect the training and performance of models not adapted for IR data. Additionally, these cameras are typically more expensive and require more power compared to regular cameras due to the added components like IR illuminators. This which may limit their practicality for some applications. Furthermore, their effectiveness can be compromised by any obstructions that block IR light, such as windows or thick glasses. This makes them less preferable in aquariums filled with fish tanks.

**Radio Frequency Identification** Radio Frequency Identification (RFID) is highly effective in medical and museum settings for tracking objects and people. This was seen in a study discussed in Section 2.1, where RFID was used to track museum visitor movements through a museum. RFID technology can have a range of 3-4 meters(Healey et al., 2020) which is sufficient for many indoor applications. Unfortuneately, they require wearable devices for tracking, which is a significant drawback for non-intrusive tracking. RFID's potential extends to applications like automatic medication adherence systems (see Figure 20), which could reduce the workload on healthcare staff by ensuring patients take their medications without direct supervision.

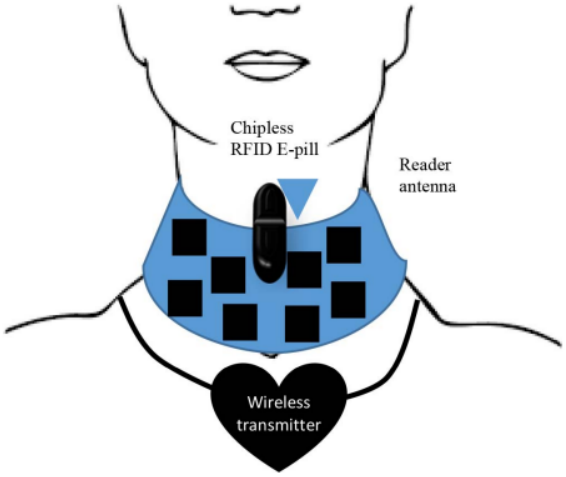



Figure 20: Proposed Design of a Pill System (Healey et al., 2020)

**Sensors for Complex Environments** In environments with obstacles that might block the line of sight, sound navigation and ranging (SONAR) can be particularly useful. SONAR sensors emit high-frequency sound waves (ultrasonic)[30] and detect their reflections to determine the changes in the environment and thus infer the presence of people. In 2009, Tarzia et al. proved human presence could be sensed with SONAR (ultrasonic) technology that was already present in most laptops in 2009. Further, it is worth mentioning that technologies like pressure sensors or systems detecting devices with Bluetooth or WiFi can also complement the presence detection/positioning in complex environments.

[30]SONAR may also emit low-frequency sound waves, which is necessary in water but unlikely to be useful in an indoor application.

# 3 Methodology

Two devices were deployed in the aquarium of "Fiskeri- og Søfartsmuseet" in Esbjerg. The devices were set in the corner of the room. The environment, angles and final achieved image quality can be seen in Figure 21.

(a) Image taken from the 'left' device

(b) Image taken from the 'right' device

Figure 21: Images Displaying the Camera Deployment Environment and Angle

## 3.1 Project Outline

A dataset was collected from the devices in the aquarium to investigate the effects of dataset quality in fine-tuning of models on the performance. The details of dataset construction is found in Section 3.2. Here, we explain how the dataset consists of three partitions: *Inconsistent*, *Consistent-1*, and *Consistent-2*.

### 3.1.1 Hardware

The selected hardware for the project was the Raspberry Pi 4B Revision 1.5 with 4 processors, with Raspbian OS and Python. A Sony IMX219 camera, which sits on the Raspberry Pi Camera Module V2.1, was used to capture images. A E337-325 4G USB modem was used to connect the SBC to the internet for remote communication with the device.

### 3.1.2 Object Detection Models

Multiple models were utilized to evaluate the effects of dataset quality. These include the following:

- Pre-trained "standard" models (DETR, YOLOv3, YOLOv9).
- YOLOv9 models fine-tuned on the inconsistent partition.
- YOLOv9 models fine-tuned on *Consistent-1*[31].
- YOLOv9 models fine-tuned on the external dataset PRW.
- YOLOv9 models fine-tuned on the external dataset CrowdHuman.

[31] Whilst all the other models were evaluated on the *Consistent* dataset, this model was only evaluated on *Consistent-2*

All models were evaluated on *Consistent-1* and *Consistent-2*. They are hereby collectively referred to as *Consistent*. Additionally, the standard models were evaluated with differing hyperparameters for input image size, and the YOLOv9 models that were fine-tuned on the inconsistent partition was evaluated with 5, 15 and 50 epochs. They were all adjusted to detect only the class *person*.

**Licenses** A successful effort was made to create a system that is free and open to use, but some conditions apply. *YOLOv9* is under a GPL-3.0 License. This is a copy-left licensing, meaning it is free to use but has the requirement that any derivative works are released under the same rights. Another algorithm discussed in this thesis, the YOLOv3, is under a free-to-use license (AGPL-3.0), but changing the code is not allowed. The final object detector algorithm discussed, DETR, is under an Apache-2.0 license, which permits users to use and modify the code to fit their needs. This could thus provide a solution for a company interested in keeping their solution hidden from competitors.

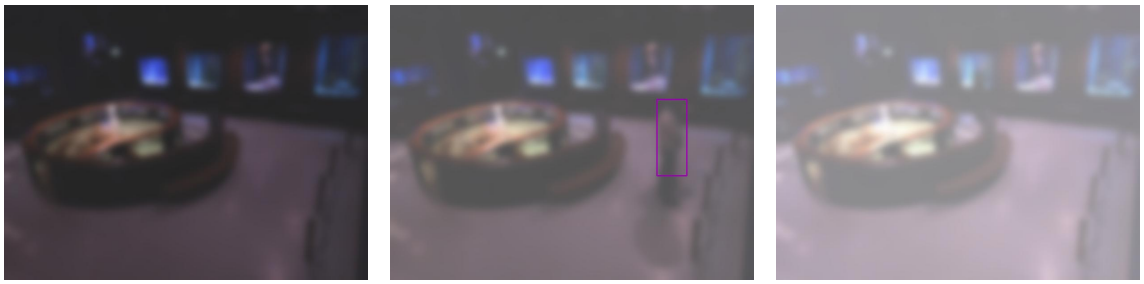
## 3.2 The FIMUS Dataset

This subsection provides an in-depth explanation of the FIMUS dataset construction, including the camera configurations, the image capturing process, and the labeling process. As previously mentioned, the dataset is split in three partitions: *Inconsistent*, *Consistent-1*, and *Consistent-2*. *Consistent* is used to denote Consistent-1 *and* Consistent-2. The camera configurations are found in the below section, the details regarding the image capturing process is found in Section 3.2.2, and the resulting image characteristics and differences are further detailed in Section 3.2.4.

### 3.2.1 Camera Configurations

**Mechanical adjustment of the aperture was ignored** The Raspberry Pi Camera v2.1 aperture can be modified by rotating the lens with a mechanical tool,configuring its depth focus. This, however, is for very close focuses. In its default position at 0 degrees, the focus is set at "infinity". Turning the lens to 45 degrees will focus the camera at 32cm. This might be neccessary for some applications, i.e. production line systems or automated recycling facilities, but it is vastly shorter than the types of applications discussed in this thesis. All the rest of the camera settings are configured programmatically through the picamera API class. The camera settings used for the consistent images are detailed in Table 2.

**Camera Configurations: Inconsistent** For the inconsistent images, the exposure mode and auto white balance mode were set to 'auto'. The automatically set values resulted in variation of image color temperature, exposure speed, and brightness, and generally lower quality images. To achieve an adequate level of brightness to label the images, the postprocessing-property *brightness* of the picamera was utilized. This resulted in artificially bright images. The brightness value was found experimentally and remotely by applying brightness and obfuscating the images before transmission. The brightness values of 60, 70, and 80 are displayed in Figure 22. A brightness value of 65 was used for the images in the inconsistent dataset. Finally, the last value set for the inconsistent dataset was resolution of the images, which was set to maximum (3264x2464).



(a) Brightness 60 (b) Brightness 70 (c) Brightness 80

Figure 22: Brightness values experimentation.

Only one person is present in each image of the inconsistent partition of the dataset. This is to simulate the probable real-world scenario of having a single technician tasked with fine-tuning a detector. The inconsistent partition is suitable for experimenting on the effects of a poorly captured but highly relevant dataset as training data on model fine-tuning. Most research in the field trains the model on a vast amount of available data, not specific to the real-world scenario. The transfer learning experiments where models are fine tuned, the images for fine tuning are almost always of great quality. The results from this analysis aims to measure the performance when highly relevant but low quality images are utilized for fine-tuning.

**Camera Configurations: Consistent** The "Consistent-1" and "Consistent-2" partitions have consistent image characteritics[32]. For these images, the camera settings were explicitly set to experimentally proven values to achieve the best image quality. The camera settings may be seen in Table 2. The consistent partition of the dataset contains images with 1-4 persons in each image. The consistent images are split in two partitions to facilitate experiments using one partition as training data and the other for evaluation. This is suitable for testing how a well-captured and highly relevant dataset may function as training data for model fine-tuning.

| **Raspberry Pi Camera Property** | **Value** |
|---|---|
| *awb_gains* | (1.5, 1.5) |
| *awb_mode* | off |
| *brightness* | 55 |
| *contrast* | 0 |
| *exposure_mode* | off |
| *exposure_speed* | 79989 |
| *framerate* | 6 |
| *iso* | 640 |
| *sensor_mode* | 3 |
| *shutter_speed* | 80000 |
| *resolution* | (3264, 2464) |

Table 2: Camera Settings for the Image Capture of Consistent Images

All the settings are explained in greater detail in appendix B. Note that the ordering matters when setting the picamera properties. The ordering used to achieve consistent image capturing for project of this thesis is illustrated in Figure 43 in appendix A.

[32]Their differences are detailed in Section 3.2.2

### 3.2.2 The Image Capturing Process

Images were captured using a script that sequentially captured images, storing them directly onto a 32GB micro SD card installed in the device. This local storage approach was adopted to eliminate data transmission costs and potential security risks associated with potentially transmitting sharp, identifiable images over the internet. Instead, should unwitting individuals wander into the frames during the capturing process, these images were deleted manually once they had been transferred to the computer.

The class *Image* from the python package *PIL* was used to store the images, and to address the limited storage capacity on the computer storing the dataset images, the images were stored with a save quality value of 90.

The dataset was built capturing images while no other visitors were present in the aquarium expect those who'd volunteer to participate. This was due to the restriction detailed in the project scope (Section 1.3). A way to cancel image capturing was needed in case visitors entered the room. The simplest way of achieving this would be to pull the plug. This was challenging, however, as the devices and their power supplies were mounted high on the wall. The selected approach was to SSH[33] into the devices to start and stop the image capturing process.

An attempt was made to implement the lightweight messaging protocol MQTT for the devices to subscribe to a topic, providing them commands to perform preset operations such as capturing and storing sharp images. Through the use of a phone application such as *EasyMQTT* (for iOS), this could've allowed for command transmissions to the devices from the phone, simplifying the image capturing process and allowing for image capturing without having to bring a laptop. However, the development progress stagnated due to significant resources being diverted to resolving issues related to authentication token generation and configuring a broker accessible via a public domain name. The SSH tunneling approach was thus deemed to be sufficient.

Every once in a while, when a lot of visitors entered the room, the devices were demounted and the SD card plugged into the computer to extract the captured images. This process resulted in slight variations in camera angles upon remounting, as replicating the exact original setup proved difficult.

**Setbacks in the Image Capturing Process** There were a few setbacks for the image capture process that are worth mentioning. These are highlighted in bold and discussed in the following paragraphs.

Firstly, **settings had to be completely readjusted** between the environment the device was tested to the environment the devices were deployed. See Figure 23 to see how a shutter speed of 20 000 is in the office versus in the aquarium. This was after auto-settings were disabled and the images in theory should only vary slightly due to the slightly lower light levels in the aquarium. However, the Raspberry Pi Camera Module seems highly sensitive to its settings.

[33](Secure Shell (SSH) is not detailed in this thesis, see Section 1.3)

(a) Shutter Speed of 20 000ms in the Office

(b) Shutter Speed of 20 000ms in the Aquarium

Figure 23: The Effect of the Same Shutter Speed in Different Environments

**A few images were corrupted** and thus were not be used. This was another minor setback. The occurences of corrupted images seems highly random, and only happened 4-5 times in total for the dataset capture. Examples of corrupted images can be seen in Figure 24. These were easy to detect and thus did not create further trouble.

(a) Corrupted Image 1

(b) Corrupted Image 2

Figure 24: Examples of Rolling Shutter Artefacts

The image in Figure 24a reveals the camera has a rolling shutter[34]. The way images are captured with a rolling shutter, is by capturing the image from the top to the bottom. The camera is constantly renewing rows of image data. The corrupted rows in 24a are purple, while at the bottom the camera has resumed its correct operation. The corrupted images are similar in size to the normal images, so the way to remove them and retake an image would be to inspect the image data values instead to determine it is mainly purple. The corrupted images were removed from the dataset. Today, Raspberry Pi offers a camera with a global shutter to alleviate the aforementioned bug. As mentioned on their website, the global shutter camera is *a specialised 1.6-megapixel camera that can capture rapid motion without introducing artefacts typical of rolling shutter cameras. Ideal for fast motion photography and machine vision applications.*

The final significant challenge encountered in the image capturing process phase was the unexpected **nonlinear increase in image capture time** following adjustments to camera settings. This change occurred when transitioning from capturing the "Inconsistent" dataset partition to the "Consistent" dataset partition. Initially, capturing an image required approximately two seconds; however, this duration increased to about eight seconds per image post-adjustment.

[34]The fact the camera has a rolling shutter was confirmed by looking in the camera hardware documetation.

The increased capture time significantly affected the volume of data collected, resulting in a larger "Inconsistent" dataset partition of 2,637 images, compared to the "Consistent" partition, which comprised only 757 images. The disparity in dataset sizes stemmed directly from these increased capture times.

Due to time constraints and the prioritization of data collection over script development, the underlying causes of the increased capture time were not fully investigated. The decision to prioritize image collection was driven by the need to secure a sufficient quantity of data for effective model training, despite the suboptimal capture conditions. Regardless, the image capturing process resulted in the dataset partitions named Inconsistent, Consistent-1, and Consistent-2, which are described in the following paragraphs.

**Image Capturing Process: Inconsistent** *Total number of images: total 2637. 1 subject.*

The first iteration of image capture, and what resulted in the *inconsistent* partition of the dataset, was made with non-optimized camera configurations. To sufficiently brighten the images, the picamera.brightness attribute was set to 65. This is a postprocessing operation, which gave brighter but also artificially lit images. Also, the camera would sometimes focus on the bright fish-tanks in the museums, rendering the rest of the image rather dark. This was an effect of the awb mode and exposure mode being set to auto, and led to images of varying brightness and color. These images were still included in the dataset however, as images seen as suboptimal to the human eye may still be useful to the training of detectors. These images are useful to inspect the impact of captured image quality on inference performance.

**Image Capturing Process: Consistent-1** *Total number of images: 292. 1-4 subjects.*

For the second image-capturing session, the camera configurations had been more thorougly tested to obtain more consistent images in terms of colors and brightness. This means using non-auto auto white balancing and exposure settings, and reducing the amount of post-processing brightness adjustment. Also, some friends were invited in this session. Due to a reduced post-processing brightness augmentation, the exposure speed had to be increased to get sufficient light in the images. This meant more unclear outlines of moving subjects in the frame. It also meant more time was spent capturing and storing each image. This increased from $1.3\frac{s}{image}$ to $6.3\frac{s}{image}$, which means the time available for image capturing was spent less productively than with the previous camera configuration. Depending on the impacts of image consistency on inference accuracy vs. amount of training data, capturing with a higher exposure speed and then post-processing the images to be brighter might be the better solution. Also, augmenting the brightness might only slightly impact the model performance. This is because a model may see slight differences in pixel-level values invisible to humans, thus enabling it to still recognize the patterns of human outlines. A sufficiently bright image would still be required for the sake of model verification and ground truth obtainment (by human annotators).

