NTNU
Kunnskap for en bedre verden

Department of Mechanical and Industrial Engineering

Master's Thesis - robotics & automation

Efficient, accurate, and privacy-preservant object detection in edge devices

Student:
Hallvard Enger Bjørgen

Supervisor at NTNU:
Amund Skavhaug

Trondheim/Esbjerg Spring 2024

Table of Contents

List of Figures

List of Tables

Abstract

Placeholders:
What
Why
How
Principal contributions
Principal conclusion

1 Introduction

On-device processing is emerging as a vital component of modern human detection and tracking systems as an approach to ensure privacy. The ability to detect and track humans in real-time is crucial for a wide range of applications, from security surveillance to visitor analytics in cultural institutions. However, the deployment of such systems raises concerns about privacy and data security. Particularly in sensitive environments like museums and aquariums. This thesis details the development of a privacy-preserving human localization system. The developed system was then deployed and tested in a acquarium in Denmark.

The method of human detection and tracking in public spaces has evolved significantly over the past decade, driven by advancements in computer vision and machine learning. Traditional surveillance systems relied on centralized processing, where video feeds were transmitted to a remote server for manual human analysis. However, this approach raised privacy concerns as it involved transmitting raw video data over the network, potentially exposing sensitive information. Additionally, it also required a human to manually analyze the video feed, which was time-consuming, prone to errors, lacking of scalability, and not privacy-preserving.

On-device processing addresses this issue by performing analytics locally on the edge device, removing the need to transmit raw video data and thus enhancing privacy.

A device was deployed in the acquarium of "Fiskeri og Søfartsmuseet" in Esbjerg, Denmark. This was done to demonstrate the feasibility and effectiveness of such on-device human detection and localization in a practical and realistic setting. The system not only faced the inherent challenges of running the analysis on-device, but also faced the challenges of a suboptimal lighting environment. Such challenges are typical in museums and acquariums where the art may be damaged, the mood may be spoiled, or the fish may be disturbed if the light is to bright. The project in this thesis demonstrates how to overcome said challenges. A Raspberry Pi 4 with a camera V2.1 module and a pre-trained yolo object detection model was used. The thesis further explores how using labeled images from the museum environment as the training dataset may affect the models performance. The implementation is described in detail. The dataset is available on Google Drive: todo.

todo amund er det over nok om background and motivation, problem description?

1.1 Scope

1.1.1 Research Questions

1.1.2 Research Objectives

1.2 Structure

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 object detection and visitor behaviour analysis in cultural institutions, focusing on privacy-preserving techniques, dataset specialization for enhanced object detection accuracy, and case studies of technology implementation.

2.1 Analysis of Traditional and Technological Methods for Visitor Behavior Studies

Studies on traditional methods of analyzing visitor behavior[1] reveal both their use cases and limitations.

Lanir et al. explored alternative technological approaches in their study on museum visitor behavior and its perceived value to museum curators (2017). Wearable RFID trackers 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. They noted that technology-based visitor behavior analysis was generally well-received by museum curators, offering valuable data that could enhance the visitor experience.

However, the requirement for visitors to wear RFID trackers represents a significant drawback, as it may be perceived as intrusive (although completely privacy preservant).

The study of Lanir et al. also highlighted divergent views on the utility of visitor behavior 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 behavior would be to satisfy my curiosity." This contrast underscores the varied perspectives within museums regarding the value and implications of behavior analysis technologies.

2.2 Privacy

Section 2.2, is a rewritten and improved version of the same section in an unpublished pre-thesis project written by the same candidate as this master thesis.

2.2.1 The General Data Protection Regulation

This section serves as a summarization of some aspects of the GDPR relevant to the thesis.

The GDPR entered into applicability in the EU on 25th of May 2018 as a way for people to have more control over their data, and for having a level playing field for all companies. There is now one set of data protection rules for all companies operating in the European Economic Area (EEA). The EEA consists of all EU countries plus Iceland, Liechtenstein and Norway. The most relevant parts of the GDPR are the regulations regarding personal data.

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:

[1] Traditional methods include surveys, manual counting, and direct observation

Definition of personal data, as given by EUs GDPR

"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).

Regulations regarding personal data also applies to the events where pieces of information are aggregated to identify a person. It is possible to store information about individuals without it being personal data. This can be done in several ways, one of them being by the method of differential privacy. Differential privacy is explained in section 2.2.5.

Another approach to the personal data is to process it the right way.

Legal Bases for Processing Personal Data Processing of personal data is permissible under the GDPR only when it satisfies at least one of the following legal bases:

1. The data subject has given explicit consent.
2. It is necessary for the performance of a contract to which the data subject is a party.
3. It is necessary for compliance with a legal obligation to which the controller is subject.
4. It is necessary to protect the vital interests of the data subject or of another natural person.
5. 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.
6. 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 1) responsible for compliance with the 3 requirements summarized below, and 2) should be able to demonstrate this compliance at any given time.

1. 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).
2. 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).
3. 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.

The 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 throughout the Union. 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.

