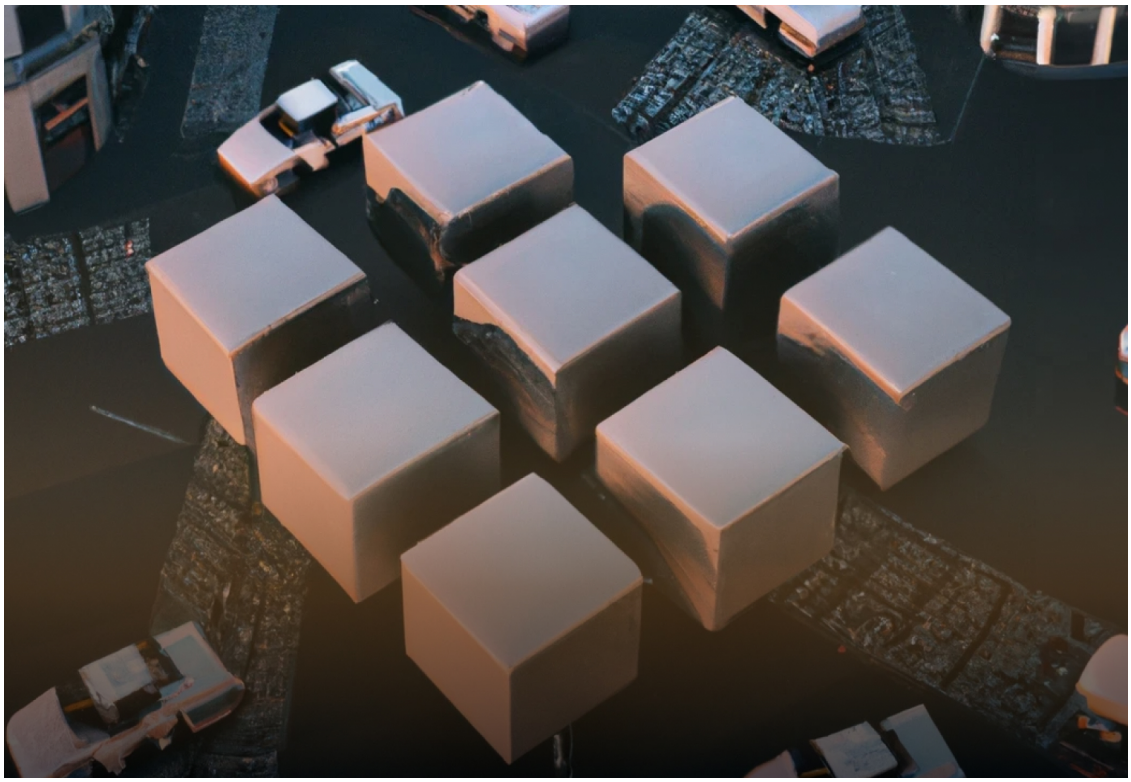
CHALMERS
UNIVERSITY OF TECHNOLOGY



Towards Federated Fleet Learning Leveraging Unannotated Data

Reimagining Autonomous Driving: A Study of Federated Learning Using Unannotated Data

Master's thesis in Data Science & AI

Alexander Viala Bellander, Yazan Ghafir

DEPARTMENT OF ELECTRICAL ENGINEERING

CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden 2023
www.chalmers.se

Master's thesis 2023

Towards Federated Fleet Learning Leveraging Unannotated Data

Reimagining Autonomous Driving: A Study of Federated Learning Using Unannotated Data

ALEXANDER VIALA BELLANDER
YAZAN GHAFIR

CHALMERS
UNIVERSITY OF TECHNOLOGY

Department of Electrical Engineering
Data Science & AI
Chalmers University of Technology
Gothenburg, Sweden 2023

Towards Federated Fleet Learning Leveraging Unannotated Data
Reimagining Autonomous Driving: A Study of Federated Learning Using Unannotated Data
ALEXANDER VIALA BELLANDER
YAZAN GHAFIR

Supervisor: Johan Östman, AI Sweden
Supervisor: Mina Alibeigi, Zenseact
Examiner: Alexandre Graell i Amat, Department of Electrical Engineering

Master's Thesis 2023
Department of Electrical Engineering
Data Science & AI
Chalmers University of Technology
SE-412 96 Gothenburg
Telephone +46 31 772 1000

Cover: A Generative AI generated illustration of a grid of boxes in an urban automotive scene.

Typeset in LaTeX
Printed by Chalmers Reproservice
Gothenburg, Sweden 2023

Towards Federated federated learningeet Learning Leveraging Unannotated Data
Reimagining Autonomous Driving: A Study of Federated Learning Using Unannotated Data
ALEXANDER VIALA BELLANDER
YAZAN GHAFIR
Department of Electrical Engineering Chalmers University of Technology

Abstract

The rapid advancement of autonomous driving technology poses new challenges, including the efficient management and use of large volumes of data generated by autonomous vehicles. federated learning, which allows for distributed, on-device learning, has emerged as a potential solution. However, the effectiveness of federated learning in the context of autonomous driving, particularly when faced with scarce or non-existent labelled data, is still an open question. This thesis explores this issue, employing semi-supervised and imitation learning methodologies within the federated learning framework for autonomous driving tasks such as ego-road segmentation and trajectory prediction. This approach deviates from the conventional assumption of abundant labelled data, aiming instead to maximise on-device learning from unlabelled data. While our experiments demonstrate the potential of federated learning in autonomous driving, results indicate that its performance is currently on par with or slightly less effective than traditional methods for the tasks we studied. Furthermore, this research underscores the largely untapped potential of self-supervised learning methodologies within the federated learning framework for autonomous driving. We argue that further exploration in this area could result in significant breakthroughs and contribute to a future where autonomous vehicles can collectively learn without compromising privacy and efficiency.