The camera was repositioned three times during this iteration of image capturing. This is a drawback as it complicates the process of mapping the person positions in the images to real world locations in the aquarium. This is because the positions are represented as x,y values from the corner of the image, and for a person standing at exactly the same position in two images, the x,y-values will differ if the camera position has moved. Serving as a dataset for machine learning applications and not for analytics generation based on real world positions, this was not an issue.

**NoIR Camera** 60 images were captured using a Raspberry Pi NoIR camera module version 2.1 to determine its efficacy in enhancing human detection under low-light conditions. The "No" in NoIR signifies its lack of an infrared filter. It was hypothized by the author that this meant the camera could then operate with a lower shutter speed, which showed promising results in initial tests. However, once deployed in the aquarium, this proved to be wrong. The NoIR camera is said to give the ability to look in the dark *with infrared lightning*. Despite its potential, the noir camera was used as a regular camera module thereafter, capturing a different angle than the first device,

for the remaining image capturing iterations. The 60 images were not used in the project, as the models trained on inconsistent data had already been trained.

**Image Capturing Process: Consistent-2** *Total number of images: 465. 1-2 subjects.*

In this iteration of image capture, a second camera capturing the left side of the aquarium was also deployed. This *left* device has some differences in colors, due to using a no-IR camera instead of the regular. These differences are seen in Figure 25. Additional differences between *Consistent-1* and *-2* are described below the figure.

(a) *Consistent-2 Left* Image, Captured with a noIR RPi camera

(b) *Consistent-1 Right* Image, Caotured with a regular RPi Camera

Figure 25: Example Images From *Consistent-1* and *Consistent-2*

*Consistent-1* is from a group of 4 people in almost all the images, moving around the aquarium and talking in a group in various location. All images are taken only from the right device, all with close to the same angle. The subjects are 3 persons of approximately 1.80m height, 2 female and 2 males and various light and dark clothes, all wearing pants. *Consistent-1* are the closest representation of the images the device will be capturing in the experimental setting.

*Consistent-2* has many images with a single subject, then two subjects. The images are primarily from the right device, but contains some taken from the left device as well (an example image from the left device is seen in Figure 25a). The subjects in *Consistent-2* are both approximately 1.80m, one wearing glasses. One person wore a white tshirt, the other wore a black tshirt. Both wore shorts.

### 3.2.3 Labeling

The detector requires precise ground truth positions of persons for training, validation, and testing. This data is obtained through labeling, also called annotating, the images. For object detection, this means to extract (typically) 5 values for each object of interest in the image.

There are multiple formats in use. The YOLO format is to save the *class id*, *x center*, *y center*, *width*, and *height* for each detection. These are saved on separate rows of a txt file, each corresponding to an image. All images have one txt file containing the labels for the image. All the label values (except class id) are normalized, so

$$(x, y) = (0.0, 0.0)$$

would correspond to the upper left corner of the image, and

$$(x, y) = (1.0, 1.0)$$

would correspond to the lower right corner of the image. The labels are stored in a folder called *labels*, and the images are stored in a folder called *images*.

Other formats may store all the bounding boxes in one json for the whole images folder, and may contain more information regarding the detections[35]. The format of this thesis project and dataset is the YOLOv5-format[36].

[35]Other information on detection may include if the object is occluded or not, or if the person id should be saved to facilitate person reidentification and tracking.

[36]YOLOv5 introduced annotation file format *txt* instead of *xml*, and has five parameters per line.

To expedite the labeling process, the images were initially processed using a pre-trained YOLOv9 model on the COCO dataset, rather than manually labeling each image. Out of the 2939 images in the first-iteration dataset, the model prosuced 1863 detections that needed verification. This includes modifications, deletions, and additions to the annotations. The remaining 1076 images, which had no initial detections, required manual labeling from scratch.

Additionally, validation of the annotations uncovered specific errors: in 74 images, moving seaweed in one of the fish tanks was mistakenly identified as a human due to its human-like movement and shape (see Figure 26). In another instance a person carrying a ladder was incorrectly recognized as one person carrying another.


Figure 26: Sometimes, the Seaweed is Deemed More Likely to be a Person than the Human

**Label Studio** "Label Studio" was used to label the images. This online tool has multiple uses. One is to set up a machine learning backend for automatically generating predictions for unlabeled images. An attempt was made to set up the machine learning backend, using a Grounding DINO model (Liu et al., 2023). The setup of this backend was not trivial, however, and this use of label studio was foregone. Instead, the images were inferred on outside label studio, the labels converted, and then imported. This was a less 'automatic' approach but nevertheless much more effective than troubleshooting the machine learning backend solution. The label-studio tool was then used to manually modify, delete, and add annotations to the images. Finally, the annotations were exported and converted to the YOLO format. Today, it is possible to convert directly to YOLO format. This was not previously possible, and was done manually with a script which can be found along other utility-scripts under Other/Code/Utils on the GitHub of the author of this master thesis.

### 3.2.4 Dataset Characteristics and Applications

The FIMUS dataset consists of in total 3394 images, of which 2637 are in the inconsistent partition, 292 are in consistent-1 and 465 are in consistent-2 (*Consistent* total: 757 images). The dataset is well suited for the task of measuring and analysing the impact of image quality on the performance of object detectors, and whether a model performance on a general dataset is a better or worse indication of real-life performance than a specialized but poorly captured dataset.

The standard for train-validation-test splits is 60-80% training data, 10-20% validation data, and 10-20% test data. If the images in Inconsistent are used for training and validation, and the images in *Consistent* are used for testing, we would get a have 78% data for training and validation, and the remaining 22% of data for testing.

Another application of the dataset would be to use *Consistent-2* for training and *Consistent-1* for testing to measure the impact of a small, but highly relevant dataset from two different angles on a fine-tuned model performance on images from the same environment and with the same settings. This would give a split of 61% for training and validation, and 39% for testing.

## 3.3 External Datasets

This project utilizes multiple external datasets for developing and testing the object detection models. Each dataset was selected based on its relevance to the project, specifically for containing labeled images of the person class, and they vary in the number of images, capturing angle, and image diversity.

### 3.3.1 Common Objects in Context (COCO)

The COCO dataset is a large dataset of 118 000 images and 80 different classes. The COCO-2017 train dataset was used to pre-train the models. The COCO-2017 validation dataset was used to evaluate the performance of the finalized models, as is industry standard.

Figure 27: COCO Dataset Example Images

Figure 27 is a great example of the widespread nature of the COCO dataset images. This makes for a great dataset for pre-training, as the trained model will have knowledge of a wide array of objects. It may then be wise to fine-tune such a model to a more specific use case, so the model can see more of the specialized data.

COCO was introduced in the article of Lin et al. (2015).

### 3.3.2 CrowdHuman

CrowdHuman, the largest dataset used, focuses exclusively on images where people are the main subject, contrasting with COCO's broader class range. This dataset was employed to assess how additional data might enhance model performance, with experiments conducted across various training data volumes.

Figure 28: CrowdHuman Dataset Example Images

The CrowdHuman dataset was presented in the article of Shao et al. (2018).

### 3.3.3 Person Reidentification in the Wild

Person Reidentification in the Wild comprises 11,816 images of pedestrians and aligns closely with our application needs as it exclusively contains images of people. This dataset's relevance is heightened by the presence of occlusions and the similar scale of persons to those detected in the aquarium setting. The dataset contains 932 individuals, annotated in 34 304 separate annotated boxes. Although designed to facilitate the development of reidentification applications, this functionality was not utilized in this project (refer to the project scope in Section 1.3 for details).

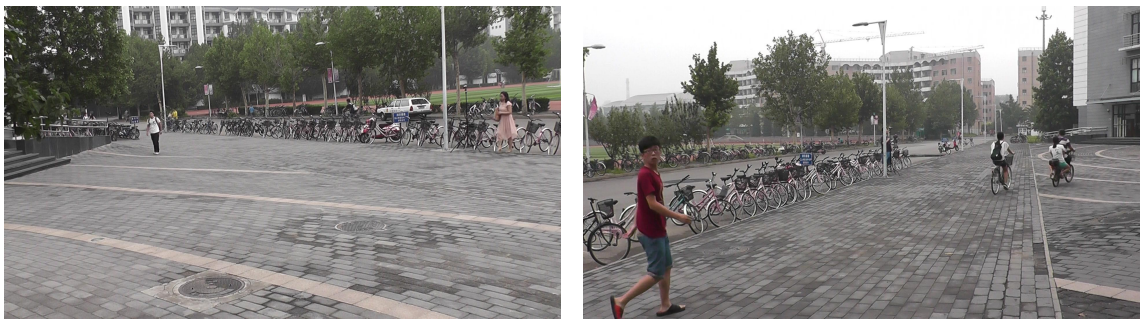
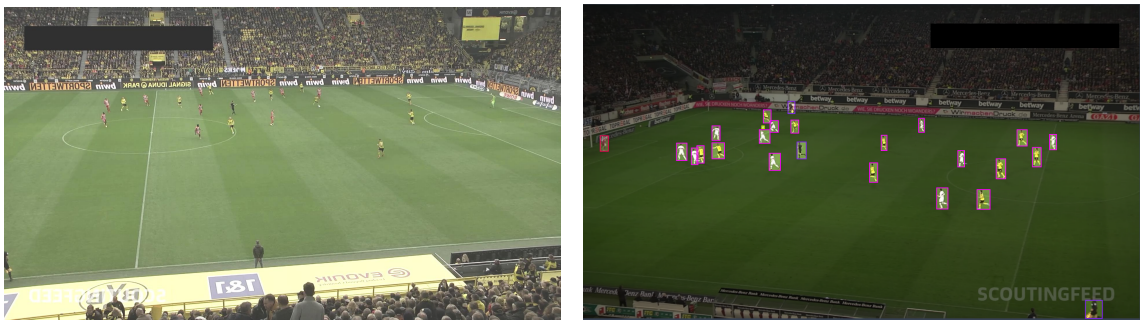


Figure 29: PRW Dataset Example Images

The PRW dataset was presented in the article of L. Zheng et al. (2017).

### 3.3.4 Football Players Detection

The football-players dataset, characterized by its uniform perspective and consistent lighting and quality, was introduced to determine whether model improvements derive solely from specializing to single-class data or if the specialization's quality and relevance are crucial. It also provides a clear contrast in dataset characteristics, aiding in attributing performance differences to dataset nature rather than other confounding factors. A weakness that may confuse a model under training may be that the audience are not labeled. This is illustrated in Figure 30b. Therefore, fine-tuning a model on this dataset may result in a model that ignores persons of such tiny scale.




(a) Example Image (Unlabeled) (b) Example Image (Labeled)

Figure 30: Football Players Detection Dataset Example Images

The Football Players Detection dataset is available at Roboflow.

## 3.4 Model Training

In the thesis project, YOLOv9 was pre-trained on the COCO dataset and then optimized by fine-tuning on specialized data. The process of fine-tuning a pre-trained model is known as transfer learning (see Section 2.7.5). The rest of this section outlines the processes and choices taken in the training process of these models.

### 3.4.1 Hyperparameter Tuning

One approach to tune the hyperparameters is to utilize Autogluon, an auto machine learning library. Installation may be tricky, but one may use this guide for installation: AutoGluon guide. Note that in some cases, 'pip install autogluon' must be evaluated twice. Further, this guide could be used to tune model hyperparameters.

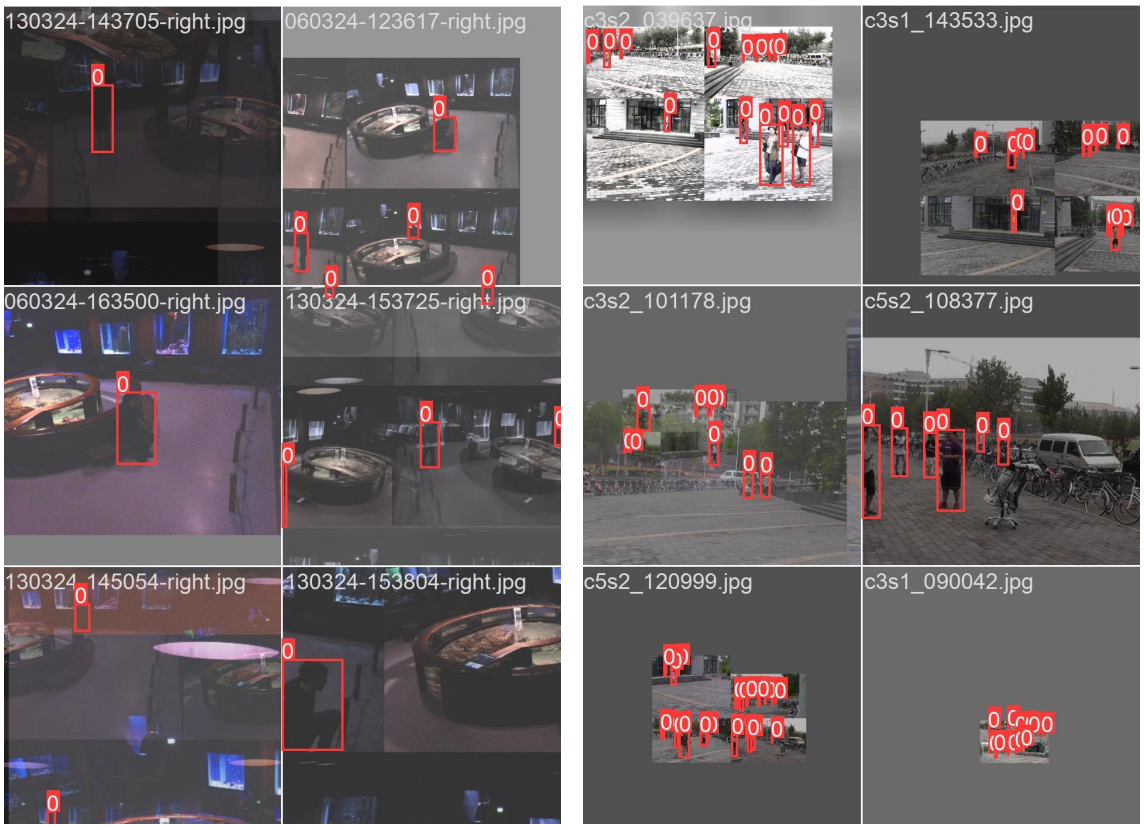
Another great guide for hyperparameter tuning was found on Kaggle. However, this would entail optimizing the hyperparameters for each of the datasets used in this project. Hyperparameter tuning is a time-consuming process, and the author of this thesis did not have the necessary means to perform this task.

For this project, a simpler solution and less effective approach to hyperparameter tuning was adopted. This was to use the standard out-of-the-box hyperparameters. This is one source of bias in this project, as some standard parameters might be optimized for a certain dataset size. The fine-tuned models would achieve better performance had the hyperparameters been tuned. Due to the large scope of this project however, hyperparameter tuning/hyperparameter optimization was not prioritized.

One modification was made to the hyperparameters however, which was the number of epochs. A major challenge and setback for the model training was a mistake made in choosing a slightly premature YOLOv9, for which the validation process was not yet implemented. This means that the models were essentially trained blindly, without providing the data to indicate whether the

models overfitted or could benefit from more training. This is apparently still (24-05-2024) a github issue for YOLOv9. A method to manually compute losses for the models was not implemented. Instead, the models were trained for 5, 10, 20 and 50 epochs, which are quite short periods.

Mosaic data augmentation was used to enhance the datasets for the training process. This consists of several processes to create more training data from the the available images. Example images resulting from mosaic data augmentataion are seen in figure 31.




(a) Example From the FIMUS Dataset (b) Example From the PRW Dataset

Figure 31: Mosaic Data Augmentation

### 3.4.2 Google Cloud Services

Google Colab was used to train the models on VMs with GPUs to speed up the training process. The training thus took about 4.5 minutes per epochs for the FIMUS inconsistent dataset partition with 2636 images, and nearly 15 minutes per epoch for the 11815 images dataset PRW. Training on the CrowdHuman dataset took nearly 20 minutes per epoch, with its 15000 training images.