2.2.2 Preservation of Individual Privacy

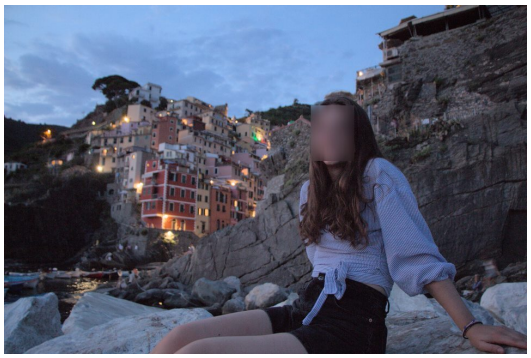
Preservation of individual privacy refers to maintaining the personal space and confidentiality of individuals, ensuring that their private lives and personal integrity are not invaded or exposed without consent. This involves considerations beyond just data, including behaviors and surveillance. Protection of personal data specifically 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. While protection of personal data is important due to the regulations, preserving the individual privacy is essential in object detection of persons.

There are multiple methods, both pre- and post-processing, for preserving individual privacy. 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 1, demonstrating practical implementations.

K-anonymity claimed to be a mathematically proven method for anonymyzation of personal data, but has been critizised by it's successor, the l-diversity criterion, for not being robust in the events where attackers have background data (ma2007l-diversity). 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 2.2.5.

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

Figure 1: Six methods to enhance individual privacy in images

There's also been much research with regards to preserving privacy in the context of machine learning (Ravi et al., 2023). Common to all use cases is the principle that deleting the data is the most definitive method of ensuring privacy, provided that such an action is feasible. When only non-personal data remains stored, the application is unequivocally secure in terms of privacy.

In addition to not protecting the identity of individuals, it is often important that the individuals *feel* their privacy is being preserved.

2.2.3 User perceptions of smart home IoT 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 (Zheng et al.). What the researchers found, shows promise to IoT edge computing visual devices; the convenience and con-

nectedness of the devices surpasses the desire to preserve privacy. Another research question of the study was user perceptions of obsolescence of the IoT devices, as there are frequent upgrades and new products on the market.

> **Responses regarding privacy and obsoleteness of IoT devices (Zheng et al., 2018)**
> "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).

2.2.4 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).

The concept of FL can be seen in figure 2, and is best described in the article of Antunes et al. (2022): "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). 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".

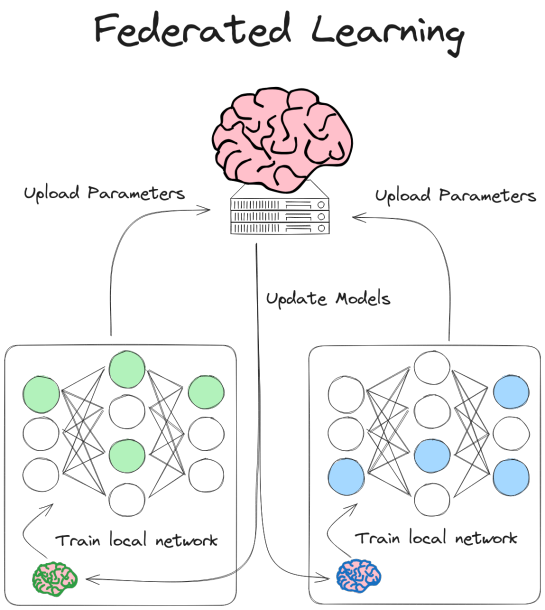


Figure 2: The federated learning process
Federated learning (FL), also known as collaborative learning, is a decentralized approach to training machine learning models. It doesn't require an exchange of data from client devices to global servers.

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[2] 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 2.2.5.

2.2.5 Differential privacy

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 3.

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)."

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.

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

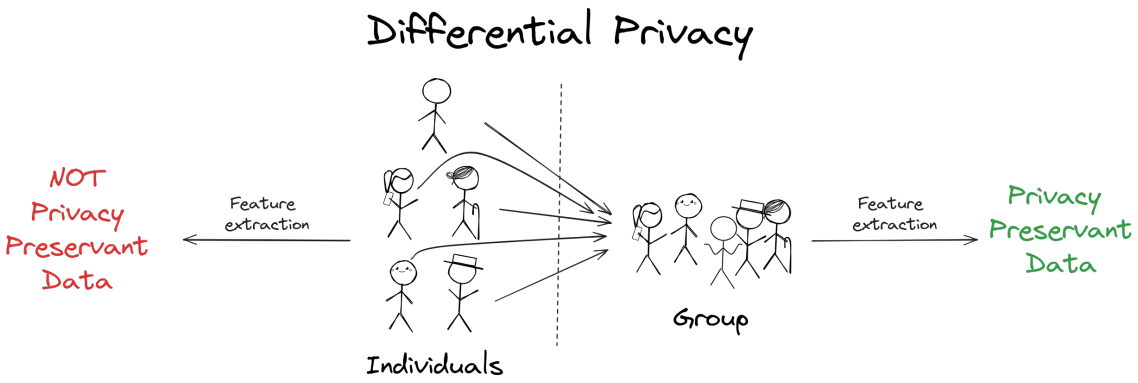


Figure 3: The differential privacy concept

2.2.6 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 Fast calculation	Much bandwidth
Traditional ML	Little resource consumption	Relying on the Internet Poor robustness
Task sharing	Reducing stress on a single device	Much bandwidth Large latency
Deep learning	Privacy protection 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 must be done in a losses way to maintain the image quality to preserve the accuracy of the detectors. Doing so is not trivial, and is a research field on it's own. A few methods are discussed in section **??**.

Traditional machine learning The second method of running traditional machine learning methods, might not the greatest solution either, as they have been less accurate than the deep learning models (see figure **??**). They may, however, be a good option for devices with low computing power and memory resources as they are generally low-demanding. The methods need less data, are more transparent, but 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. Figure 4 illustrates a road map of what have been the most popular machine learning approaches to object detection. 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, discussed in section **??**.