Keywords: Autonomous Driving, Federated Learning, Machine Learning, Semi-Supervised Learning, Imitation Learning, Ego-Road Segmentation, Deep Learning, Computer Vision, Trajectory Prediction, Fleet Learning.

Acknowledgements

Collectively, we would like to express our deepest gratitude to our thesis advisors, Dr. Johan Östman and Dr. Mina Alibeigi. Their continuous support, patience, and motivation, combined with their immense knowledge, were invaluable to the completion of this project.

We further extend our sincere appreciation to Zenseact for providing us the opportunity to be part of their research project and for allowing us to immerse ourselves in their profound machine learning culture. Special thanks are due to William Ljungberg, an Industrial PhD Student at Zenseact, for his assistance with our inquiries related to the Zenseact Open Dataset, and to Pavel Lutskov from Zenseact, for his invaluable insights that enhanced our understanding of trajectory prediction.

Our heartfelt appreciation also goes out to AI Sweden for fostering a dynamic community and providing an ideal platform for our thesis work. Their support was crucial to the realisation of this project.

As Yazan, I am deeply grateful to my wife, Razan, whose unwavering encouragement, faith, and cheerfulness lightened my journey. To my parents, whose teachings fostered my perseverance and diligence, my deepest thanks. Their enduring support and guidance have been invaluable.

We are both profoundly thankful for all the guidance, assistance, and opportunities that have been extended to us during the course of this thesis work.

Alexander Viala Bellander & Yazan Ghafir, Gothenburg, June 2023

Contents

1

Introduction

Artificial Intelligence (AI) has emerged as a transformative field in modern technology, enabling the development of intelligent systems that can learn, reason, and adapt to solve complex problems. Machine Learning (ML), a subfield of AI, focuses on the development of algorithms that allow computers to learn from data and make decisions or predictions based on that information. Deep Learning (DL), a subset of ML, utilises artificial neural networks to model complex patterns and solve intricate problems. These techniques have rapidly advanced the capabilities of AI systems in a wide range of applications across various industries.

Road safety is a pressing global concern, with human errors being the root cause for a staggering 94% of accidents, as per the report by the National Highway Traffic Safety Administration (NHTSA) [1]. An effective countermeasure to this crisis presents itself in the form of Automated Driving Systems (ADSs), designed to not only mitigate accident rates but also facilitate numerous additional benefits, ranging from emission reduction to alleviating driving-induced stress [2]. Furthermore, with the potential to yield nearly $800 billion in societal benefits annually by 2050, ADSs may revolutionise the way we comprehend transportation by mitigating congestion, reducing road casualties, conserving energy, and optimising productivity through efficient time management [3].

The advent of AI-powered vehicles holds the promise of overcoming the limitations inherent in human-operated transport systems. This artificial intelligence is immune to common human frailties, such as distraction, fatigue, and flawed judgement, as stated in [4]. The AI-driven transport era would augment safety by reducing accidents and providing instantaneous post-accident responses. The substantial autonomy conferred by these AI systems could drastically improve the mobility of the elderly or individuals with physical impairments, thereby enabling independent travel [4]. The transition towards fully autonomous transportation is a complex and lengthy process, marked by varied predictions concerning its timeline [5]. The integration of fully automated systems is expected to be feasible on a large scale in 15+ years, however, contingent on significant advancements in technology, infrastructural modifications, and comprehensive revisions to existing regulatory frameworks [5].

The use of ML and DL in autonomous driving has had significant contributions

to the field of computer vision (CV). New techniques and methodologies are being developed to address the unique challenges associated with autonomous driving. Some of these challenges include but are not limited to, recognition of objects, mapping, and planning [3], [6], [7]. A new area of research is Federated Learning (FL), a distributed learning approach that enables multiple devices to collaboratively learn an AI model without sharing their data, enhancing the possibility of privacy. This emerging field holds the potential to further enhance the performance and accuracy of AI systems in autonomous driving where data might be challenging to come by, particularly where data privacy is of great concern.

One promising application of federated learning is in driver monitoring systems. Privacy is a critical factor in such systems as they gather sensitive data about the driver's behaviour and characteristics. Using federated learning, a model can be trained to monitor the driver's state without the necessity to share individual driver data with a central server, thereby ensuring privacy. The application of federated learning presents the exciting possibility of collaboration in the automotive industry, a concept where companies cooperate in learning while maintaining their competitive edge. Multiple Original Equipment Manufacturers (OEMs) can collaborate to train a shared model, without revealing their unique and valuable data. A similar concept has already been explored in the pharmaceutical industry for drug discovery, under the Melloddy project, where various companies cooperate in model learning while keeping their sensitive data private [8].

Federated learning works by allowing individual devices to perform machine learning tasks on their local data while periodically aggregating the results on a central server. The central server coordinates the learning process and ensures that the individual devices contribute to a joint global model. This approach allows the devices to learn collaboratively without having to share their sensitive data. While the consensus among researchers is that federated learning generally performs slightly worse in contrast to centralised learning on the same available data, the gains in additional security and privacy can be worth more [9].

federated learning presents potential applications in the field of autonomous Driving, extending beyond addressing data privacy concerns. One of its significant benefits is the potential to reduce costs associated with data transmission and storage. By allowing local training of machine learning models on individual devices, federated learning minimises the need to send data to a central server. Furthermore, this methodology can be applied across different devices or data sources. For example, vertical federated learning enables collaboration among multiple organisations with unique datasets on a shared task [10]. Cross-silo and cross-device federated learning can facilitate collaboration among devices across various organisational units or device types by allowing access to more data [11].