The pros with Google Colab is the possibility to borrow computational power without having to invest in expensive computers. For this project, 200 compute units, costing about 25USD, were sufficient for the training process, making it a cost-effective solution for developing a specialized system. Google Colab also allows for seamless file exchange by mounting a Google Drive to the VM, though the solution necessitates giving explicit consent to Google Drive that they may see and download all your data, including photos on Google Foto. For the convenience the service brings, however, this potential privacy infringement may be acceptable[37]. Another pro is the possibility to co-operate on notebooks. However, personal experience suggests that collaboration can be problematic, as changes made by one user require saving and refreshing the notebook for others to see

However, Google Colab has its drawbacks. Like many cloud services, it requires a stable internet connection to the VM, which can frequently be lost, resulting in data loss if the session disconnects unexpectedly. This led to training models in 10-epoch intervals and regularly saving weights to Google Drive to prevent significant data loss and repeat work, highlighting a critical challenge in relying on stateless cloud computing environments for machine learning projects.

### 3.4.3 Validation Data

During the development of this thesis, the focus was not on hyperparameter tuning. Consequently, a strategic decision was made to forego the use of validation data during the training phase. Typically, validation data serves a critical function in monitoring a model's training progress and diagnosing issues such as overfitting or underfitting. However, given that the connectivity issues outlined in the preceding paragraph, it was deemed more efficient to utilize all available data for training to ensure results. Another reason was the intention to fine-tune on the *Consistent-2* dataset partition, which comprises only 465 images. Given this limited dataset size, it could be disadvantageous to maintain a separate validation dataset.

Omitting the validation step allowed for uninterrupted training sessions but required careful management of the training process to avoid potential overfitting. As a result, the trained models were directly evaluated using the test set. This approach inadvertently increased the overhead by generating multiple models, each requiring individual evaluation. However, this method also eliminated the need for a final retraining phase, which would typically combine training and validation datasets to optimize the model post-validation.

## 3.5 Model Overview

Following is an overview of the rationale behind the selection of models for this thesis project:

1. YOLOv3 (not fine-tuned): Included as a baseline to demonstrate the evolution in object detection technology. YOLOv3, being a widely recognized and earlier version, allows for a direct comparison with its more advanced successors, highlighting improvements over time in detection accuracy and processing speed.
2. YOLOv9 (not fine-tuned): Chosen for its state-of-the-art performance and real-time detection capabilities. As the latest iteration in the YOLO series at the time of this research, YOLOv9 brings enhancements such as better handling of varied object sizes and improved generalization from more complex background contexts compared to its predecessors.
3. YOLOv9 Fine-Tuned on *FIMUS Inconsistent*: This model variant was fine-tuned on a specifically challenging subset of the dataset to assess how well the model can adapt to lower-quality data.

[37]This perspective is ironic and hypocritic, see Section 5.4

4. YOLOv9 Fine-Tuned on *FIMUS Consistent-2* (evaluated on *Consistent-1*): This setup tests the model's effectiveness in a controlled experiment where it is fine-tuned and evaluated on high-quality, consistent datasets, providing insights into the best-case scenario performance in ideal conditions.
5. YOLOv9 Fine-Tuned on *CrowdHuman*: Fine-tuning a model on this data provides insights in how a dataset full of crowded scenes may influence model performance in busy public spaces.
6. YOLOv9 Fine-Tuned on *PRW*: The PRW may be characterized relative to the deployment scenario by the similar scale of persons. A good performance of this model could indicate fine-tuning on data with similar scale objects may be advantageous.
7. YOLOv9 Fine-Tuned on *Football Players Detection*: With a totally different scale and the additional challenge of audience persons not being counted as persons, this dataset may see how a slightly less-relevant than the other two datasets may influence the performance of object detection models.
8. DETR with a ResNet50 backbone (not fine-tuned): DETR represents a different architectural approach using transformers on top of of a CNN backbone. Including this model allows for evaluating how transformer-based models compare against the more traditional approaches like YOLO in handling complex object detection tasks.
9. DETR with a ResNet101 backbone (not fine-tuned): This variant, equipped with a more powerful backbone, provides a deeper insight into the scalability and performance improvements possible with more extensive neural networks.

The diverse selection of models and datasets is designed to cover a broad spectrum of image relevancy to the deployment scenario, enhancing the findings validity and robustness. This approach not only tests the limits of current technology but also sets a solid foundation for identifying areas of improvement in future research and development of object detection systems.

## 3.6 Model Evaluation

As mentioned in 2.6.5, there have been multiple ways for object detection model evaluation. The most widely used has been to fix the confidence threshold, and average over 10 IoU thresholds from 0.5 to 0.95 in steps of 0.05. This is hereby denoted as COCO AP. For this thesis, both COCO AP and the more computationally expensive where both confidence and IoU thresholds are varied has been implemented to see if there's a different outcome for model evaluation based on which version of the evaluation metric is chosen. The more computationally expensive version is denoted as Vary-Both AP.

What input image size is optimal depends on the dataset and use case, and should be tested for a given scenario. According to James Gallagher, to increase the input image size will augment the accuracy of a model:

> We trained our model on images with a size of 640, which allows us to train a model with lesser computational resources. During inference, we increase the image size to 1280, allowing us to get more accurate results from our model. (James Gallagher, 2024)

This postulates that even though a model is trained on images with size 640, more accurate results may be obtained by increasing the input image size during inference. This hypothesis was tested in the project, see Section 4.1.3. Here, the input image sizes 320, 640 and 1280 were compared in terms of accuracy and inference latency. Except from this experiment, the models in this project inferenced with input image size of 640.

The pre-trained weights were available in multiple sizes. The largest weights-file, called 'yolov9-e', is what has been used for this project. These weights are available for download on the Yolov9 Github repository. An assessment regarding the differences in accuracy of the different available pre-trained weights was also made. This includes the available pre-trained weights as of May 2024, yolov9-m, yolov9-c, and yolov9-e.

## 3.7 Ethical Considerations

In the deployment of advanced machine learning technologies for visitor positioning and engagement analysis, this research proactively addresses privacy concerns through the implementation of image obscuration techniques. These methods ensure that no personally identifiable information is captured or communicated, thus significantly reducing privacy risks associated with visitor tracking in cultural spaces such as museums and aquariums.

### 3.7.1 Privacy by Design

At the forefront of our ethical approach is the principle of "privacy by design." This concept involves integrating privacy into the development and operation of our tracking technologies from the outset, rather than as an afterthought. By employing image obscuration techniques, such as real-time pixelation or silhouette generation, we ensure that the visual data processed by our system remains anonymous. This method effectively eliminates the possibility of identifying individual visitors from the captured data, thereby safeguarding their privacy.

The application of these privacy-preserving techniques negates the need for explicit consent from visitors for two primary reasons. First, the anonymization process occurs instantaneously as the data is captured, meaning no identifiable information is ever stored or analyzed. Second, the focus of the research is on aggregate behavior patterns rather than individual actions, further distancing the study from privacy concerns.

### 3.7.2 Ethical Use and Data Protection

Ensuring the ethical use of technology extends beyond privacy considerations to include the responsible handling and protection of any data generated by the system. Although the data is anonymized, we are committed to maintaining high standards of data protection. This includes secure data storage, limiting access to authorized personnel, and employing robust data management policies that comply with relevant data protection laws and guidelines.

The utilization of anonymization techniques also reflects the commitment to minimize any potential impact on visitor behavior and the overall museum or aquarium experience. By ensuring that the tracking system is unobtrusive and does not compromise privacy, we aim to maintain the integrity of the visitor experience, allowing individuals to engage with exhibits without concern for their privacy.

### 3.7.3 Transparency and Accountability

While the technical approach effectively addresses privacy concerns, maintaining transparency about the use and purpose of tracking technologies is still essential. Information about the tracking system and its privacy-preserving nature will be made available to visitors, ensuring they are informed about how data is used to enhance the visitor experience.

Furthermore, the project will adhere to an ongoing ethical review process, ensuring that all aspects of the research remain aligned with ethical best practices and respond to evolving technological and societal standards.

In summary, by prioritizing privacy through the use of image obscuration techniques and adopting a comprehensive ethical framework, this research aims to advance the understanding of visitor engagement in a manner that is both innovative and respectful of individual privacy rights. This approach sets a precedent for the ethical application of machine learning technologies in cultural institutions, balancing the benefits of visitor behavior analysis with the imperative of protecting privacy.

## 3.8 Heatmaps

Heatmaps are a powerful visualization tool that can provide insights into visitor behavior patterns and engagement levels within a museum or aquarium setting. By aggregating anonymized data from the tracking system, heatmaps can reveal areas of high visitor activity, peak visitation times, and popular exhibit locations. These visual representations offer valuable information for museum staff and curators, enabling them to optimize exhibit layouts, plan interactive experiences, and enhance visitor engagement. For this project, heatmaps were attempted to be created using 3 different python packages.

**On the Heatmap Creation Attempts** The first attempt was made asking chatGPT 4 to provide the code. The AI chose to draw circles using the python package "OpenCV", which without modifications did not render satisfactory results due to the lack of depth visualization (see Figure 32a). In this heatmap, red circles were drawn at every detection. This renders a heatmap where it is impossible to see the scale of detections, apart from simply where the detection was made. An idea was formulated to apply a Gaussian filter to the detections to spread the detections out on a grid-like surface of the floor of the room.. But instead of tweaking a suboptimal solution, the attempt was then made to use the modules from Ultralytics to create the heatmap.

(a) Unsatisfactory ChatGPT-4 Heatmap

(b) Unsatisfactory Ultralytics Heatmap

Figure 32: Heatmap Development Drafts

Ultralytics is a company from Los Angeles and the same company that developed YOLOv5 on which the YOLOv9 is built upon. Amongst their many ML application are the modules specifically for creating heatmaps. However, the solution neccessitates a detector model to make inferences live, and optional arguments to pass pre-made inferences was not found. An attempt was made to modify the code and pass the detection in the format the modules were expected, but the result was unsatisfactory. This may only be a misunderstanding in how to use the module, but getting a visualization of more than the few spots shown in 32b was difficult. Again, instead of struggling further with a third-party module, an attemt with another module was made.

### 3.8.1 Supervision Heatmaps

Supervision is a module created by Roboflow, to make reusable and user friendly computer vision tools. It is designed to be model agnostic. The github repo is found here.

The solution incorporating Supervision rendered satisfactory results. An example image is provided in Figure 33 where a heatmap has been created for a single day in May. This solution supports generating heatmaps from data in a pandas dataframe, allowing for filtering the dataframe to generate the preferred heatmaps based on any variable. This could be the interesting times of the

day aggregated over a month, (e.g. every weekday from 10-11), or for a given time interval (e.g. week 39, 2024).

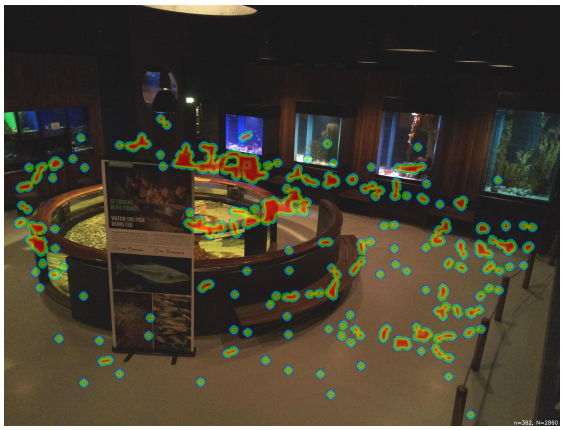


Figure 33: Final Heatmap Example

# 4 Project Results

This section presents the results of the person localization system implemented in the thesis project work. This includes a diverse range of experiments relevant for the objectives of this thesis. The results are discussed with regards to their implications and significance in the context of the research questions and objectives outlined in Section 1.5, and in the broader context of the field.

**Metrics and Notations** The tables 3, 4, and 5 shows the AP scores when the models are evaluated with "COCO AP". This means to only average over 10 IoU thresholds, while keeping the confidence threshold fixed. As mentioned in Section 2.6.5, this has been the de-facto method of calculating average precision since its introduction with the COCO dataset challenge in 2017. The confidence threshold is kept fixed at 0.5, which is the default value for the COCO challenge. In this thesis, we denote COCO AP as *AP*, and the other method of calculating average precision where both the IoU and COCO thresholds are varied is denoted as *vary-both AP*.

Additionally, for simplicity, the models that are not fine-tuned are denoted as *standard* models. The weights of these models are directly available for download online. The fine-tuned models are fine-tuned on these standard weights.

## 4.1 Object Detection Model Evaluation

The characteritics of the dataset partitions allow us to execute several experiments. The experiments are designed to answer the research questions and objectives outlined in Section 1.5. Various evaluation jupyter notebooks used in this thesis to evaluate the models are found at this github.

### 4.1.1 Fine-Tuning On *Consistent-2*, Testing On *Consistent-1*

The test-set should cover as many situations as possible that the detector may face after deployment, and be as diverse as possible. The *Consistent-1* and *Consistent-2* are similar, but have some minor differences (differences were detailed in Section 3.2.2). These differences may affect the results, and obscure whether the results are due to the differences in the dataset sizes or the differences in the dataset content. Regardless, using a smaller test-set reduces validity of the results. It is necessary, however, when a portion of the test-set is needed for fine-tuning. Since the same data must not be used for training and validation or testing, the models displayed in 3, which were fine-tuned on *Consistent-2*, are only evaluated on *Consistent-1*.

| Model | AP50 | AP75 | AP90 | mAP50-95 |
|---|---|---|---|---|
| Standard | 0.982 | 0.959 | 0.935 | 0.959 |
| Fine-Tuned 5 Epochs | 0.969 | 0.881 | 0.645 | 0.815 |
| Fine-Tuned 10 Epochs | 0.976 | 0.889 | 0.660 | 0.823 |
| Fine-Tuned 20 Epochs | 0.971 | 0.883 | 0.648 | 0.821 |

Table 3: Comparison of Various YOLOv9 Models Fine-Tuned on Consistent-2 (465 images) and Evaluated on Consistent-1 (292 images)

The scores in Table 3 shows the standard model outperformed the fine-tuned models. This may be due to the low number of epochs they are trained for. Another reason to the fine-tuned models performing poorly may be due to the simplicity of the dataset. The worsened results by fine-tuning is partly unexpected, but may be due to catastrophic forgetting, which was tried to be avoided by freezing the backbone of the YOLO models. More discussions regarding the results are found in Section 4.1.4, which displays similar results. We will first further investigate the characteristics of the test-set.

### 4.1.2 Test-Set Exploration

The following table presents the performance metrics of YOLOv3 and YOLOv9 on different test sets composed from the images of the FIMUS dataset. These models are pre-trained with COCO images, and was never trained nor fine-tuned with FIMUS data. The table highlights the Average Precision (AP50-95) and Average Recall (AR50-95) metrics across these test sets.

| Model | Test Set | Sample | Max Persons | AP50-95 | AR50-95 |
|---|---|---|---|---|---|
| YOLOv3 | Consistent-1 | n=292 | 4 | 0.826 | 0.727 |
| YOLOv3 | Consistent-2 | n=465 | 2 | 0.795 | 0.847 |
| YOLOv3 | Consistent | n=757 | 4 | 0.681 | 0.734 |
| YOLOv3 | Inconsistent | n=2637 | 1 | 0.840 | 0.851 |
| YOLOv3 | FIMUS dataset | n=3394 | 4 | 0.770 | 0.801 |
| YOLOv9 | Consistent-1 | n=292 | 4 | **0.959** | 0.826 |
| YOLOv9 | Consistent-2 | n=465 | 2 | 0.910 | 0.949 |
| YOLOv9 | Consistent | n=757 | 4 | 0.945 | 0.865 |
| YOLOv9 | Inconsistent | n=2637 | 1 | 0.929 | **0.971** |
| YOLOv9 | FIMUS dataset | n=3394 | 4 | 0.936 | 0.921 |

Table 4: Comparison of YOLOv3 and YOLOv9 on Various Quality Test Data

Like previous research has shown, YOLOv9 outperforms YOLOv3. The YOLOv3 model was included in this experiment to see if it would outperform its successor for any of the data, which it did not.

The consistent test sets (*Consistent-1*, *Consistent-2*, *Consistent*) represent subsets within the *Consistent* partition. Analyzing the results on these sets shows some variability. For instance, the AP50-95 for YOLOv9 drops from 0.959 in *Consistent-1* to 0.945 in the combined Consistent set, while the AR50-95 increases from 0.826 to 0.865. This suggests that even within high-quality images, the model's evaluation will vary significantly based on which evaluation metric is used, and on which test-set. The increase of AP and simultaneus decrease in AR for the same confidence thresholds means that the differences stems from the variation of the dataset partitions.