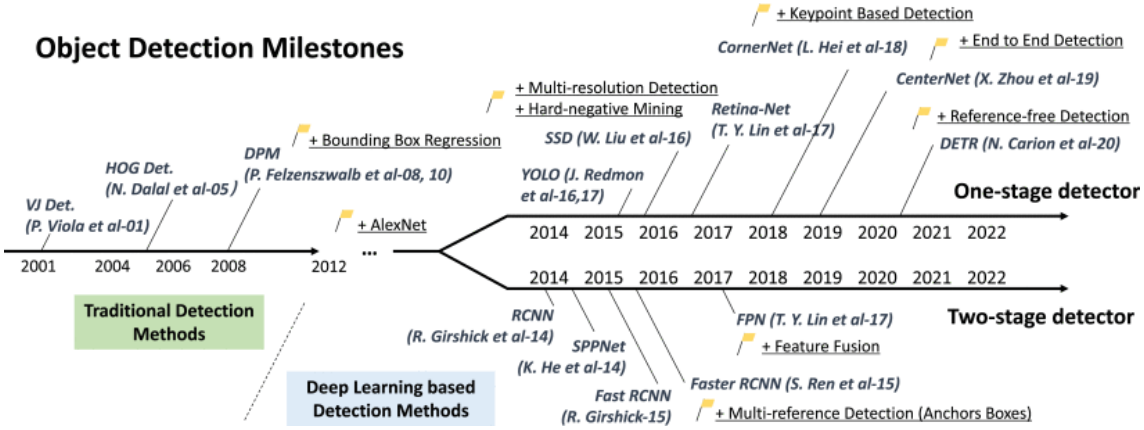


Figure 4: The evolution of object detection (Zou et al., 2023)

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 **??**), where camera devices take images and send them to a more powerful computer for doing the heavy computing. This reserves privacy as the images are never sent outside of the local network, and can be achieved by simple TCP/IP[3] communication between the nodes. This gives 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 scarcity in specialized hardware such as the GPU, this could be a nice solution, as one GPU per facility may be sufficient and 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.

2.2.7 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.

2.2.8 Deletion of images

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. This may be achieved by the following approach called edge computing.

Edge computing is to have the visual device perform the analysis directly on the device (so called 'on-device processing'). With this approach, the image is obscured or deleted right after analysis without ever leaving the edge device, and 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

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

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.

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

2.2.9 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, implying it has been done intentionally. 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, suggesting something is unclear or out of focus. One might say an image has been obscured by blurring the image, or it may be done by other methods such as masking or pixelate the faces of individuals. These methods are illustrated in figure 1.

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 enhancements 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[4] scale from 1-5. The obfuscation methods, or privacy enhancements as they called them, and the results are displayed in figure 5. 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.

[4]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.

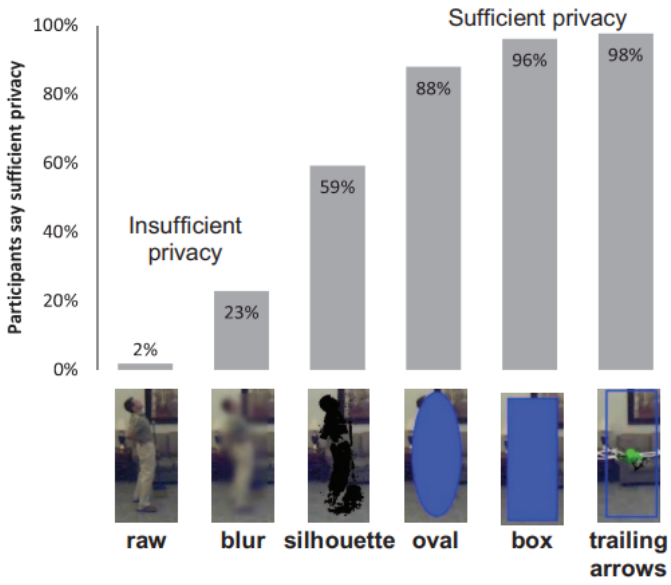


Figure 5: Privacy enhancements methods in the study of Edgcomb and Vahid (2012) The study indicates that the methods of blurring an image is not perceived privacy preservant. However, as we will see in section **??**, the results of this study might not be a reason by itself to throw the method of blurring an image out the window.

2.2.10 Ethical Considerations in the Development of Localization Technologies

As we advance the capabilities of technologies such as YOLOv9 for human localization, it becomes imperative to consider the ethical implications of our developments. The narrative of George Orwell's dystopian novel *1984* serves as a stark reminder of the potential societal consequences of pervasive surveillance and control. 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.

Historical Precedents and Personal Accountability The decision by Joseph Redmon, the creator of the initial versions of YOLO, to cease his work on the project due to its military applications illustrates a profound ethical stance. Redmon's choice underscores the responsibility that developers bear in considering the broader impacts of their work, paralleling Orwell's caution against allowing technological advances to override ethical judgments. Joseph Redmon's resignation from the development of YOLO due to ethical concerns 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 AI development is still very much ongoing.