The availability of labelled data forms the backbone of successful machine learning and deep learning models [12]. In the context of autonomous vehicles, labelled datasets provide the necessary information for the models to learn, such as distinguishing pedestrians from other vehicles, identifying road signs, understanding

lane markings, and more. However, creating these labelled datasets is often labor-intensive and time-consuming, demanding a precise annotation of countless images and video frames. Furthermore, data labelling is subject to privacy and legal constraints, particularly when dealing with data that captures public spaces or private information. This makes the process of data acquisition, labelling, and usage for training autonomous driving systems particularly challenging. Federated learning, by retaining data at the source, might help circumvent some of these issues by training models across multiple data owners without the need for explicit data sharing [11].

1.1 Research questions

In this thesis, we explore the potential and effectiveness of federated learning within the realm of autonomous driving, particularly in relation to computer vision tasks. We pose the question: can federated learning outperform traditional centralised learning methods in this context? Further, we inquire into the feasibility of applying federated learning to this field, with particular attention to situations where labelled data are either scarce or absent due to privacy constraints. Within these constraints, can unsupervised or semi-supervised learning methodologies, employed within a federated learning framework, provide viable solutions?

1.2 Approach

To achieve our research objectives, we propose simulating and conducting a comparative analysis of federated learning and centralised learning methods on various known autonomous driving computer vision tasks. We simulate the federated learning approach using open-source tools and evaluate the performance of the federated learning models against traditional centralised models in addition to centralised training in isolation. We conduct experiments on a real-world open-sourced autonomous driving dataset to evaluate the feasibility of federated learning in a practical setting.

1.3 Expected contribution

Our proposed research can contribute to the growing body of literature on federated learning in autonomous driving computer vision tasks. Specifically, we aim to provide a comprehensive evaluation of the effectiveness of federated learning in autonomous driving-related computer vision tasks. Furthermore, we aim to identify the challenges and limitations of federated learning in this context, providing insights into potential areas of improvement and future research directions. Our work can also inform the development of innovative techniques for distributed machine learning, improving privacy, efficiency and applicability in unlabelled contexts.

1.4 Report structure and overview

This report will be organised into several chapters, each addressing a specific topic related to the research question. Chapter 2 will provide a comprehensive review of the relevant literature on federated learning and autonomous Driving. Chapter 3 will focus on trajectory prediction in autonomous Driving, describing the relevant theoretical concepts and techniques employed in solving the problem, presenting our approach using federated learning, and discussing the experimental results and analysis. Chapter 4 will focus on road prediction in autonomous Driving, following a similar structure to Chapter 3. Finally, Chapter 5 will provide a discussion of the results obtained and draw conclusions from the experiments conducted.

2

Theory

In this chapter, we lay down the fundamental theoretical frameworks which underpin the methodologies adopted in our research. We briefly touch upon traditional, centralised machine learning in section 2.1 before progressing to the more complex and novel paradigm of federated learning in section 2.2. The latter has particularly profound implications for privacy-preserving applications such as autonomous driving, a facet which we will discuss in section 2.6.

2.1 Traditional machine learning

Machine learning, in its most traditional form, is a subfield of artificial intelligence that leverages statistical techniques to enable computers to improve at solving a task from experience. It involves training models using a large volume of data in order to infer the underlying patterns and correlations [13]. This approach is centralised in nature as it typically requires the aggregation of data from different sources at a single location for the training process.

The foundational algorithmic process of traditional machine learning involves a *model* learning a function that maps an input to an output based on example input-output pairs. These pairs constitute a dataset, where the model is trained on a subset (training set), and its performance is evaluated on the remaining data (test set). This process involves the iterative minimisation of a loss function, which quantifies the discrepancy between the model's predictions and the actual outputs [13].

However, despite its efficacy, this traditional approach raises issues related to privacy, data ownership, and transfer costs when data is to be collected from geographically distributed sources or in settings where privacy is paramount.

2.2 federated learning

Introduced by McMahan et al. in 2016 [14], federated learning provides an innovative solution to training AI models, allowing data confidentiality and processing at the source. It has emerged as a compelling approach to address the evolving regulations

around data privacy and handling. FL's unique ability to tap directly into the data stream from sensors on satellites, infrastructural components, machinery, and an increasing variety of smart devices positions it as a significant enabler for harnessing the potential of AI applications [15].

2.2.1 Introduction to federated learning

In this era where data privacy regulations are increasingly stringent, federated learning emerges as a revolutionary technique in machine learning. Federated learning allows multiple parties to collaboratively train a deep neural network without direct data sharing [9], [16]. The significance of this privacy-preserving technology is particularly evident in the realm of autonomous driving, where data privacy is paramount [17].

Federated learning can be tailored to suit two types of applications: vertical and horizontal [9]. In horizontal applications, disparate datasets share common features, providing an avenue for wider collaborative learning. Conversely, vertical applications leverage federated learning to enrich existing datasets, where additional data on the same entities can be integrated. This versatile approach allows for federated learning implementations across a variety of infrastructures, ranging from Internet of Things (IoT) devices to expansive data centres. Moreover, its applicability extends to different contexts such as cross-silo, cross-organisational, and even across geographical and jurisdictional boundaries, underlining its immense potential in the realm of distributed learning [9], [15].