*Consistent-1* has higher precisions and lower recalls than *Consistent-2*. This means that the persons in *Consistent-1* are more often missed. This may be due to the additional occlusions that are introduced with the increased number of persons compared to *Consistent-2*, making it harder to detect all persons in the image.

Generally, the models perform very well, achieving scores far beyond the scores typically achieved on bigger, generic and possibly harder to infer on datasets. This shows our choice in machine learning architecture for the given environment and scale is has been successful[38].

Finally, this evaluation of the various test sets serves as a demonstration of the accuracies of YOLOv9 for the specific environment, which surpasses the accuracies of the same model on the COCO dataset. This is further demonstrated and underscored in the following results.

[38]The high scores could also indicate something has gone wrong in our evaluation pipeline. This is considered unlikely, however, after having looked at images from the inferences and verified a lot of the inferences are indeed correct.

### 4.1.3 Input Image Size

We experimented with the input image size for model inference with the standard YOLOv9 to find the optimal value. 320, 640 and 1280 were tested. The input image size affects accuracy and inference latency of the models. The inferences were performed using an 8-core AMD® Ryzen 7 4700u CPU with 16GiB RAM[39], resulting in higher inference latencies than what are usually reported for similar models. In addition to the hardware, other simultaneuos tasks on the computer may affect inference latency. Therefore, no other user input was given during these model inference runs. However, background processes and other programs were not terminated during the process, so the numbers are only not valid outside this thesis. More valid results could have been achieved by running the inference many times and taken the average inference latency.

The weights file also affects the inference latency. The YOLOv9 is released with three sets of pre-trained weights. These are called yolov9-m, -c, and -e (as of 03.06.2024). The results of Table 5 were achieved using the yolov9-*e* weights.

| Model | Input Image Size | mAP50-95 | Inference Latency |
|---|---|---|---|
| YOLOv9 | 320 | 0.820 | 302ms |
| YOLOv9 | 640 | 0.945 | 840ms |
| YOLOv9 | 1280 | 0.877 | 3336ms |

Table 5: Comparison of Yolov9 Models With Various Input Image Sizes Evaluated on *Consistent*

Table 5 reveals the effects of input image size on model performance. Performing model inference with an input image size of 1280 was hypothesized to augment the accuracies, but this was not the case in our experiment. Introducing a much higher inference latency and poorer performance than of the 640 image size, the 1280 is clearly the worst option. This is likely due to the scale of the persons in the FIMUS Consistent dataset partition, which does not necessitate a higher input image size than 640. Inferring with size 320 may be a viable option in applications where speed is important.

### 4.1.4 Model Average Precisions on Consistent Dataset

The models were evaluated using both COCO AP *and* Vary-Both AP metrics, increasing reproducibility as future developments and evaluations may freely choose between the two methods. The rankings of the models are similar for both evaluation metrics, demonstrating that the choice of evaluation metric is not detrimental to the process of finding the optimal model for a deployment scenario. The results are discussed in depth in following paragraphs.

**COCO AP** The AP and the AR of the models varied with number of epochs they were fine-tuned for. The scores are displayed in Figure 34.

[39]The inference machine hardware specification is mentioned here, and not in the methodology section, as this is the only place in the thesis where speed is discussed and the inference machine hardware is of relevance. The machine's hardware does not affect inference accuracy.

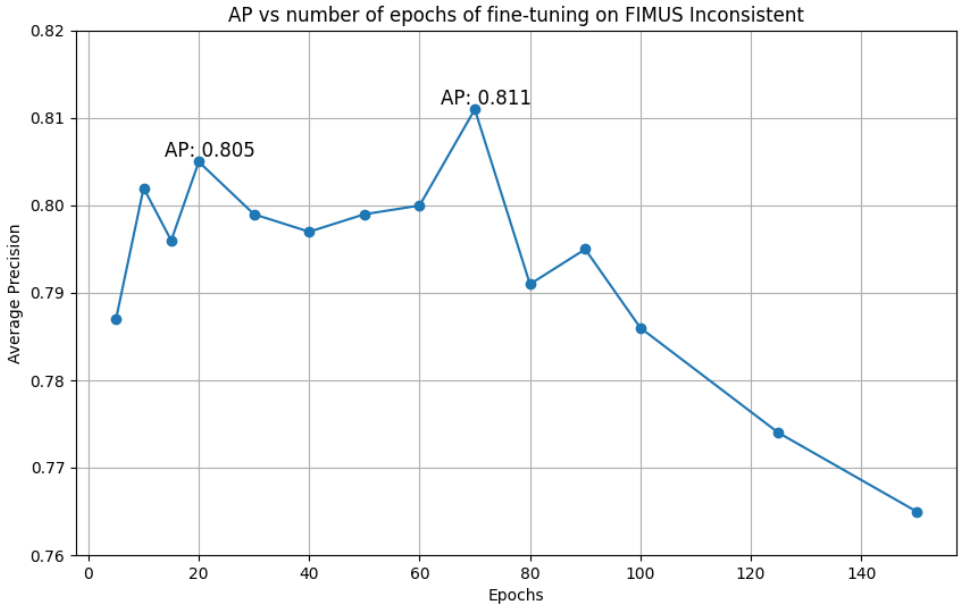



Figure 34: Average Precisions Over Number of Epochs Fine-Tuned on Inconsistent

The model achieving the best average precision was fine-tuned for 70 epochs. Notably, we observed a local maximum at 20 epochs. Had we not continued training past 60 epochs, we would not have known the model could improve at 70 epochs. The average precision scores were also calculated and are displayed in the graph of Figure 35.

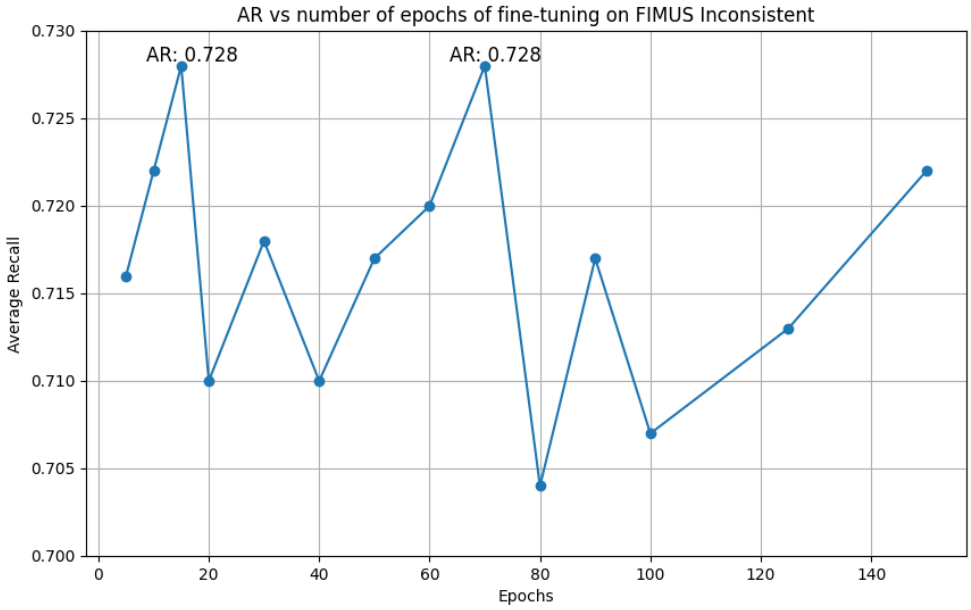



Figure 35: Average Recalls Over Number of Epochs Fine-Tuned on Inconsistent

We observed that the best models regarding average recall were achieved after 15 and 70 epochs. Considering both AP and AR for the models, it is clear that fine-tuning the model for 70 epochs produced the highest-scoring fine-tuned model. A comparison of this model and all other models

on the FIMUS Consistent dataset partition is displayed in Table 6.

| Model | AP50 | AP75 | AP50-95 | AR50-95 |
|---|---|---|---|---|
| YOLOv9-m Standard | 0.957 | 0.908 | 0.865 | 0.716 |
| YOLOv9-c Standard | 0.932 | 0.885 | 0.845 | 0.747 |
| YOLOv9-e Standard | 0.972 | 0.949 | **0.945** | **0.865** |
| YOLOv9 CrowdHuman 5ep | 0.968 | 0.916 | 0.834 | 0.692 |
| YOLOv9 CrowdHuman 10ep | 0.964 | 0.914 | 0.831 | 0.702 |
| YOLOv9 Inconsistent 70ep | 0.954 | 0.867 | 0.811 | 0.728 |
| YOLOv9 PRW 5ep | 0.802 | 0.681 | 0.605 | 0.389 |
| YOLOv9 PRW 10ep | 0.969 | 0.856 | 0.721 | 0.455 |
| YOLOv3 | 0.804 | 0.742 | 0.681 | 0.734 |
| DETR-50 Conf 0.50 | 0.404 | 0.347 | 0.307 | 0.675 |
| DETR-50 Conf 0.99 | 0.776 | 0.711 | 0.624 | 0.549 |
| DETR-101 Conf 0.50 | 0.606 | 0.535 | 0.468 | 0.689 |
| DETR-101 Conf 0.99 | 0.818 | 0.762 | 0.665 | 0.559 |

Table 6: Models APs on *Consistent* (757 images)

**Results Discussion** These results show no accuracy improvements from fine-tuning. The standard models performed best on the Consistent dataset. The results indicate that freezing the backbone and fine-tuning the head on inconsistent, highly relevant data was a destructive practice. A notable discovery was the already high scores the YOLOv9 achieves on the *Consistent* (and the rest of the FIMUS dataset). C.-Y. Wang and Liao report a score of 0.556 on the COCO dataset, which is substantially lower than the scores we achieved on FIMIST Consistent. This may seem odd, as the model weights are pre-trained on COCO data, which should be more similar to the test data for the same dataset. However, as discussed in the section introducing the COCO dataset (see 3.3.1), it is a very difficult, highly diverse dataset. In our Consistent dataset, if a model is able to correctly infer one image, it is likely to also infer correctly on the rest of the images because they are very similar in nature.

Out of the fine-tuned models, the YOLOv9 model trained on the CrowdHuman dataset achieved some of the highest scores. This is not unexpected: The dataset contains a lot of diversity and probably more instances of humans to learn from than the Inconsistent FIMUS dataset. The models that were fine-tuned on Inconsistent were trained for a tenfold more epochs to try to make up for this fact, but the accuracies failed to improve. These results indicate that rather than producing a sub-optimal specialized dataset for fine-tuning a model, the better option may be to use a larger and better dataset.

Should the positioning of the bounding boxes be of less importance, we see that fine-tuning on the PRW dataset is still a better option than fine-tuning on our FIMUS Inconsistent dataset partition. This is revealed by the AP50 score of 0.969 after just 10 epochs, while training on FIMUS Inconsistent did not achieve similar results in its 150 epochs of training.

The attentive reader may have noticed the Football Players Detection Fine-Tuned models are not included in the results presented in Table 6. This is because both the 5 epochs and 10 epochs model both completely missed the persons in the aquarium. This could have been due to the difference in scale, resulting in a fine-tuned model not able to predict the humans in the aquarium setting. A review of the labels indicate that the model had close to no inferences with a confidence score higher than 0.1, and the highest confidence at 0.425. An attempt was made to improve the scores by detecting at confidence threshold 0.001, then normalizing the detections, but the poor model

performances were not salvageable. An example from the labeled images is displayed in figure 36. Interestingly, it does not infer the roof lamp to be a person, like we saw with the standard YOLOv9.

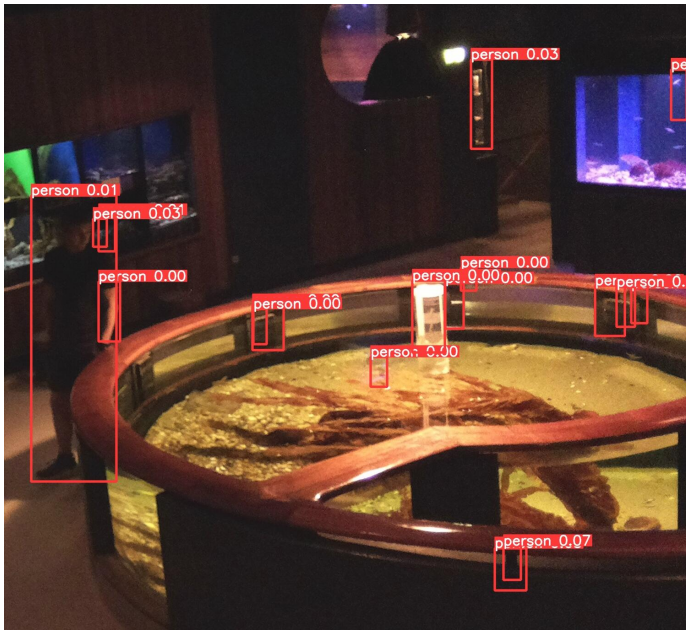



Figure 36: Example Image of the Unsalvageable Football Players Model Inferences

The reasons for this model failure was due to a mistake made when freezing the backbone. Only the 10 first layers were frozen, instead of all 28, resulting in a model that has suffered from the catastrophic forgetting problem which, as mentioned, is especially prevalent in tiny ML applications. The model would then have to be retrained, for which there were no resources left to do in this thesis project.

The fixed confidence threshold was raised for the DETR models since they report vastly higher confidence scores than many other model architectures. Therefore, the confidence scores were fixed at 0.99. A comparison of the inference latencies of the models were not conducted, as the YOLO model is still clearly the better option. We see, however, that this raise in confidence threshold naturally also results in a reduced AR score for the models. For the calculations of Vary Both AP, the confidence threshold was varied similar to every other model.

Another surprising result from Table 6 is the YOLOv9-m performance relative to the larger weights of c and e. These letters correspond to different pre-trained weights, with various number of parameters. The converted version[40] of YOLOv9-e weight file has a size of 117.2MB, YOLOv9-c and -e fills only respectively 51.4MB and 40.7MB of space.

**Vary-Both AP** Figure 37 displays how the AP of the models varied with the number of epochs. Equally to the COCO AP, the best model was the standard YOLOv9-e. We recall the Vary-Both AP references to the AP score achieved when evaluating models with a varying IoU threshold of 10 values from 0.50-0.95 *and* confidence threshold from 0.10 to 0.90 in 20 steps.

[40]There's also converted vs not converted models. The reparameterization functionality to convert a model consists of trimming layers meant to speed up and augment model training, which is not needed for running model inference. The converted models achieve the same results but with lower inference latency, smaller size. They should not, however, be used for fine-tuning.