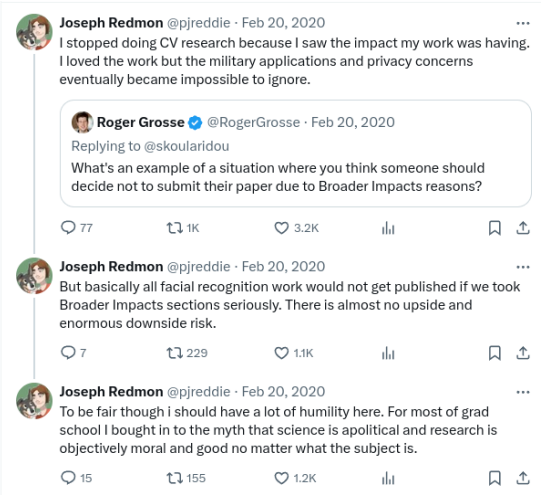


Figure 6: Redmon Quit Computer Vision Research Due to Ethical Concerns

Greed's Grief

For the love of money is a root of all kinds of evil. Some people, eager for money, have wandered from the faith and pierced themselves with many griefs. -1 Timothy 6:10, The Bible.

Ethical Framework for Development In developing technologies capable of tracking and analyzing human behavior, we must establish robust ethical frameworks that prevent misuse and ensure that advancements enhance societal welfare without infringing on individual rights and freedoms. This involves transparent development processes, clear privacy safeguards, and continuous monitoring of technology deployment.

Learning from History Just as Orwell warns against the dangers of forgetting or altering history, 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.

Conclusion 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.

2.3 Object Detection

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 **??**. 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[5]. 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.

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). These are explored in greater detail in section 2.3.2 and 2.3.3

Machine learning can be seen as a gamified[6] 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.

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

[6]apply typical elements of game playing (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)

2.3.1 Performance benchmark

Dataset 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 it's introduction in 2014, with it's 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 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.

Accuracy of Model Inferences 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. According to Zou et al., these 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.

However, 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 sum of precision values under various recalls. The mean is when this is averaged for all the object classes in the dataset. The IoU brings bounding box positioning into the equation, representing how well the predicted box fits to the ground truth. A detailed explanation of what this means is given in the following paragraphs.

First we need 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 2.3.1 and 2.3.1, 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[7], the metrics usually used to evaluate it's 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.

[7]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.

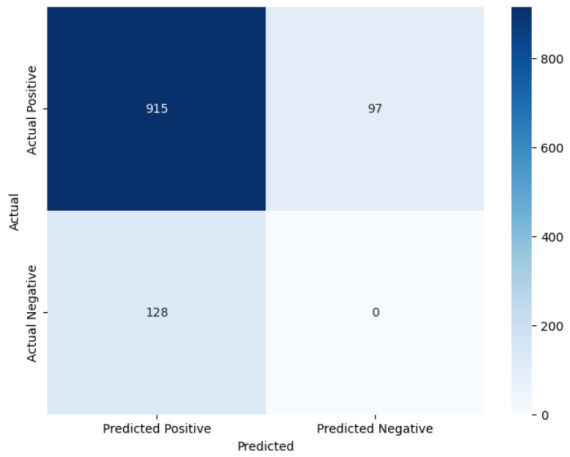


Figure 7: An example of a confusion matrix.

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.
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 7. This model has identified 915 people correctly, failed to identify 128 people, and incorrectly identified 97 instances as people when there were none.

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.

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

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.

$$textRecall = \frac{TP}{TP + FN}$$

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.

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

To assess a regular machine learning model's performance, the precision-recall curve is still a common practice. The precision-recall curve is a graph that shows the trade-off between precision

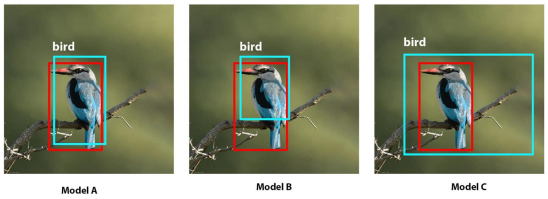


Figure 8: Intersection over Union (IoU) metric

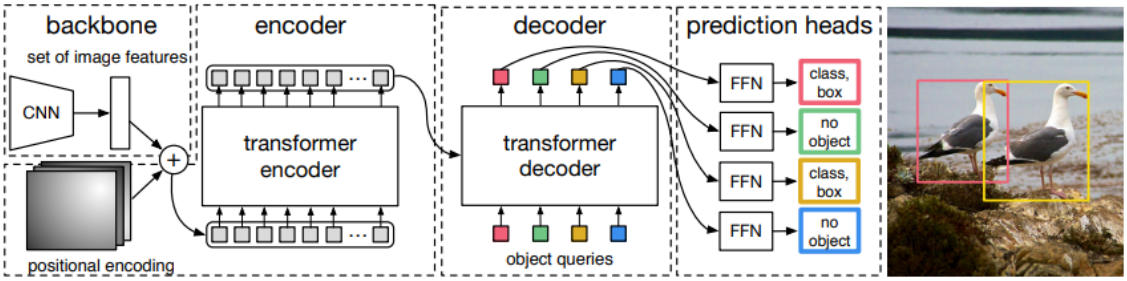


Figure 9: The architecture of the DETR model (Carion et al., 2020).

and recall for different thresholds for confidence in the object class. As you allow your model to be more uncertain in it's inferences on the image, the number of hallucinations will also increase and thus the precision drops.

todo sett inn precision recall curve

Understanding the IoU metric Accuracy in object detection refers to both detecting the object *and* it's 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 8 for an illustration. The equation is simple:

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

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[8], MS-COCO AP is averaged over multiple IoU thresholds between 0.5 and 0.95, which encourages more accurate object detector localizations (Zou et al., 2023.