The conventional centralised learning model, while effective, is fraught with significant data transfer overheads and potential violations of data privacy laws. The inherent design of FL, which promotes the sharing of machine learning models rather than the raw data, ensures privacy and adherence to regulations [9].

2.2.2 Advantages of federated learning

FL presents multiple advantages over the traditional approach, especially in terms of data confidentiality, bandwidth, and storage optimisation. The transfer of knowledge, rather than raw data, to the central server eliminates the need for massive data transfers, thereby conserving local device resources while preserving data insights [18]. This pivotal attribute aligns FL closely with the principle of data minimisation, a cornerstone of many global data protection and privacy laws. It is noteworthy that the compliance of Federated Learning with the General Data Protection Regulation (GDPR) is currently under investigation by relevant authorities, such as The Swedish Authority for Privacy Protection (IMY) [19].

The concept of data minimisation dictates that organisations should restrict the collection, storage, and usage of personal data to the bare minimum necessary for specific tasks or processes. This principle is notably embodied in the European Union's General Data Protection Regulation (GDPR) [20]. Therefore, FL stands out as an embodiment of this principle, offering a compelling solution that minimises the

need for centralisation of data during processing, thus further bolstering its position in privacy-preserving machine learning.

As that the raw data remains confined to the original device, federated learning minimises the exposure of sensitive information during the learning process, thereby mitigating the risks associated with data breaches. This is further fortified by the implementation of privacy-enhancing technologies such as differential privacy and secure aggregation [21]. Since the data remains on the local device, there's no need for large-scale data transfers, which can be both costly and potentially risky. This also means that data doesn't have to be stored centrally, which can reduce storage costs and further lower the risk of data breaches and the risk of having a single point of failure [9].

In some industries and regions, data privacy regulations such as the GDPR or Health Insurance Portability and Accountability Act (HIPAA) require data to be handled in specific ways. Federated learning can help organisations comply with these regulations by keeping data decentralised and limiting data processing to what's strictly necessary [9], [20]. Federated learning presents a unique opportunity to utilise more data for training purposes, even amongst rival OEMs. This may significantly enhance autonomous driving solutions and accelerates progress towards full autonomy.

Further, FL leverages distributed data and computing resources available at edge devices, a key feature for autonomous driving where each vehicle functions as an edge device. Consequently, a larger fleet can contribute to the learning process, potentially enhancing ML/AI application performance in less time and facilitating quicker development of advanced applications [16], [18]. FL also encourages data sharing among different car OEMs and transport service providers securely, thereby optimising road usage, enhancing delivery speed, reducing traffic congestion, and ensuring safer transportation [17].

2.2.3 Challenges of federated learning

Despite its noteworthy advantages, especially in sensitive data contexts where centralisation is infeasible, federated learning also has inherent challenges. The performance of FL models are often inferior to that of their centralised counterpart, primarily due to the distributed and potentially heterogeneous nature of the data. This leads to a delicate balancing act between preserving privacy and achieving optimal model accuracy [9].

Moreover, federated learning entails a unique set of network-related challenges. The very structure of a federated network, with its diverse and distributed nature, may introduce vulnerabilities to different types of attacks, such as model poisoning or data inference attacks [22]. Given the relative novelty of this research field, not all possible threats have been thoroughly examined or addressed.

One crucial challenge in the FL setup lies in discerning between beneficial and detrimental clients. Identifying the contributory significance of a client is non-trivial,

particularly as direct inspection of client data is typically not possible in order to maintain privacy. Therefore, devising robust and private methods for assessing client contribution and identifying potential adversaries in an FL setting remain open areas of investigation in this burgeoning field [23].

Specifically in autonomous driving, FL faces distinct challenges. It requires in-vehicle generation of ground truth data to enhance the performance of machine learning and deep learning models. Potential solutions such as self-supervised learning, which doesn't rely on labels, have shown promise, although their current performance has not yet reached parity with supervised approaches that use human-labelled data [24], [25]. In addition, semi-supervised learning represents another approach to tackle these challenges. This learning paradigm leverages both labelled and unlabelled data, which can be particularly beneficial in autonomous driving where acquiring labelled data may be costly or impractical. However, like self-supervised learning, semi-supervised techniques are still under active development to improve their performance in comparison to fully supervised methods [26].

Despite these challenges, the potential benefits of FL make it an exciting area of research, especially in the development of technologies for autonomous vehicles. Its ability to balance privacy and utility could drive significant advancements in the field.

2.2.4 Federated learning workflow

The FL workflow starts with an orchestrator (server) distributing a global model to the federation of training nodes (clients). These clients further train this model using their local data and intermittently submit their locally trained models to the central server for aggregation. The server then synthesises these models into a consensus model, which is then sent back to the clients for continued training. This process, known as a global round, persists until the model attains satisfactory performance see fig. 2.1 [9], [14], [16].