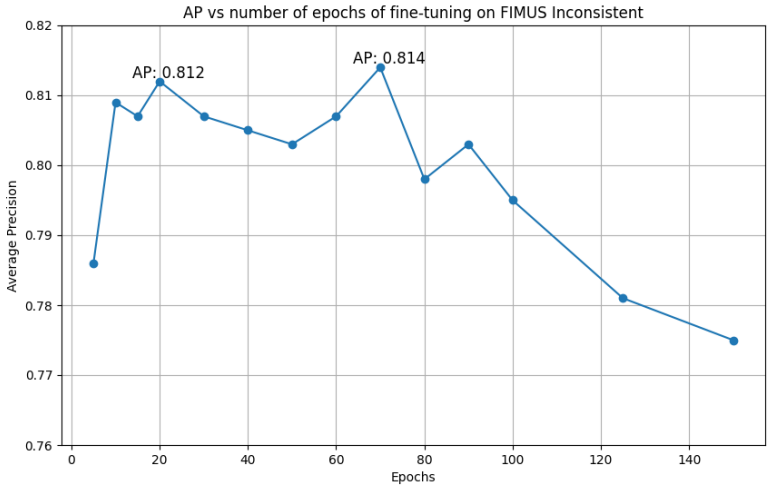



Figure 37: Vary-Both APs Over Number of Epochs Fine-Tuned on Inconsistent

The following table (7) displays how well the models performed on the Consistent dataset. Similar to the results in Table 6, the YOLOv9-e was the most accurate, achieving an AP50-95 of 0.913.

| **Model** | **AP50** | **AP75** | **AP90** | **AP50-95** |
|---|---|---|---|---|
| YOLOv9-m Standard | 0.895 | 0.855 | 0.746 | 0.821 |
| YOLOv9-c Standard | 0.881 | 0.842 | 0.740 | 0.810 |
| YOLOv9-e Standard | 0.942 | 0.918 | 0.876 | **0.913** |
| YOLOv9 on Inconsistent for 20ep | 0.948 | 0.854 | 0.680 | 0.812 |
| YOLOv9 on Inconsistent for 70ep | 0.948 | 0.867 | 0.674 | 0.814 |
| YOLOv9 on PRW for 5ep | 0.808 | 0.706 | 0.427 | 0.633 |
| YOLOv9 on PRW for 10ep | 0.924 | 0.792 | 0.252 | 0.677 |
| YOLOv9 on CrowdHuman for 5ep | 0.908 | 0.863 | 0.688 | 0.788 |
| YOLOv9 on CrowdHuman for 10ep | **0.944** | 0.902 | 0.740 | 0.822 |
| YOLOv9 on Football for 5ep | 0.221 | 0.177 | 0.175 | 0.179 |
| YOLOv9 on Football for 10ep | 0.514 | 0.457 | 0.443 | 0.452 |
| YOLOv3 | 0.780 | 0.728 | 0.528 | 0.672 |
| DETR ResNet50 | 0.414 | 0.355 | 0.164 | 0.315 |
| DETR ResNet101 | 0.594 | 0.526 | 0.245 | 0.460 |

Table 7: Comparison of Detectors APs on Consistent (757 images), Multiple IoU and Confidence Thresholds

Table 7 includes AP90 instead of AR50-95 to highlight a separate discovery from what is discussed around Table 6. We see the standard YOLOv9-e is still the highest scoring model, even when calculating the AP with varying IoU and confidence thresholds as opposed to calculating COCO AP where the confidence is fixed. However, we also notice the YOLOv9 fine-tuned on the CrowdHuman dataset for 10 epochs is better performing than the standard model when the accuracies of the bounding boxes are less strict, i.e. the IoU value is lower. This indicate that our training data may have been imprecise in bounding box placements. Human errors are not uncommon. The

YOLOv9-e is best performing not only due to the high precision for IoU threshold at 0.50, but the maintained precision as the IoU threshold is tightened up to the final value of 0.95. This suggests the training data for fine-tuning the models may have been suboptimal in its bouding box placement.

Another difference from the COCO AP evaluation is that here, the DETR models were also varied from 0.1 to 0.90 in confidence thresholds, resulting in much worse average precisions than when the confidence threshold for these models were fixed at 0.99.

**DETR Fixed Confidence Thresholds' Effects on Average Precision** The poor performances of the DETR models motivated the an investigation of the labels. The DETR models infer with much higher confidences, making the fixed threshold at 0.5 way too low to score well on the COCO metric. An assessment was made thus made to find the optimal confidence level for the model. This was found at 0.991, nearly doubling the AP score from 0.315 to 0.625 for the DETR with a ResNet50 backbone, and improving the score from 0.460 to 0.668 for the model with the more complex ResNet101 backbone.

| Model | AP50 | AP75 | AP90 | mAP50-95 |
|---|---|---|---|---|
| DETR ResNet50 Conf 0.50 | 0.414 | 0.355 | 0.164 | 0.315 |
| DETR ResNet50 Conf 0.95 | 0.755 | 0.660 | 0.326 | 0.585 |
| DETR ResNet50 Conf 0.99 | 0.776 | 0.711 | 0.369 | 0.624 |
| DETR ResNet50 Conf 0.991 | 0.773 | 0.713 | 0.373 | 0.625 |
| DETR ResNet101 Conf 0.50 | 0.594 | 0.526 | 0.245 | 0.460 |
| DETR ResNet101 Conf 0.95 | 0.795 | 0.719 | 0.353 | 0.627 |
| DETR ResNet101 Conf 0.99 | 0.818 | 0.762 | 0.408 | 0.665 |
| DETR ResNet101 Conf 0.991 | 0.818 | 0.767 | 0.419 | 0.668 |

Table 8: APs for DETR When Fixing the Confidence Threshold at Various Values (757 images).

The results presented in Table 8 are in accordance with the results of Carion et al. They claim it is performing well on panoptic segmentation, a task where pixel-level detail is important. This is the reason to the confidence values are high for the task of object detection. For the experiments in Table 6, a threshold of 0.99 was used instead of 0.991, because experimentally finding the optimal confidence threshold post-inference on a test-set is likely overfitted to the testing data and not so easy to optimize for in a practical setting.

The results of this section are put into a broader context in the Discussion section, see 5.8.

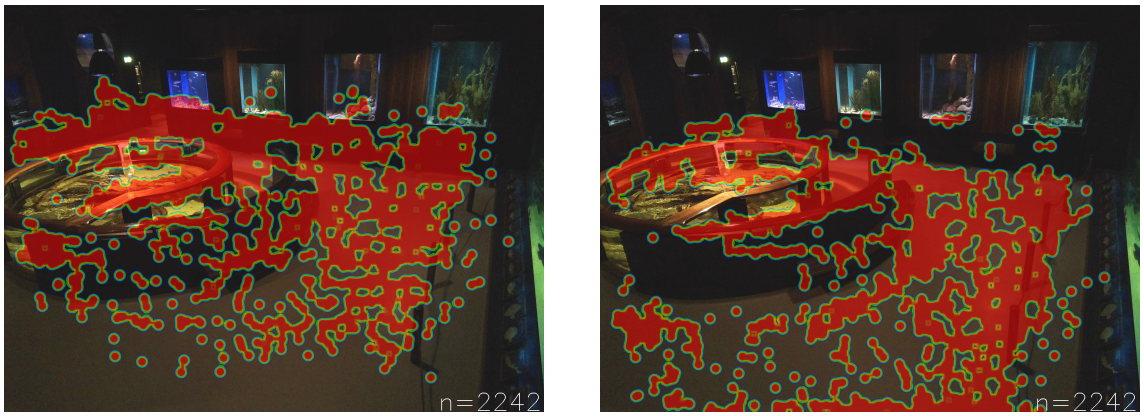
## 4.2 Data Visualization

The obtained localization data of visitors may be visualized a multiple of ways. The explored methods in this thesis are by creating heat maps and bar charts to visualize the data.

### 4.2.1 Visitor Localization Heatmaps

As mentioned in Section 3.8, heatmaps are a powerful visualization tool that can provide insights into visitor behavior patterns and engagement levels within a museum or aquarium setting. Heatmaps for the month of may are illustrated in 38. These heatmap generation code for the two heatmaps are identical, apart from one variable: the position where detections are mapped to. In 38a and b, the detections are mapped to the respectively the middle and the bottom center of

the detection bounding box. This single modification has the largest difference on the edges of occlusions, such as (for the images in Figure 38) the railing of the fish tank in the center.

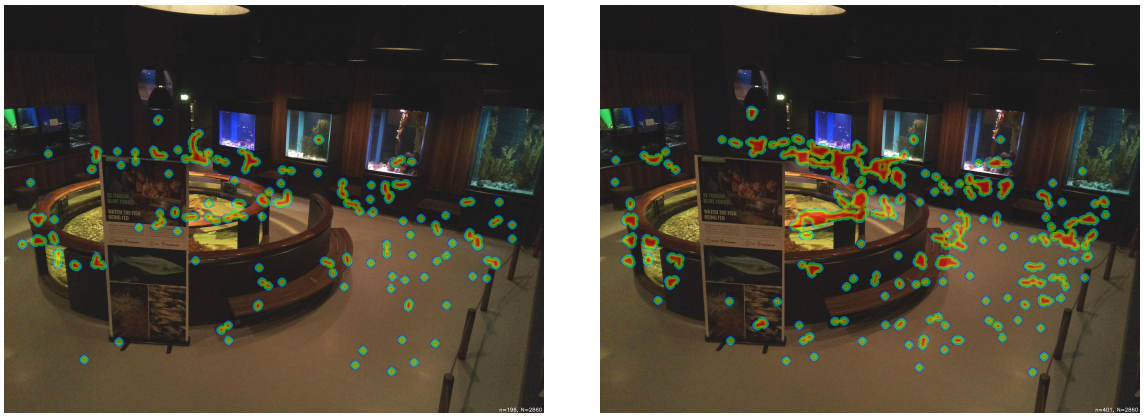



(a) Sampled From the *Center* of Detections (b) Sampled From the *Bottom* of Detections

Figure 38: Weekly Heatmaps: Positions Sampled From Center vs Bottom of Each Detection

There's another, more important takeaway from the small modification. While seemingly similar, the heatmap sampling detections from the middle of the bounding boxes (38a) reveal a weakness in our detector which is invisible in the other heatmap: the lamp is sometimes classified as a person. On the other hand, the other heatmap (38b) reveal another weakness. The seaweed in the second fish tank from the right is sometimes also classified as a person.

Apart from revealing weaknesses from the detector models[41], these heatmaps may provide valuable insights with regards to which areas of the facility are being used the most. There may be difficulties, however, in correctly inferring what are the reasons for the variations. For periods less than a day, these variations are likely due to randomness. The more interesting numbers in this context would be to see the total number of visitors throughout the day, which is better visualized in the bar charts in Section 4.2.2. Two heatmaps for separate days are illustrated in Figure 39.



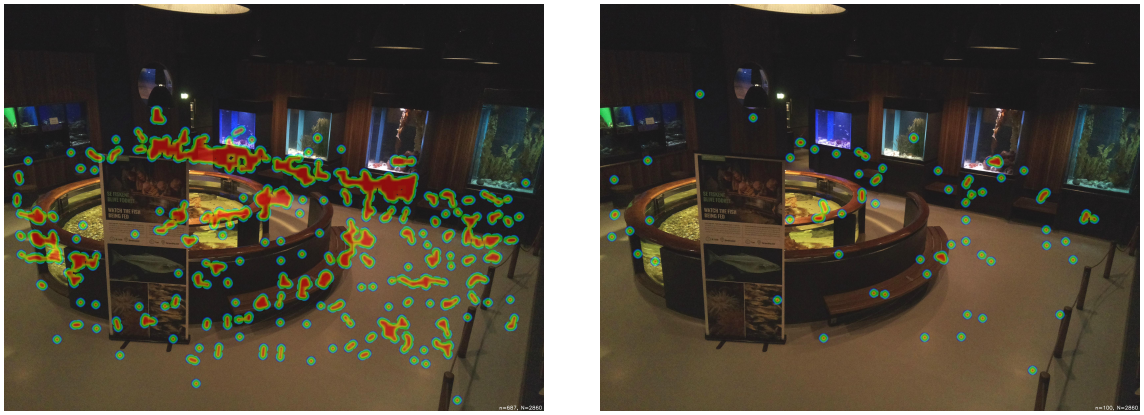
(a) Wednesday 8th of May, 2024 (b) Saturday 11th of May, 2024

Figure 39: Daily Heatmaps

Heat maps may also visualize the hours throughout a day, accumulating all data for a specific time period each day to see if the visitor engagement changes based on the time. This is one example of introducing a variable, namely the time of day, to filter the detections. For an area

[41]These weaknesses in our detector models could be revealed by looking at the annotated images. However, looking at the annotated images is not possible for a on-device processing image-deleting device. In this case, one would need to display/plot the detections onto a base-layer image (heat map), or make use of obfuscation discussed in Section 2.5.7 to illustrate and reveal model weaknesses.

where other variables such as the temperature, the noise level or the weather is also known, this could be used instead to filter the detections and illustrate how visitor engagement changes based on these factors. This usage would naturally, require some months-worth of data to be valid. For this project, only a months-worth of positional data has been stored to make the analysis. An illustration of heatmaps where the time of day has been used to determine which detections are presented in the heatmaps are displayed in Figure 40.

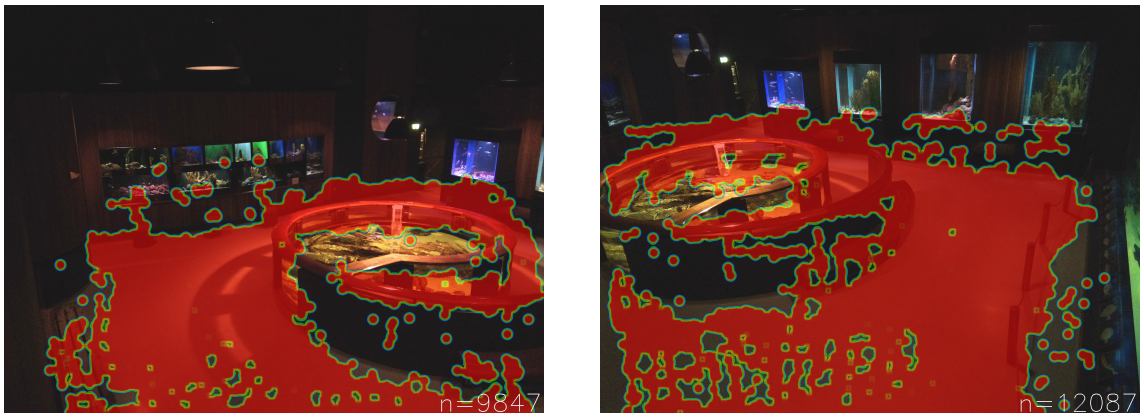


(a) 13:00-14:00 (b) 16:00-17:00

Figure 40: Hourly Heatmap

The heatmaps in Figure 40 reveal another use case for heatmaps. The relative difference between the two heatmaps is likely due to randomness, but with a larger number of detections one might be able to look for patterns. This could be that the heatmap for 13:00-14:00 could show a higher number of detections in front of the fish tanks, while the heatmap for 16:00-17:00 could show a higher number of detections on the benches. This could have easily been overlooked, had a manager of the museum only passed through the museum in the day and never in the evenings, resulting in him not thinking so many benches were neccessary.

The project experiment with the devices in the aquarium lasted for the entire month of May, and heatsmaps from all of the gathered data are displayed in Figure 41.




(a) Left, n=9847 (b) Right, n=12087

Figure 41: Monthly Heatmap

From the heatmaps in Figure 41, we can clearly understand the need to filter the data in one of the proposed ways to gather more insightful information.

### 4.2.2 Visitation Count Bar Charts

Another tool is to analyze the average number of detected persons per hour. This provides insights into room utilization during different times of the day. Figure 42 provides an easy to grasp representation of what the peak hours of the aquarium has been. The figure displays that most visitors are in the aquarium from 13:00-13:59. This is not unlikely, as it is the time of the day the aquarium staff feeds the fish. It is also possible to see that the first hour after opening is quite busy, while the last hour from 16:00-16:59 is the hour with the least amount of visitors. The numbers in Figure 42 are subject to the previously described uncertainty of the model in inferring person localizations.

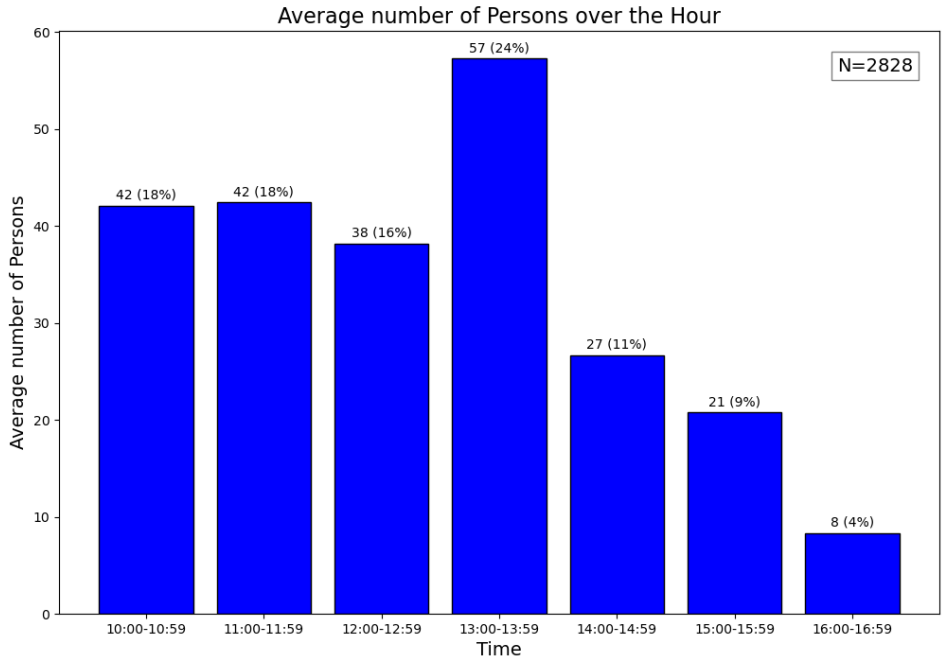



Figure 42: Peak Hours Analysis

Some insights may be extrapolated from this data. For instance, a lower number of detections during early opening hours, despite high visitor entry, could indicate that the room temperature is not yet optimal, affecting visitor comfort. This hypothesis could be tested by using the number of visitors as the dependent variable and adjust the room temperature to see the effects. In the absence of other confounding variables[42], this could be an interesting causal relationsship to investigate to infer the perfect room temperature. However, due to the requirement in such an investigation for the high volume of data to rule out the possibility of randomness confounding the results, this investigation is likely unrealistic.