2.3.2 Real-Time Detection Transformer

The Real Time Detection Transformer (RT-DETR) is a real-time object detection model developed by the Facebook (Meta) Research team. The model uses a Transformer encoder-decoder architecture similar to a large language model. See figure 9 for an illustration of the architecture.

In the paper first introducing the model, Carion et al. claim it outperforms competitive baselines on panoptic segmentation[9] with a simple segmentation head trained on top of a pre-trained DETR.

[8] A 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 having the value the ability to adjust the threshold is a good idea when implementing an object detector

[9] This is a challenging pixel-level segmentation task. In segmentation tasks, an image is divided into meaningful regions. Altough different from object detection.

The conclusions of the paper are as follows: "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, in particular 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).

And

2.3.3 YOLOv9

The YOLO (You Only Look Once) object detection algorithm is a popular choice for real-time object detection. YOLO processes images in a single pass, making it faster than traditional object detection algorithms that require multiple passes. YOLO divides the image into a grid and predicts bounding boxes and class probabilities for each grid cell. The YOLOv9 is an improved version of the original YOLO algorithm, incorporating various enhancements to improve detection accuracy and speed. The YOLOv9 model is pre-trained on the COCO dataset. This is previously introduced dataset contains 80 different classes which the model is able to detect.

YOLOv9 author's comment on the DETR series

"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." (C. Wang et al., 2024).

2.3.4 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. 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. Park et al., 2020.

todo les YOLO in the Dark - Domain Adaptation Method for Merging Multiple Models

2.3.5 Effectiveness of Training Dataset Specialization

Comparison of various algorithms performance enhancement when data has been optimized for the use case. What is the role of a training dataset in the task of determining the weights in a yolov9 artificial neural network, and how can would special training data optimize the weights for a specific use case?

Transfer learning is typically when a trained model is used as a starting point for a new model, and then fine-tuned on a new dataset. This is a useful to extend the trained models capabilities to for example learn how to correctly identify a new object class. In this thesis, the goal is to improve the performance of the YOLOv9 model by training it on a dataset that is specialized for the museum environment. The dataset used for training the YOLOv9 model in this thesis is the COCO dataset, which is a large dataset of images with 80 different classes. The COCO dataset is a general dataset, and the YOLOv9 model trained on this dataset may not perform optimally in the museum environment. By training the YOLOv9 model on datasets that are more similar to the acquarium environment, we will not extend the models capabilities in the same way as what is typically done in transfer learning as we are trying to reduce the number of classes to 1 (person), and we are trying to improve on it's already learnt capability of identifying persons.

The datasets of this project are slightly different from each other The CrowdHuman dataset is bigger than the others, but is not very different than the dataset (COCO) that the pretrained model has already been trained on apart from only containing images where people are the main focus. This dataset is big, and it's performance tells us to which extent adding more data would improve the model. The model was trained with various amounts of training data, to

The football-players dataset contains images that are all from the same perspective and in the same light and quality. This dataset is added to see if the model improves just by specializing to single class data, or if the specialization quality and relevancy matters. It's also used as a matter of having more datasets in the experiment where differences in the datasets are clear enough to attribute differences in performance to the difference in nature of the datasets and not to other confounding factors.

The person reidentification in the wild dataset is the 2nd largest one. This dataset is more similar to our use case than the others, as it exclusively contains images of people. The persons are sometimes occluded, and in many cases they are of approximately the same scale as the persons we will be detecting in the acquarium.

2.4 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.

2.4.1 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 todo denne beskrevet tidligere?. Users can train custom models using this dataset, fine-tuned for scenarios such as monitoring museum traffic. On Roboflows website, there are multiple guides for how such applications may be implemented.

2.4.2 GPT-4 with Vision on Object Detection

The well known large language models (LLM) have been generalized to perform more tasks, and are thus applicable to than just text. The ChatGPT-4 with Vision is one such large *multimodal* model (LMM). Numerous solutions already incorporate OpenAI's chatGPT as a fundamental component of their product. Expanding the role of GPT to include visual processing could potentially yield additional benefits. LLMs with vision may enable applications capable of semantically understanding scenes. This could mean the application may for example understand when a riot is about to break out in a bar street in England, or when a fish tank feeding is taking place in the acquarium, and what the crowds general reactions are to the show [10]. This may allow automated applications to provide insights to their users so they don't need to analyze the data. The resulting solution may be faster, less error prone and more scalable than the "surveillance-system with human interference"-paradigm we have today for public surveillance and intelligence.

One issue arises from the generative nature of the GPTs. It is not given that a model performing well one day will be as good the next. Many experiments are performed to measure the performance of the LMM, and some show promising results. However, most experiments are frozen in time and will not reflect how well the model may perform from one day to the next. This may result in models performing well when tested, but no longer doing their jobs post-deployment.

To tackle this issue, a website has been dedicated to measure how the GPT-4 with Vision [11] 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.

2.4.3 Drawbacks of Utilizing Third-Party Services

The two preceding sections highlights the strenths 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.

[10] Detecting the mood of people is also much researched and could be implemented for images with high enough resolution and quality. One model would then detect people or faces, and another would get cut-outs of those faces to detect the mood of each individual.

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

1. **Complete Control Over the System:** Developing your own application allows for full customization in terms of software architecture, data processing, and system integration. This total control facilitates the optimization of the system to meet specific performance and operational requirements. In addition, a system built separately would have the benefit of being independent from the performance and existence of Roboflow.

2. **Data Privacy and Security:** On-device processing ensures that all data processing is kept on-device, enhancing data security and privacy. Roboflow offers local deployment, but this comes as part of their more expensive business-level subsxription plan.