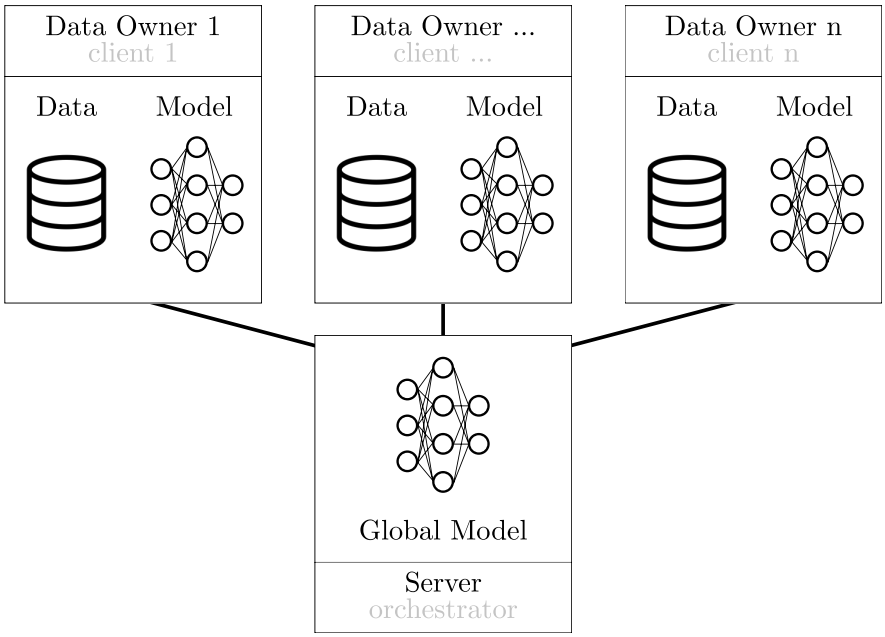


Figure 2.1: The following illustration is a depiction of the procedure in federated learning where data owners, termed as clients, use their own data to train local models. The central server then collects these models and forms a consolidated global model. This global model is subsequently returned to the clients, allowing further refinement on their local data. This iterative process forms the essence of federated learning, fostering cooperative learning while maintaining data privacy.

The Federated Averaging (FedAvg) algorithm is the most common as it was introduced by the original federated learning paper [14]. The algorithm, provided in algorithm 1 incorporates several variables that control its functionality and affect its performance. Among them, K represents the total number of clients involved in the federated learning process. The learning procedure takes place independently on each client, identified by the index k. The local minibatch size, denoted by B, signifies the number of training samples that a client processes during one learning iteration. The number of local epochs, symbolised by E, describes the number of times the learning algorithm passes through a client's local dataset.

Another crucial component is the learning rate, η, a hyperparameter controlling the step size at each iteration while minimising the loss function. The model weights, represented by w, are learned during the training process. The server initialises these weights, and they are updated after each round of the algorithm.

The algorithm undergoes multiple rounds until convergence, with t standing for the current round. In each round, m clients are randomly chosen to participate in the learning process. The selection count, m, is determined as the maximum of $C \cdot K$ and 1, where C is a fraction of the total clients. $\mathcal{S}_t$ symbolises the set of clients randomly chosen to participate in round t.

For each client k, n_k is the number of local data points, and P_k represents the entire set of local data points.

Each of these variables plays an essential role in the FedAvg algorithm, shaping the learning process and ultimately impacting the algorithm's overall efficacy.

Algorithm 1 FederatedAveraging (FedAvg). The K clients are indexed by k; B is the local minibatch size, E is the number of local epochs, and η is the learning rate.

Server executes:
initialise w_0
for each round $t = 1, 2, \dots$ **do**
  $m \leftarrow \max(C \cdot K, 1)$
  $S_t \leftarrow$ (random set of m clients)
  for each client $k \in S_t$ in parallel **do**
    $w_{t+1}^k \leftarrow \text{ClientUpdate}(k, w_t)$
  end for
  $m_t \leftarrow \sum_{k \in S_t} n_k$
  $w_{t+1} \leftarrow \sum_{k \in S_t} \frac{n_k}{m_t} w_{t+1}^k$
end for

function ClientUpdate(k, w)
  $B \leftarrow$ (split P_k into batches of size B)
  for each local epoch i from 1 to E **do**
    for each batch $b \in B$ **do**
      $w \leftarrow w - \eta \nabla \ell(w; b)$
    end for
  end for
  return w to server
end function

2.3 Exponential Moving Average

Exponential Moving Average (EMA) is a concept used widely in fields such as financial analysis and signal processing. More recently, EMA has also found applications in the domain of machine learning, specifically in the training of deep learning models [27].

The EMA at time t can be defined as follows:

$$EMA_t = (1 - \alpha) \cdot EMA_{t-1} + \alpha \cdot X_t \tag{2.1}$$

In the equation 2.1, EMA_t is the exponential moving average at time t, EMA_{t-1} is the exponential moving average at the previous time step $t - 1$, X_t is the raw value at time t, and α is a smoothing factor that typically ranges between 0 and 1.

The factor α determines the degree of weight decrease for older observations, i.e., the higher α, the more weight is given to recent observations. This allows the EMA to be more sensitive to recent changes in the data. In the context of machine learning, this sensitivity can help the model adjust more quickly to changes in the underlying patterns of the data.

During the training of a deep learning model, it is typical to update the model parameters iteratively based on a loss function. The standard approach involves directly using these updated model parameters for validation and testing. However, this can lead to issues such as fluctuations in the performance metrics of the model, which happens especially in the context of large-scale models or when training with stochastic optimisation algorithms like stochastic gradient descent (SGD). These issues arise from the inherent noise in the parameter updates, which can lead to high variance in the model performance [28].

EMA offers a solution to this problem. In the context of machine learning, the idea is to create a shadow model, which is an EMA of the original model parameters. The shadow model maintains an EMA of the trained model's parameters. The key advantage of this approach is that the EMA model smoothens the noise in the model updates, providing a more stable and reliable estimate of the model parameters for validation and testing [27].