Further, comparing visitor detections of summer vs winter months, normalized for the total visitors in the facility, could provide deeper insights. It would enable the possibility of gauging the relative popularity of different areas. Indoor environments, typically maintained at constant temperatures, might offer different levels of comfort compared to the naturally fluctuating conditions of outdoor areas. More data visualization future work potential avenues are discussed in the Future Works section.

Understanding these dynamics can guide decisions on environmental controls, such as adjusting

[42]Removing all other confounding variables from a real world setting may be a very difficult task. To do so, one must collect data over a long time period in order to average out the random events and factors that may influence the experiment.

heating levels to enhance visitor comfort and potentially increase engagement in specific areas of the facility. Such adjustments could directly influence the overall visitation experience, making the whole facility more favorable for a visit regardless of seasonal factors.

A hypothesis was formed during the image capture process by observing the few visitors during the final opening hours of the aquarium. Most late visitors would sit more on the bench, rather than walk and run around the room like many of the earlier visitors would. This is a find that the heat map could have revealed, had it been the case. However, the data-driven visualization (the heatmaps), tell us a different story.

### 4.2.3 Data Visualization Future Directions

Through the development of more sophisticated models, possible future use cases for the technology could be to develop a model to segment the users based on their age group. This could provide graphs to infer e.g. that children are most represented early in the day, and this data could thus be used to have children activities when most of the user group are present to hit the biggest audience. One could also use the data to have children-activities in the afternoon to draw more young visitors to the unpopular visitation hours. The age groups would most succintly expressed as different colored bars in a barchart.

# 5 Reflections & Overall Discussions

This section explores the broader implications of the findings, summarizing both the results and their significance in the development of comparable systems. It evaluates the feasibility of the approaches presented in this thesis as viable solutions to the addressed problems. For coherence, the research questions introduced at the beginning are revisited and answered conclusively in this section.

## 5.1 Privacy in Images

The developer has complete control over post-analysis image handling in systems with on-device processing. The most privacy-preserving method is to delete the image after analysis and transmit only the anonymized results in form of positional coordinates of the persons. However, image obfuscation may sometimes be preferable. While deleting the images is advantageous when the technology is proven or some level of analysis error is tolerable, the retaining obfuscated images facilitates easier technology development and verification, as verification of analysis results is possible. Additionally, obscuring images can provide transparency about the origin of analysis results, which may be beneficial in situations where end-users seek to verify and understand the data. Techniques for image obfuscation, such as blurring, were discussed, noting that blurring the entire image is a straightforward method to face blurring to avoid relying on facial recognition to work perfectly.

Edgcomb and Vahids findings suggests blurred images are not be considered privacy preservant (2012). However, one could argue the results of this study are invalid for multiple reasons:

- Public perceptions of privacy have evolved since 2012, potentially diminishing the study's relevance.
- Internal validation: The design of the questionnaire may have biased participants to perceive blurred videos as more invasive than they might actually be. This could be because other, more privacy-preserving methods were presented alongside blurred videos, leading participants to comparatively view blurring as the least effective option. Had blurred videos been evaluated in isolation, they might have been judged as adequately privacy-preserving.
- External validation: The demographic profile of the participants might constrain the studys relevance to regions like Scandinavia, where a generally higher trust in public institutions might lead to greater acceptance of obfuscation methods. Exploring how demographic variations influence perceptions of privacy protection methods, however, falls outside the scope of this thesis.

Having discussed the nuances of privacy preservation in image processing, let us now turn our attention to the implications of using third-party services and products, which present a different set of challenges and considerations.

## 5.2 Third-Party Services & Products

This subsection delves into the pros and cons of integrating third-party services and products into object detection systems, highlighting the trade-offs between ease of development and control over data and functionality. As discussed in Section 2.8, third-party services offer significant advantages over creating a solution from the bottom. However, these advantages must be balanced against several potential drawbacks.

#### Potential Drawbacks of Utilizing Third-Party Services or Products

1. **Control Over the System:** Developing an in-house application provides unparalleled customization opportunities, from software architecture to data processing and system integration. Such autonomy allows for system optimizations specific to performance and operational needs, and independence from the continuance and performance of external services.
2. **Data Privacy and Security:** On-device processing ensures data remains confined to the device, thereby enhancing security and privacy. Although some providers, such as Roboflow, offer local deployment, this often requires a costly business-level subscription.
3. **Cost Efficiency:** While third-party services may reduce upfront development costs, they can incur ongoing charges like subscription fees, usage rates, and costs for premium features or enhanced support. For large or long-term projects, these expenses can be substantial.
4. **Performance Optimization:** By owning the entire system pipeline, one can tailor both hardware and software for optimal performance, achieving faster processing speeds and reduced latency that third-party services may not match.
5. **Scalability and Integration:** Custom solutions facilitate easier scaling and integration with existing IT infrastructure, aiding in the seamless expansion of data workflows and supporting business growth without the constraints of external platforms.

In conclusion, while third-party services can accelerate development in the field of object detection, it is crucial for researchers and practitioners to meticulously consider these aspects, particularly in sensitive applications like person detection where privacy is paramount. Exploring both proprietary and third-party solutions will provide a balanced perspective on flexibility, control, and innovation potential.

While third-party services offer certain benefits and drawbacks, the application of person localization systems, especially in environments like museums, presents unique opportunities that affect various stakeholders differently.

## 5.3 Applicability of Person Positioning Systems

Here, we explore how person localization systems can be applied in cultural settings, examining the differing priorities and potential benefits for various stakeholders such as curators and administrators.

Curators may prioritize the enhancement of visitor engagement and educational experiences. Person positioning systems can provide valuable data on visitor traffic patterns, dwell times, and interest areas, enabling curators to optimize exhibit layouts and tailor informational content to visitor behavior. This data-driven approach can significantly enhance the educational impact of exhibits and improve overall visitor satisfaction.

On the other hand, administrators might focus more on operational efficiencies and security enhancements. The insights gained from person localization technologies can streamline staffing needs, enhance security monitoring, and manage crowd control more effectively. By understanding peak visitation times and the flow of visitor movement, administrators can allocate resources more efficiently, potentially reducing operational costs and improving the safety and comfort of museum environments.

Beyond the practical applications in cultural institutions, the development of person localization systems also raises profound ethical questions, particularly concerning public surveillance and data privacy.

## 5.4 On the Ethicality of Person Positioning Systems Development

This subsection discusses the ethical considerations and societal perceptions related to the deployment of person localization systems, particularly in the context of public surveillance and

privacy. History has shown Kant's categorical imperative to function as a guiding principle in smaller groups, but it often falters in larger societies where in- and out groups are forming. This is apparent, since wars and failure to provide basic humanitary aid to those in need is still an issue in global society. Enabling further mass public control through automated person localization devices raises significant ethical concerns.

**Public Surveillance Perceptions** People of the general public often argue against mass surveillance, stating it is inherently wrong without being able to fully articulate the consequences. As seen in Section 2.1.3, these objections often give way when convenience outweighs privacy concerns. This constitutes a need for a stronger motivation to uphold the individual privacy, once technology capabilities surpass these desires. Most agree to surveillance technologies in public spaces for security reasons, but the mass general public has also accepted having devices in our homes and pockets listening for a "Hi Siri", "Okay Google", or "Alexa", which displays a gradual increase in acceptance of devices which are potentially chipping away at our privacy. The potential for more intrusive technologies that track individual movements, gestures and activities within private spaces poses even greater risks to personal privacy.

**The Deontological Perspective and Its Limitations** The statement by Redmon, "[...] I bought in to the myth that science is apolitical and research is objectively moral and good no matter what the subject is", reflects the limitations of a purely deontological approach[43] to ethics in technology. This perspective often overlooks the broader impacts of scientific advancements.

**Balancing Benefits and Risks: A Utilitarian Approach** Conversely, utilitarianism suggests that we must weigh the potential benefits of e.g. increased safety, efficiency, and convenience against the risks to privacy, individual freedoms, and other negative impacts. The key challenge for developers and policymakers is to maximize the net positive impact—enhancing societal well-being while minimizing adverse outcomes. This perspective underscores the importance of impact assessments and ethical considerations in the development and implementation phases of these technologies.

**Modern Philosophical Contributions to AI Ethics** Contemporary philosophical discourse, as highlighted by thinkers like Harris, Russell, and Yudkowsky, stresses the importance of effective governance and informed regulation of AI technologies. These discussions are particularly relevant in the context of developing regulatory frameworks that can manage the ethical challenges posed by advanced technologies without stifiling innovation.

**Conclusions on the Ethicality Person Positioning Systems Development** In conclusion, the ethical deployment of computer vision technologies requires a multi-faceted approach. A deontological stance would be to not accept visual person localization technologies, but such positions often struggle to counteract the compelling convenience these technologies offer. Instead, a more robust ethical framework must be adapted by society. The utilitarian principles, which call for maximizing societal benefits while minimizing harms, and modern philosophical discourse on AI regulation highlight the necessity for informed policymaking. Since we can not expect every person to scrutinize the potential consequences of every new IoT device, we need governmental control to decide which technologies can make it into our homes. This control through regulations also allow for the heuristic of blindly trusting devices on the market, easing the decision-making process and reducing public resources spent individually to investigate products for their privacy characteristics. This is why products need CE marking exists, to ensure consumers the product conforms with European health, safety, and environmental standards.

The necessity for regulation becomes particularly evident when considering societal behaviors that undermine collective well-being; continued littering, even though we have proof of trash destroy-

[43]Deontological approach, meaning the morality of an action should be based on whether the action itself is right or wrong.

ing ocean life, and people choosing to not recycle their metals, even though we have scientific proof of Earth warming at a higher rate than ever previously measured[44]. These continued behaviours advocate for stronger governmental control, capable of rewarding well behaved citizens or sanctioning the ill behaved. This may motivate the use and implementation of public control technologies which may be capable of identifying undesired behaviour. The person localization systems discussed in this thesis is meant for visitor behaviour analysis in cultural insitutions, but could be extended to such applications. Redmon's development of the YOLO algorithm which was later used as a military weapon exemplifies the actual use of technological advancements can diverge significantly from their intended purposes. Regulations are necessary in a well-functioning society for controlling citizen behaviour, and they are necessary in a well-functioning society for controlling the crowd-control capabilities of future data science and computer vision technologies may facilitate.

Having examined the ethical landscape surrounding person localization technologies, we now shift focus to the practical aspects of system implementation, specifically the selection of appropriate hardware and software.

## 5.5 Hardware and Software Choices

In this subsection, we review the hardware and software choices made for the project, assessing their effectiveness and the lessons learned that could inform future projects. The selection of hardware and software for this project was made based on availability, cost, and compatibility with the project goals. While some choices might not have been optimal in hindsight, they were adequate for the project's scope. For instance, the chosen hardware handled the low-light conditions of the aquarium setting satisfactorily, and the software solutions provided the necessary functionality without excessive complexity. However, it is always valuable to reflect on these choices to identify potential improvements for future projects. It is likely that the advancements in mobile phone technology eventually will translate into more efficient low-level systems and better suited hardware for visual mobile systems.

There are many options for sensors to detect the presence and positions of persons in a room (IPS). The option to use a visual sensor was due to the opportunity to extend the application to perform more sophiticated tasks which are inaccessible to other sensors. These applications could include fall detection, approximation of age, emotion detection, and pose estimation. These improvements could be crafted by improving the software, not needing to reconstruct hardware or change the already deployed cameras. The extension of the person localization system to detect falls in public spaces where individuals sometimes travel alone could have the potential to save lives. The option to focus on visual sensors and the applicability in terms of ethicality and practical dilemmas was motivated by this extendability to meaningful and helpful devices that may save lives in the future[45].

**Programming Language** The chosen programming language of Python is typical for data science projects because of the broadly adopted libraries for computer vision, data management and machine learning. This motivated the option to write the code for the project in Python. Using a programming language closer to machine code, i.e. C, C++, or Rust, would prove beneficial to optimize device efficiency and speed and is a consideration that could have been taken more seriously if the goal of the application was real-time object detection or similar. Micropython could've been an option, but was foregone as there was no need for it since the Raspberry PI 3 is more than powerful enough to run regular Python.

After addressing the technical decisions regarding hardware and software, it's crucial to reflect on

[44]The dean of Information Technology and Electrical Engineering at NTNU mistakenly claimed Earth's temperature was higher than ever before, at a speech given to the graduating students of 2024. This is a common misconception, even among the highly educated population.

[45]A clear application of these devices once the ethical principles and on-device processing procedures of deleting the image post-analysis would be to have them detect if an elderly person living at home has fallen, has taken his/her medications, or if the person has stayed in bed for longer than a pre-defined maximum period.

the results of our system's implementation and how it meets the project's research questions and objectives.

## 5.6 Results

This subsection presents a detailed discussion of the performance metrics used to evaluate the object detection models, highlighting the implications of these findings on model selection and system effectiveness.

The evaluation metrics employed, COCO AP and Vary Both AP, yielded similar rankings for the tested models, suggesting that the choice of metric is not a crucial factor in model selection for this project. However, COCO AP is particularly beneficial as it mirrors the practical demands of object detection systems, where confidence thresholds are fixed to optimize real-world performance. This feature makes COCO AP a more representative metric for gauging how models might perform post-deployment.

**Impact of Fine-Tuning on Performance** The fine-tuning process, intended to refine model performance by adjusting model parameters to better fit specific datasets, paradoxically led to poorer outcomes. This degradation in performance could stem from several factors:

- The evaluation setup for the model accuracies was not tested sufficiently to completely outrule biased results. This might have led to results that does not reliably reflect the impact of dataset quality and relevance.
- Potential missteps in the training process, such as inappropriate training protocols or sub-optimal configurations for single-class inference, could have skewed the results.
- The architectural decisions regarding the number of anchors and layers in the neural networks may not have been optimized for the scenarios tested, resulting in pre-trained model weights being more optimized to the scenario. Adjusting these factors could potentially enhance model performance, suggesting that future studies should explore more nuanced modifications beyond simple fine-tuning without regards to other factors.
- The training process contained too few instances of persons. Either the datasets for fine-tuning would need to contain a higher number of relevant instances for the model to learn the necessary features, or it would need to train for many more epochs. As previously stated in the Results section, we already identified we hit a local maximum of model accuracy when it was fine-tuned on FIMUS *Inconsistent* for 20 and 70 epochs. As a dataset with few instances per image and a relatively low number of images, the models that are fine-tuned on *Inconsistent* may necessitate more training epochs to converge on better model weights. This is still the case altought the model will already know basic features from pre-training the backbone.

Freezing the backbone was most likely not the cause to worsened performance. The models were trained on the COCO dataset, which contains a vast amount of images and 80 classes, including person. The high accuracy of the standard YOLOv9 witness there's no issues in an unoptimized model to infer correctly on the data.

Investigating the effects of model fine-tuning on varying degrees of dataset relevancy remains an interesting topic for the future. Leveraging neural network search techniques and comprehensive hyperparameter tuning could yield more definitive insights with regards to how well the most optimized model may perform, should the technology be used for an application where its performance is critical.