3. **Cost Efficiency:** Managing your own system can be more cost-effective in the long run, particularly if the application demands extensive processing power or high throughput, as it eliminates recurring costs associated with third-party platforms. Roboflow's plans include costs related to "inference credits", making the system great for small applications but less likely to be a good fit for bigger enterprise solutions looking to leverage the margins. GPT4-V may be accessed via Azure's OpenAI service, which is also priced by how much the service is used and how

4. **Performance Optimization:** Owning the inference system allows for hardware and software optimizations that are not possible when using third-party services. This can lead to better performance, especially in terms of processing speed and latency.

5. **Scalability and Integration Flexibility:** Implementing your own solution allows for easier scaling and integration with existing IT infrastructure, which is beneficial for maintaining seamless data workflows and supporting business growth without being limited by external platform constraints.

2.4.4 Conclusion

In conclusion, third-party services offer convenient solutions that may align perfectly with specific requirements for object detection systems, providing a quick and efficient path to implementation. However, despite their benefits, these services come with inherent drawbacks such as dependency on external providers, potential privacy concerns, and limitations in customization and control. While leveraging third-party services can expedite development, it is imperative for researchers and practitioners in the field of object detection, particularly in contexts such as person detection where privacy may be of cencern, to carefully weigh these considerations. Exploring alternative methods of implementation, including developing systems from scratch, can offer greater flexibility, control, and potential for innovation.

2.5 Summary of Literature Review

A summarization of the current state of research, and where my thesis aims to contribute.

3 Methodology

Two cameras were deployed in a room of acquariums at "Fiskeri- og Søfartsmuseet" in Esbjerg to take images for building a specialized dataset and to evaluate the effects of developing a highly specialized detector rather than using a general.

3.1 Dataset Construction

3.1.1 Camera

Hardware camera can be tweaked by screwing the lens with a mechanical tool to modify it's aperture, which influences it's depth focus. Aperture mechanical setting (camera focus adjustment), depth control. Default not found... images well sharp enough... 50cm to infinity...

3.1.2 Technical Challenges and their Solutions

Pi out of memory on image capture Context: tweaking camera settings, leading to larger size images Error: mal port enable failed to enable connected port... Out of resources. Solution: allocating more memory to the GPU by going through raspi-config Performance Options -¿ GPU Memory

AWB Gains not setting Context: setting picamera.awg gains has no effect Error: no change Solution: set awb gains after awb mode has been set to off, and capture an image. The control seems to not set before after capturing an image. Thus, setting these values and then checking the values, it might seem the modification has not been made altough it will show on the images.

3.1.3 Image Capture

The dataset was built by capturing images while no visitors were present in the acquarium. Due to the constraint to only operate within opening hours when the facility was open to everyone, a way to cancel image capturing was needed in the case if someone entered the room. One of the goals of the dataset was to have the images taken from the same angle as the device will be used in the future. The device was therefore mounted in the corner of the room, and ssh was used to access the device remotely from a pc in the acquarium. Then, a script was ran to capture images in sequence, storing them on the SD card in the device. The choice to store the image locally rather than transmitting it was to not have to worry about data transmission costs and issues.

All picamera configurations used in the image capture iterations can be seen in figure TODO. Example images to display the image qualities and differences are found in appendix 9

1st iteration *Total number of images: 1312 (day 1), 986 (day 2) and 641 (day 3), total 2939. 1 subject.*

The first iteration of image capture was made with non-optimized camera configurations. To sufficiently brighten the images, the picamera.brightness attribute was set to 65. This is a post-processing operation, which gave brighter but also artificially lit images. Also, the camera would sometimes focus on the bright fish-tanks in the museus, 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 may be used to inspect the impact of captured image quality on inference performance.

The images were then used to build a proof of concept for the project pipeline, verifying and developing the steps needed for a successful project. The following steps in the project pipeline are described after the description of the 2nd and 3rd iterations of image captures.

Due to many technical difficulties the first few times images were being captured for the dataset, only the developer and author of this thesis is present in the images[12].

2nd iteration *Total number of images: 295 (normal camera) and 60 (no infrared filter camera), total 355. 4 subjects.*

For the second image-shooting 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, as pointed out by TODO insert mikkels master, 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).

TODO send til mikkel når internett:

Hurtigt spørgsmål om masteren din... Jeg skriver måske nogle som det her i oppgaven min: "Also, as pointed out by TODO insert mikkels master, 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)." Hvor fikk du dine ground truth data ifra? Vil det ikke stemme at man uansett vil behøve billeder hvor man selv kan verifisere at modellens predictions stæmmer? in any case when developing the model

While there, the raspberry pi noir camera v2.1 was also tested to see if it could capture better images in the rather dark setting of the acquarium. The configurations were tested in the office with all lights turned off and curtains shut. However, the heavily overexposed scene in the office turned really dark in the acquarium as illustrated in figure TODO.

This was discovered, and 60 images were captured using the noir camera and half the exposure speed so the persons in the images were possible to see for the human eye. One of these images is displayed in figure TODO.

Out of the ordinary camera images, some turned out faulty, to which the reasons are unknown. For this iteration of image capturing, this only happened once and the faulty image can be seen in figure TODO.

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 acquarium. 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. The positions had to be corrected for the generation of heat maps.