The EMA model is typically not used for training. It only observes and follows the training of the original model. The EMA model is usually updated after each step of the training, and the degree of smoothing is controlled by a decay factor. This decay factor determines the weights given to the recent observations, allowing the model to adapt quickly to the most recent changes in the model parameters [28]. This technique has been found to be particularly effective in improving the generalisation performance of the model. This is because the EMA model is less sensitive to the specific conditions during the training, such as the random initialisation and the particular minibatches sampled during the training [28].

2.4 Practical AI constraints in AD

Autonomous driving is a rapidly advancing field, but it is also one fraught with numerous challenges and constraints, many of which arise from the practical application of AI techniques [29].

One of the most notable constraints is the issue of data privacy. Autonomous vehicles generate an enormous amount of data, some of which may be sensitive or personal. As such, it's crucial to ensure that data privacy is upheld, even as AI models are trained on this data. This issue has led to the increased interest in federated learning, which allow AI models to be trained on decentralised data, thereby maintaining privacy [16], [30].

Another significant constraint is the need for high-performance computing resources.

Training AI models, particularly deep learning models, requires substantial computational power. This demand can be a challenge for autonomous vehicles, which may not have the same processing capabilities as a data centre or cloud server. Edge computing solutions, which bring computation closer to the data source, are one potential solution to this problem, but they come with their own set of challenges, such as latency and power consumption [18].

The inherent complexity and unpredictability of the real world also pose constraints on AI in autonomous driving. Unlike in controlled environments, autonomous vehicles must be able to handle a wide range of scenarios, many of which cannot be anticipated ahead of time. Moreover, the AI models must perform reliably and consistently, despite these uncertainties. This requirement necessitates robust, generalisable models that can perform well across a variety of situations [24].

Furthermore, the real-time nature of autonomous driving presents another constraint. Decisions must be made quickly, often in fractions of a second, to ensure safety. This speed requirement can limit the complexity of the AI models that can be used, as more complex models generally require more computation and, therefore, more time [25].

Lastly, the regulatory environment can also present constraints. Autonomous vehicles, and the AI that drives them, must comply with a range of regulations, which can vary widely from one jurisdiction to another. These regulations can influence everything from data collection to model deployment [17].

In conclusion, while AI holds great promise for autonomous driving, it is not without its constraints. Overcoming these challenges will require continued research and innovation, as well as collaboration among stakeholders in the industry.

2.5 Zenseact Open Dataset

Zenseact Open Dataset (ZOD)[31] is a collection of annotated image frames and sequences depicting different road environments, including highways, rural areas, and urban environments. The images were collected by Zenseact developmental vehicles under a variety of weather conditions and times of day, giving the dataset a robust diversity of situations that a self-driving car might encounter.

The images in zenseact open dataset are not just raw captures but also come with a degree of preprocessing. To ensure privacy, the dataset features blurred faces and vehicle license plates. These anonymization steps are crucial in maintaining compliance with data privacy regulations, while not detracting from the dataset's utility for the development of autonomous driving algorithms.

One of the features of zenseact open dataset is that the core frames in the sequences are annotated in the GeoJSON format. This provides additional context and structure to the dataset, aiding in tasks such as object detection and semantic segmentation. These annotations can serve as an essential resource for supervised

learning algorithms.

Zenseact open dataset can be particularly useful in the context of this master's thesis project, as it provides a variety of realistic driving scenarios for testing and validating federated and swarm learning approaches. Given its comprehensive nature and the attention to privacy, zenseact open dataset is a high-quality dataset that can be instrumental in advancing research in autonomous driving.

2.6 Related work

Du et al. provided an exhaustive review on federated learning for the Vehicular Internet of Things (IoT) in *federated learning for Vehicular Internet of Things: Recent Advances and Open Issues*[32]. The authors explored the fundamental principles of FL, its classifications, and recent advancements in the discipline. They also highlighted the potential of FL in fostering emerging vehicular IoT applications, while discussing technical challenges and future research trajectories in this area [32].

Nguyen et al. introduced a decentralised federated learning approach, FADNet, in their study, *Deep federated learning for autonomous driving* [33]. Addressing the single point of failure issue in conventional federated learning scenarios, FADNet employed RGB images to execute a regression task aimed at predicting a vehicle's steering angle. Intriguingly, this peer-to-peer approach demonstrated superior performance compared to traditional *Server Based federated learning* (FL) and centralised learning [33].

Zhang et al. developed an end-to-end steering angle regressor based on a stream of RBG images and trained the regressor via federated learning. The experiments were conducted with a varying number of clients, ranging from 4 to 64. The FL results were benchmarked against centralised ML, where data from edge vehicles were first collected to a single server, and local centralised ML, where baseline models were trained directly on each edge vehicle without any federated model aggregation. Their findings suggested that FL could yield results comparable to central ML [34].

Fantauzzo et al. proposed FedDrive, a semantic segmentation benchmark for training a semantic segmentation model for autonomous vehicles. They focused specifically on the challenges of statistical heterogeneity and domain generalisation [35].

Elbir et al. attempted object detection in an FL setting from LiDAR point clouds. They identified several significant research challenges for FL in vehicular networks, covering both learning difficulties such as data labelling and model training, and communication issues like data rate, reliability, transmission overhead, privacy, and resource management. They also suggested potential future research directions [36].

3

Trajectory Prediction

In this chapter, we will explore the ego-trajectory prediction problem and its significance in the context of autonomous driving. Lane prediction plays a crucial role in the perception module of an autonomous driving system. The prediction module aims to anticipate the future state of the vehicle and its surroundings, enabling the vehicle to operate safely and efficiently in complex traffic scenarios.