**Reevaluation of Metrics** Initial assessments included a range of performance metrics, such as average precision (AP), average recall (AR), and F1-scores. Despite the initial decision to prioritize

AP50-95 based on its prevalence in the literature as a critical performance indicator, this decision was revisited. Upon reflection, incorporating AR and F1 scores into the final analysis was deemed essential, especially since the balance between accurately detecting targets (precision) and minimizing missed detections (recall) is crucial in many person localization applications. The recalculated results, incorporating a balanced consideration of both precision and recall, facilitated a more nuanced discussion about the performance of the models across different test set configurations.

## 5.7 Research Questions

The research questions have been answered throughout the sections of this master thesis. However, finding the answers to one specific question of interest in a 70+ pages thesis may be difficult. To alleviate this, concrete, short answers are provided in the list below to quickly summarize the findings of this thesis.

1. What are some privacy risks associated with traditional person localization systems in public spaces, and how may a system mitigate these privacy concerns?

   Privacy risks include unauthorized data collection, misuse of personal data, and lack of transparency in data handling. On-device processing mitigates these concerns by keeping data localized, enhancing security and privacy, and allowing users greater control over their data.

2. How does the validity of object detection model evaluations change when using data specifically from the intended deployment environment compared to using generic datasets?

   Accuracy scores were much higher in the thesis project than evaluations of the same models on generic datasets. This illustrates how scientific model evaluation consider how well a model performs given a wide array of diverse images with several challenges, as opposed to a specific use case where many of the challenges are similar for all the images.

   The validity of object detection model evaluations thus improves when using data from the intended deployment environment because the model is more precisely measured on it's ability to recognize specific features and variations present in the specific environment.

3. What are some machine learning architectures suitable for object detection in a real-world deployment scenario?

   Multiple machine learning architectures are suitable for object detection. Section 2.7 discussed multiple architectures, including those belonging to the category of *traditional* machine learning. We've seen how which architecture is optimal depends on the preferred capabilities, such as the ability to fine-tune, speed, and accuracy. The YOLO series make for robust, efficient and easy-to-implement models, and the Co-DETR is the current best-performant, but the DETR series is possibly premature, more complex, and may be more difficult to implement.

4. How do the performance metrics of object detection models compare when applied to different quality datasets?

   The consistent test sets (*Consistent-1*, *Consistent-2*, *Consistent*) represent higher quality images and show variability in performance metrics (see 4). This indicates that even within high-quality images, different partitions with separate characteristics can affect performance metrics differently. Further, this advocates for having as large and diverse test set as possible. It should cover as many of the possible encounterable post-deployment scenarios an object detector may face.

   The results also show different results for YOLOv3 and YOLOv9. For v3, the model performs best on the Inconsistent test-set, while this discrepancy is not as clear for the v9. For experiments aiming to find the best model, the test set is not selected in the way we are experimenting on the possibilities in this thesis project. Normally, the test set would be randomly selected from the available data. However, these results support the neccessity of a truly unbiased sampling of test data, as the test data itself may influence the choice of object detector model.

All the research objectives (see Section 1.6) have been successfully achieved throughout the thesis. The comprehensive Structure section (1.8) provides a roadmap, facilitating easy navigation to specific topics of interest within the document. Given this clear organization, a reiteration of the objectives and their fulfillment is deemed redundant and unnecessary. Readers seeking to verify the achievement of specific goals are encouraged to refer directly to the Structure section.

With the detailed results and responses to our research questions laid out, we now explore the broader implications of our findings, considering their impact on future technological developments and ethical considerations.

## 5.8 Broader Implications

Finally, we consider the wider societal and ethical implications of our findings, reflecting on how they contribute to ongoing discussions about technology, privacy, and the role of AI in public spaces. The practical demonstation of basic-level computer vision technology in the project of this thesis plays a tiny role in advancing technology. This aligns with the stated objectives of this work, which include demonstrating, investigating, and assessing how the quality and relevance of datasets influence the fine-tuning of models. Through this, the provided insights are not only technically informative but also contextually relevant to the ongoing evolution of AI technologies.

The theoretical framework of this thesis and especially the privacy and ethical considerations convey a message of necessity for a thoughtful and comprehensive approach in developing and implementing technology. Developers bear a significant responsibility to ensure ethical performance, as the public often fails to recognize these issues and public regulations lag behind.

By upholding the aforementioned ethical standards and evaluating the potential outcomes of the project, we can justify the development and live experimentation of the system in this thesis project. Recognizing that the benefits to societal advancement and the contributions to ethical discourse make the project acceptable and valuable as a demonstration of the feasability and effectiveness of on-device person detection in a practical and realistic setting (Primary Objective 2).

This discussion expands to the broader topic of using sensitive data for training AI models. META AI, specifically, has faced criticism for training their models on personal data without clear communication about their intentions, potentially retaining flexibility for future uses while avoiding misrepresentation. This practice highlights broader ethical concerns about transparency and trust in how AI companies manage and utilize sensitive data (Olifent, 2024).

As mentioned in the Scope of this thesis (see 1.2), the scope of this project extends beyond the immediate technical implementation to a wider examination of its implications and applications. The deployment of the on-device detection system within the FIMUS aquarium illustrates a tangible application of theoretical concepts to practical settings. This showcased the feasibility and utility of edge computing in enhancing visitor analysis without compromising privacy. Another key advantage of the system is it's non-interference with the visitation experience by having the visitors wear custom devices or answer questionnaires as they move through the facilities. This system's ability to generate actionable insights through data visualization, like heatmaps of visitor positions, provides empirical evidence of the technology's value, potentially improving visitor experiences and operational management.

A focal point of this thesis is the robust implementation of privacy-preserving methodologies, critical in garnering public trust and ethical approval for deploying surveillance technologies in sensitive environments. By prioritizing anonymization and minimal data retention, this project aligns with the highest standards of data ethics, reinforcing the importance of privacy considerations in the design and deployment of technological solutions.

Furthermore, this thesis addresses the critical topic of privacy preservation. The emphasis on anonymized data collection and the implementation of privacy-preserving measures underscore the ethical considerations essential for deploying such technologies in public spaces. This focus attempts to not let the technical advancements of the practical implementation come at the expense of individual privacy rights.

# 6 Future Work

This section outlines a comprehensive research agenda based on the findings of the thesis, addressing outstanding questions and challenges to advance the field of on-device processing person localization systems.

**Evaluation of Fine-Tuned Models on Generic Datasets** While this thesis has demonstrated the importance of using deployment-specific data for model evaluations, future research should include a comprehensive assessment of fine-tuned models on generic datasets. This will help establish a more complete understanding of the models' generalizability and their performance across diverse environments, thereby validating their robustness and applicability in varied scenarios.

**Enhancing Dataset Diversity and Quality** Expanding the FIMUS dataset to encompass a variety of environments and conditions would enrich the robustness of future models and facilitate more research. Collaboratively building a dataset consisting of many various real-world scenarios for surveillance-type positions of cameras could facilitate the development of pre-trained weights more specifically developed for person localization in public spaces. This includes gathering data from various public spaces with different lighting, architectural layouts, and population densities. Such diversification will help improve the accuracy and reliability of object detection models, ensuring their effectiveness in a broader spectrum of deployment contexts. Obtaining informed consent for participation in such datasets must be taken seriously to avoid inadvertently using photos of individuals.

**Development of Regulatory Frameworks** Future work should contribute to the development of regulatory frameworks that balance innovation with ethical guidelines. Research should explore the creation of standardized policies and best practices for the deployment of AI technologies in public spaces. Collaboration with legal experts, policymakers, and technologists will be essential to formulate regulations that protect privacy while fostering technological advancement.

**Expanding Practical Applications** Another avenue for future work and research is to ease the implementation cost of on-device processing in systems that may benefit from increased privacy. Potential areas include smart home automation, healthcare monitoring, and personalized user experiences in public venues. Evaluating the feasibility and impact of these applications will help expand the utility of person localization systems in various sectors.

**Heat Map Generation with More Variables** Future work could enhance the solution by integrating additional variables to provide deeper insights. For instance, incorporating data on temperature, weather, and light conditions could help determine how these factors influence visitation patterns. Once the initial technology for privacy-preserving person localization is established, experimenting with these additional data variables could significantly improve the accuracy and utility of the generated insights.

**Zones** Understanding visitor distribution and engagement across different exhibition zones and in front of different exhibitions can significantly enhance operational management and visitor experience. This approach would provide an easy-to-understand and practical visualization of where visitors most frequently stand, offering insights into which exhibitions are the most popular (or time-consuming). Additionally, identifying the zones where queues form can help optimize space and improve visitor flow.

The data gathered could be most useful in the event where certain zones are subject to wear and tear or cleaning based on amount of usage, e.g. pools. If a person localization system could see how many and possibly what age groups are in different pools or swimming facilities at different

times, one could possibly allocate certain time periods for usage which would better overlap with the actual usage of facilities.

# 7 Conclusions

**Summary of Findings** This thesis has explored the viability and ethical implications of developing on-device person detection systems, with a strong emphasis on privacy preservation. The research demonstrates that on-device processing significantly enhances data security and privacy, effectively mitigating the risks associated with person localization systems in public spaces, while still enabling the analysis of visitor behavior.

**Model Evaluation and Dataset Relevance** Through extensive experimentation, it has been established that the validity of object detection model evaluations improves when tested on data specific to their intended deployment environment. This was evident from the significant variation in scores when the models were tested on the specific deployment dataset vs the generic datasets they've been evaluated on before. This finding underscores the importance of using relevant and high-quality datasets for model evaluation to achieve accurate and practical results. The analysis of this thesis lack evaluations of fine-tuned models on generic datasets for a complete assessment of the fine-tuned models accuracies on deployment-specific test data relative to generic data.

**Technological Innovation and Ethical Framework** The thesis delved into the ethical considerations surrounding the deployment of person localization systems. It highlighted the potential risks of mass public control and emphasized the need for a balanced approach that weighs the benefits of technological advancements against the risks to privacy and individual freedoms. The consumer trend to prioritize convenience over privacy in smart home applications hightlight the need for regulations such as GDPR and NIS2 to control technology. The application of utilitarian principles, along with insights from modern philosophical discourse, serves as a framework for assessing the ethical deployment of these technologies. This thesis underscores the importance of a holistic approach that balances technical efficacy with privacy concerns.

**Practical Implementation** The practical implementation at the "Fiskeri og Søfartsmuseet" aquarium showcased the technical feasibility and utility of on-device detection technologies in real-world settings. The successful collection and analysis of visitor data demonstrated the system's potential for enhancing crowd management and operational efficiency in public venues. Importantly, the project's focus on anonymized data collection and privacy-preserving measures exemplifies how similar person localization systems can be integrated into societal infrastructures without infringing on individual privacy.

**Future Directions and Policy Recommendations** The rapid pace of AI development necessitates that regulators possess a nuanced understanding of the technologies they seek to govern, ensuring that laws and policies are both protective, up-to-date, and conducive to innovation. Developers should be encouraged to pursue creative solutions while adhering to ethical guidelines. A balance between innovation and ethics must be achieved through collaborative efforts between technologists, ethicists, policymakers, and the public. This thesis contributes to the ongoing discourse by bridging the knowledge gap, and can help foster collaborative and communicative efforts in this area.

**Final Remarks** *Overall, this thesis contributes valuable insights into the development and implementation of privacy-preserving person localization systems. By addressing both technical and social dimensions, it provides a comprehensive understanding of the challenges and opportunities for these systems. The insights advocate for continued innovation, informed by rigorous ethical standards and regulations, to ensure that technological advancements serve societal well-being while safeguarding individual privacy.*

# Bibliography


aiMotive Team. (2021). *Efficiency, not utilization or tops: Why it matters*. https://medium.com/aimotive/efficiency-not-utilization-or-tops-why-it-matters-a4ef1301c5e5/ (cit. on p. 38).

Antunes, R. S., André da Costa, C., Küderle, A., Yari, I. A., & Eskofier, B. (2022). Federated learning for healthcare: Systematic review and architecture proposal. *13*(4). https://doi.org/10.1145/3501813 (cit. on p. 17).

Bibbo, L., Carotenuto, R., & Della Corte, F. (2022). An overview of indoor localization system for human activity recognition (har) in healthcare. *Sensors*, *22*(21). https://doi.org/10.3390/s22218119 (cit. on p. 1).

Blum, A., Ligett, K., & Roth, A. (2011). A learning theory approach to non-interactive database privacy (cit. on p. 18).

Boesch, G. (2023). *Yolov7: The most powerful object detection algorithm (2024 guide)*. https://viso.ai/deep-learning/yolov7-guide/ (cit. on p. 37).

Brandeis, L. D., & Warren, S. (1890). The right to privacy. *Harvard Law Review*, *4*(5), 193–220. https://doi.org/10.2307/1321160 (cit. on p. 14).

Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., & Zagoruyko, S. (2020). End-to-end object detection with transformers. https://doi.org/https://doi.org/10.48550/arXiv.2005.12872 (cit. on pp. 29, 30, 67).

Chhabria, P. (2022). *Expert a 1m53s 16x9*. https://www.youtube.com/watch?v=w78U7w33NTI&t=5s (cit. on p. 36).

Dictionary, O. E. (2023, July). Gamification (n.) https://doi.org/10.1093/OED/7320229446 (cit. on p. 24).

Dwork, C., McSherry, F., Nissim, K., & Smith, A. D. (2006). Calibrating noise to sensitivity in private data analysis. *Theory of Cryptography*, 265–284 (cit. on p. 18).

Dwork, C., McSherry, F., Nissim, K., & Smith, A. D. (2011). Differential privacy - a primer for the perplexed. https://api.semanticscholar.org/CorpusID:2583736 (cit. on p. 18).

Edgcomb, A., & Vahid, F. (2012). Privacy perception and fall detection accuracy for in-home video assistive monitoring with privacy enhancements. *SIGHIT Rec.*, *2*(2), 6–15. https://doi.org/10.1145/2384556.2384557 (cit. on pp. 21, 72).

Elias, A. R., Golubovic, N., Krintz, C., & Wolski, R. (2017). Where's the bear? - automating wildlife image processing using iot and edge cloud systems [PDF available here]. *2017 IEEE/ACM Second International Conference on Internet-of-Things Design and Implementation (IoTDI)*, 247–258. https://ieeexplore.ieee.org/document/7946882 (cit. on pp. 20, 32, 33).

Eufy. (2022). *What makes eufycam 3c stand out?* https://eu.eufy.com/pages/security-eufycam3c (cit. on p. 35).

European Parliament and Council of the European Union. (2022). Directive (eu) 2022/2555 of the european parliament and of the council of 14 december 2022 on measures for a high common level of cybersecurity across the union, amending regulations (eu) no 910/2014 and (eu) 2016/679 and directive (eu) 2018/1972 and repealing directive (eu) 2016/1148 [Accessed: 2024-03-15]. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32022L2555 (cit. on p. 14).

Everingham, M., Gool, L. V., Williams, C. K. I., Winn, J., & Zisserman, A. (2010). The pascal visual object classes (voc) challenge. https://doi.org/10.1007/s11263-009-0275-4 (cit. on p. 23).

Fischer-Hbner, S., & Berthold, S. (2017). Chapter 53 - privacy-enhancing technologies. In J. R. Vacca (Ed.), *Computer and information security handbook (third edition)* (Third Edition, pp. 759–778). Morgan Kaufmann. https://doi.org/https://doi.org/10.1016/B978-0-12-803843-7.00053-3 (cit. on p. 14).

Gao, L., Tao, S., Ji, C., & Wang, B. (2016/11). Development of person localization and activities recognition system. *Proceedings of the 6th International Workshop of Advanced Manufacturing and Automation*, 257–260. https://doi.org/10.2991/iwama-16.2016.48 (cit. on p. 5).

Gu, R., Niu, C., Wu, F., Chen, G., Hu, C., Lyu, C., & Wu, Z. (2021). From server-based to client-based machine learning: A comprehensive survey. *ACM Comput. Surv.*, *54*(1). https://doi.org/10.1145/3424660 (cit. on p. 17).

Gupta, G. (2019). *What's the difference between single-, double-, multi- and mixed-precision computing?* NVIDIA. https://blogs.nvidia.com/blog/whats-the-difference-between-single-double-multi-and-mixed-precision-computing/ (cit. on p. 39).

Healey, A. J., Fathi, P., & Karmakar, N. C. (2020). Rfid sensors in medical applications. *IEEE Journal of Radio Frequency Identification*, *4*(3), 212–221. https://doi.org/10.1109/JRFID.2020.2997708 (cit. on p. 41).