3rd iteration *Total number of images: TODO. 1 subject.* For the third iteration the camera setup was better prepared. Two main differences since the last iteration was made. The first was to opt for a higher post-processing brightness value and a faster exposure speed to increase the number of images taken. The other was to spend some time beforehand setting up the camera to the

[12]Initially, an attempt was made to pass MQTT messages as a way to initialize image capture so multiple cameras could be deployed in several locations, thus speeding up and simplifying the image capturing process. This was discarded due to technical difficulties related to efficiently stopping the image capturing. For this single-deployment angle and area project, however, the approach with ssh-ing into the device worked fine.

best position before starting image capture. The image capture sessions are heavily influenced by the constraint of not photographing unconsenting individuals, thus limiting the capturing session to the windows where no other persons are in the acquarium. Therefore, spending time setting up the camera and verifying it's positioning is time which could be spent capturing images for the dataset. However, with well-enough time to complete the project, spending some extra time setting everything up correctly can save time in later stages of the process.

The third iteration was made after setting everything up, in the time waiting for a second subject arrive.

4th iteration

The 4th iteration was with the same exact setup as the 3rd iteration, but with 2 subjects instead of 1.

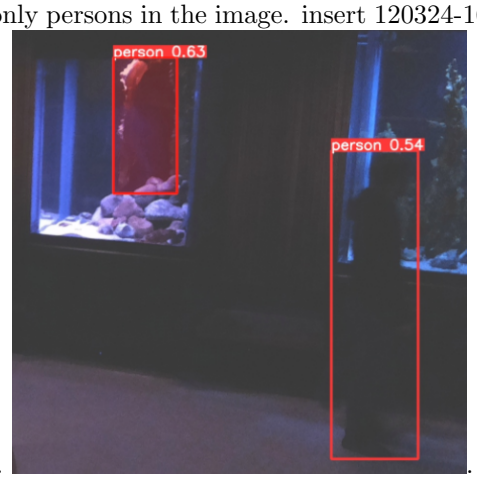
3.2 Labeling

The detector needs to know the ground truth of the correct person positions when training and validating on the images.

The images was first inferenced by a yolov9 detector pre-trained on the COCO dataset. This is a way to greatly speed up the labeling process, instead of manually labeling every image. The detector had close to zero hallucinations due to a sufficiently high confidence rate of 0.5. Out of the 2939 images in the 1st-iteration dataset, 1076 came out with no detections and had to be manually labeled. The rest of the images were then validated. In 74 of the images, a part of one the fish tanks were identified to be human, as it was a moving seaweed that sometimes had a human-like shape. In one image the person was carrying a ladder which was identified to be a person.

The 2nd iteration images were brighter, and in these images one of the roof-lamps were sometimes identified to be human.

In one image, the lamp and the seaweed were the only persons in the image. insert 120324-163917-84. In another, the seaweed was more man than I.



TODO hører hjemme i en recommendations eller discussion? selv-kritiske kommentarer.. I for-studien hadde jeg en "reflections" etter diskusjon som sa litt om hva jeg ville gjort anerledes og forklarte noen av de valgene jeg tok. The time to understand the tools and develop this pipeline was similar to what it would take to manually label all the images, but can now be used for future applications also. The approach to label the images is described in section: 3.2.

After the ground truth was identified for all the dataset images, it was then used to evaluate the pre-trained yolov3 and yolov9. Results are discussed in 4. The data was then used to train the detectors. The training process is described in section 3.3. The trained models were then deployed to the device to evaluate the inference speed and accuracy.

To visualize the improvements and highlight the areas in the image benefitting most from detector

improvements, heat maps were generated. The process of generating heat maps is described in section 3.5.

"Label Studio" was used to label the images.

Labeling speed: about 10 images per minute, when calculating in the time of deleting images without people,

3.3 Model Training

3.3.1 Hyperparameter Tuning

todo finskriv... Not really optimization.. More like finding... Since we're doing a cheeky approach to this. Done with autogluon, follow this guide for installation: https://auto.gluon.ai/stable/install.html. I had to run pip install autogluon twice for the imports to see autogluon.

This guide could be used to fine tune the hyperparameters of the model: https://auto.gluon.ai/scoredebugweight/tutoria
A simpler guide was implemented to find the hyperparameters. This was to save time, and since our models require an okay level of hyperparameters. However, this choice to not give every dataset the same "fighting chance" with their optimal hyperparameters might have led to a lower validity of this experiments results.

Experiment 1: Do we need to run inference on the whole 2939 images to evaluate our model performance or will the results of evaluation on a subset be relative to the results of evaluation on the whole?

The standard yolov3 and yolov9 models were tested on the 1st and 2nd iteration images to see their out-of the box performance. Then, the yolov9 (which performed slightly better) was ran again but with a higher imgsz which typically increases the accuracy of the model. The three inferences: 1) yolov3 with imgsz 1280, 2) yolov9 with imgsz 1280 and 3) yolov9 with imgsz 1280 were then evaluated on the 1st iteration images and 2nd iteration images.

As expected, the inferences using a hightened imgsz was better. Another key takeaway was that the scores were relative, meaning evaluating on the full dataset was unnecessary. Moving forward, the 2nd iteration images from the acquarium were used to evaluate the models, while the 1st iteration images were used for training (and validation). This decision is also motivated by the fact that the 2nd iteration images are the most similar to the real-world application of the models.