To provide a comprehensive understanding of the ego-trajectory prediction problem, this chapter will delve into the relevant theoretical concepts and techniques employed in solving the problem. We will explore the various methods and algorithms used for lane detection and trajectory prediction, highlighting their strengths and limitations. Our approach to solving the ego-trajectory prediction problem will also be presented in this chapter. We will outline the framework and methodology used in our experiments, including the pre-processing techniques, model selection, and evaluation metrics. We will also provide a detailed description of the dataset used in our experiments and the experiments' configuration and we outline our approach to solving the problem using federated learning.

Furthermore, we will discuss the experimental results obtained using our approach for solving the ego-trajectory prediction problem. The results will be analysed and compared with relevant techniques, providing insights into the effectiveness of our approach and its potential for real-world implementation.

3.1 The problem

The trajectory prediction problem in autonomous driving involves predicting the future path of the vehicle based on its current position and the surrounding environment. The goal is to anticipate the future state of the vehicle and its surroundings, enabling the vehicle to operate safely and efficiently in complex traffic scenarios. Trajectory prediction is a crucial component of the perception and planning modules in autonomous driving systems, as it helps the vehicle make informed decisions and take appropriate actions in real-time.

The motivation for trajectory prediction is to enable autonomous driving systems to operate safely and efficiently in a wide range of driving scenarios, such as changing

lanes, merging into traffic, and avoiding obstacles. Accurate and reliable trajectory prediction is necessary for several critical functions, such as route planning, trajectory planning, and collision avoidance [37].

The challenges in trajectory prediction arise due to the complexity and variability of real-world traffic scenarios. The prediction must take into account various factors such as the speed and direction of the vehicle, the surrounding environment, the behaviour of other road users, and potential hazards. Moreover, the prediction must be accurate and timely to ensure the safety of the vehicle and its passengers. Hence, developing effective solutions to the trajectory prediction problem is crucial for the development of safe and efficient autonomous driving systems.

3.1.1 The single-trajectory predictor

In certain scenarios, trajectory prediction can become ambiguous, particularly when multiple possible and plausible trajectories coexist. Consider a car at a standstill in a four-way intersection; in this situation, an ideal trajectory predictor would present various plausible solutions, such as turning left, continuing straight, or turning right. The complexity of the problem increases with the addition of multiple lanes on each road. Furthermore, an even more intricate interpretation of the issue might incorporate u-turn trajectories or escape/evasive manoeuvres, see fig. 3.7.

In this study, we employ a simplified version of the problem, the single-trajectory predictor, to test whether federated learning can be applied and beneficial. We assume that a single, potentially dominant trajectory exists. To establish ground truth, we use the vehicle's trajectory data and assume that the future trajectory of the vehicle can be predicted based on a single image frame. We utilise this data because it inherently enables self-supervised learning. While the core problem remains supervised, similar to many other self-supervised tasks, the annotations are automatically generated, which is characteristic of self-supervised learning.

As authors, we recognise several limitations inherent to this approach. We discuss these issues and their implications on the results and any conclusions drawn from them in section 3.5.

3.2 Relevant theory & related work

Planning and decision-making strategies for autonomous vehicles (AVs) predominantly fall into three categories [38]: traditional sequential planning, end-to-end control, and end-to-end planning. Traditional sequential planning, despite its common use, faces challenges due to the complex interdependencies [39] of individual components, intricate cost function design, and susceptibility to error propagation, as demonstrated by the Tesla accident in 2016 [3].

Alternatively, end-to-end methodologies aim to address these issues by providing a direct link between sensory input and actuator output, thereby reducing the risk of

error propagation. However, they also introduce new challenges, including ensuring safety guarantees and the need for abundant training data.

Imitation learning, a machine learning subfield, has found application in autonomous driving, where algorithms learn from human driving data [40]. This approach is advantageous as it can handle a wide array of driving scenarios without explicit programming for each potential situation. However, it is dependent on the quality and volume of the training data and may struggle with scenarios not well-represented in the training set.

Each of these methodologies offers potential solutions to autonomous driving challenges, but there is no universal approach. The choice of method depends on various factors, including specific driving task requirements, available computational resources, and the quality and quantity of training data. Our research contributes to this discussion by further exploring ego-trajectory prediction in autonomous driving using imitation learning.

3.2.1 Ego-trajectory prediction

Most literature has focused on the non-ego aspect of trajectory prediction, that is predicting the trajectory of surrounding objects such as other cars, pedestrians, cyclists, etc. That list is extensive and focuses a lot on higher-level orchestration using Reccurent Neural Networks or Convolutional Neural Networks over situation maps. For a comprehensive review, the reader is referred to [41].

Bergqvist & Rödholm [42] attempted to compute the ego-trajectory from images procured from a front-facing camera. They utilised steering signals and onboard sensor data to construct the ground truth. Their exploration varied from simple models using a single frame in a CNN for prediction, to more complex ones using sequences of 10 images. Moreover, they attempted combinations of CNNs and RNNs such as LSTMs. Their findings suggest that future ego-trajectory prediction is feasible and temporal data consistently enhances the accuracy of predictions. Interestingly, their work revealed that object data significantly improves prediction outcomes compared to image data alone, especially when it included the positions of other vehicles. Their hypothesis suggested that in most cases, predicting a path similar to the vehicles in front is simpler than deriving the path from the image. They utilised an image size of (182×68) and also experimented with 224x224 on a ResNet-18, but observed no significant changes.