Huang, Y., Li, Y. J., & Cai, Z. (2023). Security and privacy in metaverse: A comprehensive survey. *Big Data Mining and Analytics*, *6*(2), 234–247. https://doi.org/10.26599/BDMA.2022.9020047 (cit. on pp. 16, 17).

Huang, Z., Yang, S., Zhou, M., Gong, Z., Abusorrah, A., Lin, C., & Huang, Z. (2022). Making accurate object detection at the edge: Review and new approach. *Artificial Intelligence Review*, *55*(3), 2245–2274. https://doi.org/10.1007/s10462-021-10059-3 (cit. on p. 18).

James Gallagher, P. S. (2024). How to train yolov9 on a custom dataset [Accessed: 2024-05-21]. *Roboflow Blog*. https://blog.roboflow.com/train-yolov9-model/ (cit. on p. 56).

Lanir, J., Kuflik, T., Sheidin, J., Yavin, N., Leiderman, K., & Segal, M. (2017). Visualizing museum visitors' behavior: Where do they go and what do they do there? *Personal and Ubiquitous Computing*, *21*(2), 313–326. https://doi.org/10.1007/s00779-016-0994-9 (cit. on p. 11).

Leo Ueno, T. L. (2024). GPT-4o: The comprehensive guide and explanation [Accessed: 2024-05-23]. *Roboflow Blog*. https://blog.roboflow.com/gpt-4o-vision-use-cases/ (cit. on p. 35).

Li, Q., Niaz, U., & Merialdo, B. (2012). An improved algorithm on viola-jones object detector. *2012 10th International Workshop on Content-Based Multimedia Indexing (CBMI)*, 1–6. https://doi.org/10.1109/CBMI.2012.6269796 (cit. on p. 28).

Lin, T.-Y., Maire, M., Belongie, S., Bourdev, L., Girshick, R., Hays, J., Perona, P., Ramanan, D., Zitnick, C. L., & Dollár, P. (2015). Microsoft coco: Common objects in context. (Cit. on p. 51).

Lin, T., Maire, M., Belongie, S. J., Bourdev, L. D., Girshick, R. B., Hays, J., Perona, P., Ramanan, D., Doll'a r, P., & Zitnick, C. L. (2014). Microsoft COCO: common objects in context. *CoRR*, *abs/1405.0312*. https://doi.org/10.48550/arXiv.1405.0312 (cit. on pp. 23, 28).

Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Li, C., Yang, J., Su, H., Zhu, J., et al. (2023). Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499* (cit. on p. 50).

Luccioni, S. (2023). Ai is dangerous, but not for the reasons you think author [Accessed: 2024-05-17]. https://www.youtube.com/watch?v=eXdVDhOGqoE&ab_channel=TED (cit. on p. 24).

Ma, C., Shimada, A., Uchiyama, H., Nagahara, H., & Taniguchi, R.-i. (2019). Fall detection using optical level anonymous image sensing system. *Optics & Laser Technology*, *110*, 44–61. https://doi.org/10.1016/j.optlastec.2018.07.013 (cit. on p. 21).

Maayah, M., Abunada, A., Al-Janahi, K., Ahmed, M. E., & Qadir, J. (2023). Limitaccess: On-device tinyml based robust speech recognition and age classification. *Discover Artificial Intelligence*, *3*(1), 8. https://doi.org/10.1007/s44163-023-00051-x (cit. on p. 88).

Machanavajjhala, A., Kifer, D., Gehrke, J., & Venkitasubramaniam, M. (2007). L-diversity: Privacy beyond k-anonymity. *ACM Trans. Knowl. Discov. Data*, *1*(1), 3–es. https://doi.org/10.1145/1217299.1217302 (cit. on p. 16).

META AI. (2024). Ms coco. https://paperswithcode.com/dataset/coco (cit. on p. 27).

Murali, N. (2021). *Image classification vs semantic segmentation vs instance segmentation*. https://nirmalamurali.medium.com/image-classification-vs-semantic-segmentation-vs-instance-segmentation-625c33a08d50 (cit. on p. 4).

Neuman, S. M., Plancher, B., Duisterhof, B. P., Krishnan, S., Banbury, C., Mazumder, M., Prakash, S., Jabbour, J., Faust, A., de Croon, G. C., & Reddi, V. J. (2022). Tiny robot learning: Challenges and directions for machine learning in resource-constrained robots. *2022 IEEE 4th International Conference on Artificial Intelligence Circuits and Systems (AICAS)*, 296–299. https://doi.org/10.1109/AICAS54282.2022.9870000 (cit. on pp. 4, 88).

Olifent, L. (2024). Meta tager endnu en bid af dit privatliv: Protester virker, mener eksperter. *Ingeniøren*. https://ing.dk/artikel/meta-tager-endnu-en-bid-af-dit-privatliv-protester-virker-mener-eksperter (cit. on p. 79).

OpenAI. (2024). Chatgpt (june 2024 version) [Retrieved from https://www.openai.com/]. (Cit. on p. 5).

OpenCV. (2022). Intersection over union (iou) in object detection and segmentation [Accessed: 2024-05-29]. https://learnopencv.com/intersection-over-union-iou-in-object-detection-and-segmentation/ (cit. on p. 27).
Orwell, G. (1949). *1984*. Secker; Warburg. (Cit. on p. 22).
Paiano, M., Martina, S., Giannelli, C., & Caruso, F. (2023). Transfer learning with generative models for object detection on limited datasets. *arXiv*. https://arxiv.org/abs/2402.06784 (cit. on pp. 31, 32).
Park, J., Chen, J., Cho, Y. K., Kang, D. Y., & Son, B. J. (2020). Cnn-based person detection using infrared images for night-time intrusion warning systems. *Sensors*, *20*(1). https://doi.org/10.3390/s20010034 (cit. on p. 30).
Pérez Cortés, L. E., Ha, J., Su, M., Nelson, B., Bowman, C., & Bowman, J. (2023). Gleaning museum visitors' behaviors by analyzing questions asked in a mobile app. *Educational technology research and development*, *71*(3), 1209–1231. https://doi.org/10.1007/s11423-023-10208-1 (cit. on p. 11).
Rajapakse, V., Karunanayake, I., & Ahmed, N. (2023). Intelligence at the extreme edge: A survey on reformable tinyml. *ACM Comput. Surv.*, *55*(13s). https://doi.org/10.1145/3583683 (cit. on pp. 5, 88).
Ravi, S., Climent-Pérez, P., & Florez-Revuelta, F. (2023). A review on visual privacy preservation techniques for active and assisted living. *Multimedia Tools and Applications*, *Online First*, 1–16. https://doi.org/10.1007/s11042-023-15775-2 (cit. on p. 16).
Redmon, J. (2020a). Joseph redmon's twitter [Accessed: 2024-05-09]. *Twitter*. https://twitter.com/pjreddie/status/1230524770350817280?lang=en (cit. on p. 22).
Redmon, J. (2020b). Joseph redmon's twitter [Accessed: 2024-05-09]. *Twitter*. https://twitter.com/pjreddie/status/1230524770350817280?lang=en (cit. on p. 22).
Sandtrø, J. (2022). *Webcast: Hvordan unngå å bryte regelverket (gdpr)*. SuperOffice Norge on Youtube. https://www.youtube.com/watch?v=FB2P-ijCIKw&ab_channel=SuperOfficeNorge (cit. on p. 13).
Sasagawa, Y., & Nagahara, H. (2020). Yolo in the dark - domain adaptation method for merging multiple models. *Proceedings of the European Conference on Computer Vision (ECCV)*. https://doi.org/10.1007/978-3-030-58589-1_21 (cit. on p. 31).
Saurav, S., Saini, A. K., Saini, R., & Singh, S. (2022). Deep learning inspired intelligent embedded system for haptic rendering of facial emotions to the blind. *Neural Computing and Applications*, *34*(6), 4595–4623. https://doi.org/10.1007/s00521-021-06613-3 (cit. on p. 23).
Shao, S., Zhao, Z., Li, B., Xiao, T., Yu, G., Zhang, X., & Sun, J. (2018). Crowdhuman: A benchmark for detecting human in a crowd. *arXiv preprint arXiv:1805.00123* (cit. on p. 52).
Sharma, P. (2023). *Role of weight transmission protocol in machine learning*. https://www.tutorialspoint.com/role-of-weight-transmission-protocol-in-machine-learning (cit. on p. 17).
Sweeney, L. (2002). K-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, *10*(05), 557–570. https://doi.org/10.1142/S0218488502001648 (cit. on p. 15).
Tarzia, S. P., Dick, R. P., Dinda, P. A., & Memik, G. (2009). Sonar-based measurement of user presence and attention. *Proceedings of the 11th International Conference on Ubiquitous Computing*, 89–92. https://doi.org/10.1145/1620545.1620559 (cit. on p. 41).
Termly. (2023). Natural person [Accessed: 2024-05-15]. https://termly.io/legal-dictionary/natural-person/ (cit. on p. 14).
The European Parliament. (2016). *Eu directive 2016/679 general data protection regulation (gdpr)*. Official J Eur Union 2014. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A32016R0679 (cit. on p. 12).
Viola, P., & Jones, M. (2001). Rapid object detection using a boosted cascade of simple features. *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, *1*, I–I. https://doi.org/10.1109/CVPR.2001.990517 (cit. on p. 28).
Vizetelly, F. H. (1920). *A desk-book of errors in english*. Project Gutenberg. https://www.gutenberg.org/files/48907/48907-h/48907-h.htm (cit. on p. 2).
Wang, C.-Y., & Liao, H.-Y. M. (2024). YOLOv9: Learning what you want to learn using programmable gradient information (cit. on pp. vi, vii, 30, 64).

Wang, X., Ellul, J., & Azzopardi, G. (2020). Elderly Fall Detection Systems: A Literature Survey. *Frontiers in Robotics and AI*, *7*. https://doi.org/10.3389/frobt.2020.00071 (cit. on pp. 15, 20).

Wei, C., Wu, G., & Barth, M. J. (2024). Feature corrective transfer learning: End-to-end solutions to object detection in non-ideal visual conditions. *arXiv preprint arXiv:2404.11214*. https://arxiv.org/abs/2404.11214 (cit. on p. 31).

Westin, A. (1967). *Privacy and freedom*. Atheneum. (Cit. on p. 14).

Yang, Q., Liu, Y., Chen, T., & Tong, Y. (2019). Federated machine learning: Concept and applications. *ACM Trans. Intell. Syst. Technol.*, *10*(2). https://doi.org/10.1145/3298981 (cit. on p. 17).

Zheng, L., Zhang, H., Sun, S., Chandraker, M., Yang, Y., & Tian, Q. (2017). Person re-identification in the wild. (Cit. on p. 52).

Zheng, S., Apthorpe, N., Chetty, M., & Feamster, N. (2018). User perceptions of smart home iot privacy. *Proc. ACM Hum.-Comput. Interact.*, *2*(CSCW). https://doi.org/10.1145/3274469 (cit. on p. 12).

Zou, Z., Chen, K., Shi, Z., Guo, Y., & Ye, J. (2023). Object detection in 20 years: A survey. *Proceedings of the IEEE*, *111*(3), 257–276. https://doi.org/10.1109/JPROC.2023.3238524 (cit. on pp. 19, 24, 27, 28).

# A Code Snippets

```
class CameraHandler:
    def __init__(self, awb_gains, awb_mode, brightness, contrast, exposure_compensation,
    exposure_mode, image_format, iso, meter_mode, resolution, sensor_mode, shutter_speed, framerate)
    :
        # Create camera with the arguments
        cam = picamera.PiCamera(resolution=resolution, sensor_mode=sensor_mode,
        framerate=framerate)
        cam.iso = iso

        # Wait for the automatic gain control to settle
        time.sleep(2)
        cam.shutter_speed = shutter_speed
        cam.exposure_mode = exposure_mode
        cam.awb_mode = awb_mode
        cam.awb_gains = awb_gains
        cam.exposure_compensation = exposure_compensation
        cam.brightness = brightness
        cam.contrast = contrast
        self.image_format = image_format
        cam.meter_mode = meter_mode
        self.picamera = cam
```

Figure 43: CameraHandler Class Initialization

An attempt was made to set the AWB (Automatic White Balance) gains on the Pi camera, but the setting had no effect.

The solution was that AWB gains adjustments only take effect after the AWB mode is set to 'off'. Additionally, the new settings are not applied until after an image is captured. Therefore, to ensure that the AWB gains were effectively set, it was necessary to first disable the AWB mode, capture an image, and then adjust the gains. This sequence ensured that changes were accurately reflected in subsequent images. Figure 43 displays the ordering of settings for consistent image capture.

# B Camera Settings Explanation

- *awb_gains* - Set the auto white balance gains red and blue. Set as a (red, blue) set. Each value may range from 0.0 to 8.0. Typical is 0.9-1.9. Only has an effect when *awb_mode* is 'off'. IMPORTANT: awb and exposure mode must be set to off BEFORE setting the *awb_gains*.
- *awb_mode* - Auto white balance. Default is auto. Disabling auto white balance mode allows for manual setting of AWB gains, ensuring consistent image color temperature. 'off' 'auto' 'sunlight' 'cloudy' 'shade' 'tungsten' 'fluorescent' 'incandescent' 'flash' 'horizon'
- *brightness* - Adjusts the post-processing brightness of the image. Default is 50, representing no adjustment. 0 to 100.
- contrast - Adjusts the post-processing contrast of the image. Default is 0, representing no adjustment. -100 to 100.
- *exposure_compensation*- Adjusts the exposure compensation level. Range is -25 to 25. Default is 0.
- *exposure_mode* - Disabling auto-exposure allows for manual control over exposure settings. 'off' 'auto' 'night' 'nightpreview' 'backlight' 'spotlight' 'sports' 'snow' 'beach' 'verylong' 'fixedfps' 'antishake' 'fireworks'.
- *exposure_speed* - Indicates the effective exposure speed, which may differ from the set shutter speed after adjustments.

- *framerate* - Sets the number of frames per second captured by the camera.
- *iso* - Sets the ISO sensitivity of the camera sensor. Values: 100, 200, 320, 400, 500, 640, 800. 0 is auto.
- *metering_mode* - Sets the metering mode. 'average' 'spot' 'backlit' 'matrix'. Backlit is the largest area. Default is average.
- *sensor_mode* - Controls the sensor mode, where '3' typically corresponds to standard image capturing.
- *shutter_speed* - Sets the shutter speed in microseconds. 0 to 6000000. Default 0. 0 is auto. Max 6s.
- *resolution* - Sets the resolution of the image frame.

# C TinyML and Frugal Devices

As mentioned in Section 1.3, TinyML is when machine learning models are aimed at deployment to heavily resource constrained environments, e.g. what is called frugal devices. These are devices where the microcontroller units (MCUs) are accompanied by memory measured in kilobytes, and processor speeds measured in megahertz.

Machine learning networks applied to tiny robots are subject to challenges from size, weight, area, and power (SWAP) (Neuman et al., 2022). Many of the same challenges apply even in applications where the SWAP challenges are not the main concerns. Rajapakse et al. mentions the open challenges and future directions of the next generation tinyML. Catastrophic forgetting, which is when information from previous tasks while learning new ones are forgotten, are a result of the frugal devices' computational resources and memory. The first recommendation for future directions from the authors is to investigate fog computing as a means to offload tasks from the frugal devices.

Maayah et al. (2023) explore the ways of speech processing on microcontrollers to improve car AI systems. They employed their trained and optimized model to an Arduino Nano 33 BLE. The model achieved accuracies in above 85 percent on recognizing whether the voice was that of an adult or a child, and to detect whether the speech was a replay (synthetic) or "live".

Furthermore, Rajapakse et al. discuss some of the challenges in industrial IoT environments with several smart object devices, where having the devices share a collective dataset of anomalies within a manufacturing environment would be advantageous for utilizing collective learning to improve the ML models in each of the devices (2023). This means the devices will all learn from observations of the other devices, such that the training period from when a network of devices is deployed within a new environment to when they are fully functioning with regards to accuracy in their predictions is reduced. See more about this in section 2.5.2 about federated learning as a way of implementing a collective learning network for the edge devices.