We'd like the hyperparameter tuning process to focus on finding the hyperparameters that will best infer on images from the 2nd iteration. Therefore, an amateur would plan to tune the hyperparameters for each of the models using their respective dataset's training data and the 2nd iteration FIMUS dataset for validation data. This means the yolov9 CrowdHuman Smodel would use the same data for validation and testing, but only to find the right hyperparameters. Then, after using the testing data for tuning the hyperparameters, we replace the FIMUS 2nd iteration images for the datasets validation images and train the models using the optimal hyperparameters. This as a method to find the optimal hyperparameters may seem intuitive, as it has the strength of focusing on the right data when doing transfer learning with data other than the specialized task the model is specializing for. This means that the models would, during training, train to infer on data different from what it sees during training, which is what it will be doing after deployment.

The aforementioned approach is completely ignorant to how machine learning works, however. In supervision machine learning, the validation set is used to tune the hyperparameters, and the test set is used to evaluate the model's performance. The test set should be representative of the data the model will see after deployment, and the validation set should be representative of the data the model will see during training. The test set should not be used to tune the hyperparameters, as this would lead to overfitting the model to the test data.

Then, the . Tuning the hyperparameters for the test data means that the validation data used for training later will not at all be representative of the situation earlier, meaning the optimal

hyperparameters will no longer be valid. Such an approach would not

The CrowdHuman dataset has some data where the

What makes this project interesting is not the varying nature of datasets. Usually, one wants the best, most similar, and most data possible, and it's therefore irrelevant to see what using a less relevant dataset would do to the model. In this project the focus is more about how much of the high quality data is needed to make an impact. If 100 labeled images does the same work as 1000, then much work can be saved by only labeling 100 images. If only 100 images is needed then setting up a proper labeling tool might not even be necessary, as it may be done faster using a sub-optimal but fast-to-employ labeling tool. The insights may also be useful when

3.4 Ethical Considerations

In the deployment of advanced machine learning technologies for visitor localization 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.

3.4.1 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.

3.4.2 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 our commitment to minimizing 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.

3.4.3 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.

3.5 Heatmaps

Notes

Tried to download/use model from Roboflow, but either image has to be sent to an API which would not retain privacy, or the device has to host an API itself to run the inference... Seems unlikely to be the most preferable solution, as the device would have to set up the service and run it locally. Possibly an interesting solution would be to do this with multiple devices. This supports the master-slave pattern of having multiple weaker computers and have them send to the stronger unit. Setting up private TCP connection between the weaker units and the strong unit and have the images sent to the stronger, so it can detect on them and send information etc... How many weak units do we need in order to make it profitable to have a strong GPU unit to do the processing? This whole systems sounds to be complicating processes, not making the product modular and easy-to-use. Includes a lot of connection/networking to make the weaker units find and connect to strong, physically close device. This task would mean setting up a strong device to host a network to which the weak units might connect to, and send images to. The issue is whenever images are sent, a lot of transmission is used... But the model takes image input size of 416x416. Would it be similar to just downscale the image before sending, or would this give the model less detail to work with?

Will now run several models on datasets from the web, i.e. the CrowdHuman dataset to see their accuracies. Will then deploy the models to device in acquarium to see if the best-performing model is an option in terms of size and inference speed. If it is preferable, I will attempt to increase it's accuracy by accumulating and annotating a specialized dataset for that setting, and training the final layers on the data. Can this be done with a

4 Results

This chapter will present the results of the human detection and tracking system, including the system's performance in the museum environment, the effects of adding labeled images from the museum environment to the training dataset, and the system's ability to detect and track humans in real-time.

The FIMUS 2nd iteration images were used as the test set for all of the evaluations. This dataset consists of 295 images of similar light condition and image quality. They are the closest representation of the images the device will be capturing in the experimental setting. All images are of 3264x2464 resolution (which is the maximum for the hardware).

The following models were tested:

"Standard models:" Yolov3 Yolov9 using imgsz 640 for inference Yolov9 using imgsz 1280 for inference

"Specialized models:" FIMUS trained 5 epochs FIMUS trained 50 epochs CrowdHuman trained 5 epochs

Model Evaluation

The full model evaluation jupyter notebook can be seen in appendix todo insert model evaluation ipynb.

5 Discussion

This chapter will discuss the implications of the results, summarizing the results and their significance for the development of similar systems, and whether or not the approach in this thesis is a viable solution for the presented problem.

6 Conclusion

Summarization of the thesis and its contributions to the field.

7 Recommendations

If I were to do it all over, what would I do differently and why? What are the key takeaways from this project, and what recommendations do I have for anyone looking to do similar work?

8 Ethical Implications

What are the ethical implications of the development of automated visual tracking?

9 Future Work

Recommendations for future work include the following:

Bibliography

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 pp. 6, 7).

Blum, A., Ligett, K., & Roth, A. (2011). A learning theory approach to non-interactive database privacy. (Cit. on p. 7).

Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., & Zagoruyko, S. (2020). End-to-end object detection with transformers. (Cit. on pp. 15, 16).

Dictionary, O. E. (2023, July). Gamification (n.) https://doi.org/10.1093/OED/7320229446 (cit. on p. 12).

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. 7).

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. 7).

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. 10, 11).

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: date-month-year]. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32022L2555 (cit. on p. 4).

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. 13).

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. 6).

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. 4, 7).

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. 8).

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. 2).

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. 12).

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 p. 13).

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. 10).

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. 16).

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. 5).

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. 3).

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. 13).

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. 7).

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. 4).

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. 3).

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. 12).

Wang, C., Yeh, I.-H., & Liao, H.-Y. M. (2024). Yolov9: Learning what you want to learn using programmable gradient information. (Cit. on p. 16).

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. 4, 9).

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. 7).

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 pp. 5, 6).

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. 9, 12, 13, 15).