Expanding upon Bergqvist & Rödholm's foundational work, Cai et al. [43], [44] used the car's path from a GPS/inertial solution, captured at a frequency of 50 Hz, as the ground truth. They also used images from a front-facing camera. Their goal was to predict the vehicle's path up to 3 seconds into the future using a combination of a CNN and LSTM. They used MobileNet-V2 as the feature extractor and also included the driver's intentions in the training data (i.e., turn-left, turn-right, keep straight). They used an image size of (1247, 380) and conducted an extensive review of different scenarios affecting the image, including various weather conditions

and times of the day. Their overall framework was designed to take as input the front-view image and movement sequences from the past 1.5 seconds along with the intention command, to generate feasible trajectories 3 seconds into the future. They trained the entire network end-to-end using imitation learning. However, they observed that the discrete high-level driving commands are not suitable for more complex road topologies, such as intersections with more than one possible direction.

Khakhlyuk [45], another contributor in this field, employed Zenseact data in his research. He used a MobileNet-V2 architecture as a feature extractor, similar to Cai et al. For the ground truth, he used GNSS/IMU data sampled at 100Hz. This data was downsampled to 17 (xyz) points at distances ranging from 5 to 165 meters. This approach represented a significant extension over previous work, where prediction distance was limited. Thus, his model attempts to predict 17 points, the furthest of which is 165m ahead of the vehicle. For his single frame predictor, he used 4536 samples to train the model. He also demonstrated the relationship between training set size and evaluation metric, suggesting that increasing the dataset size improves performance exponentially, plateauing around 8000 images. Interestingly, he achieved an L1 loss of approximately 20 (m)*. His model used a batch size of 16 for 100k iterations, utilising the Adam optimisation algorithm with a learning rate of 1e-4 and weight decay of 1e-6. All computations were executed with 32-bit precision. Interestingly, Khakhlyuk found that changing the values of batch size, learning rate, and weight decay did not significantly improve validation performance. He also tried several learning rate scheduling strategies, ultimately discovering that a constant learning rate offered the best results.

Kilichenko [46] with a similar contribution to Khakhlyuk used the same target distances for prediction. His model employed the MobileNet-V2 architecture as the feature extractor, following the precedent set by both Cai et al. and Khakhlyuk. Kilichenko found that using an L1 loss function yielded improved results in contrast to RMSE loss, providing another valuable insight for the design of ego-trajectory prediction models and achieved an L1 validation loss of <1 m. In his work, Kilichenko successfully demonstrated the utility of the MobileNet-V2 architecture in extracting useful features for trajectory prediction, while also highlighting the importance of loss function selection in model performance. These findings align with the broader trends observed in the works of Bergqvist & Rödholm, Cai et al., and Khakhlyuk, further solidifying the foundation upon which our research is built.

In essence, these studies depict an evolution in the field of ego-trajectory prediction. Starting from the work of Bergqvist & Rödholm, who demonstrated the feasibility of ego-trajectory prediction from image data and highlighted the importance of object data, to the advanced studies by Cai et al., who combined GPS/inertial solution and driver's intentions with image data for more accurate predictions. Finally, Khakhlyuk's work stands out for extending the prediction distance and revealing the impact of training set size on prediction accuracy.

*After a conversation with the supervisor of the project, we conclude that the plot used shows the sum of the batch loss rather than the mean, thus achieving a loss of aprox. $\frac{20}{16}$, however, they did not publicly disclose the sequences or images used for validation.

3.3 Methodology

3.3.1 Dataset

The methodology for this thesis experiment involves the use of Zenseact's Open Dataset [31], see section 2.5. The dataset includes 100000 frames with a 90/10 train/validation split. Each image is associated with high-precision GNSS/IMU data, specifically the OxTS RT3000 inertial and GNSS navigation system [47]. This navigation system features a six-axis L1/L2 RTK, a 100 Hz frame rate, 0.01m position accuracy, and 0.03° pitch/roll and 0.1° heading accuracy.

The high-precision GNSS/IMU data is logged at 100 Hz and stored as HDF5 files, including UTC timestamp in seconds, WGS84 geographic coordinates (latitude, longitude, and altitude), ECEF Cartesian coordinates, heading, pitch, roll, velocities, accelerations, angular rates, and satellite information. This data can be used as ground truth for training different machine learning models, such as ego-vehicle trajectory prediction. A comprehensive description of the fields, coordinate transformation, and visualisation functionalities can be found in the development kit provided by ZOD [48].

In this experiment, the frames dataset is utilised, with the high-precision GNSS/IMU data serving as the ground truth. It is assumed that the recorded path of the vehicle is among the most plausible based on the associated image.

3.3.2 Ground truth

The GNSS/IMU data is processed to create a coordinate system relative to the vehicle, positioning the GNSS/IMU device at the origin (0,0,0) for a given frame. This transformation is crucial as it enables the prediction of the vehicle's future state by creating a localised reference frame for analysis. To build our ground truth, we sample 17 points from the GNSS/IMU data, which are carefully selected based on their spline distances from the vehicle. This sampling strategy ensures that the selected points adequately capture the vehicle's trajectory while minimising computational complexity.

The target distances used in this experiment is the ordered set (in meters)

$$T = \{5, 10, 15, 20, 25, 30, 35, 40, 50, 60, 70, 80, 95, 110, 125, 145, 165\}, \tag{3.1}$$

chosen based on Kilichenko's (2023) work [46], which made use of a Zenseact proprietary dataset, a portion of which is included in the ZOD dataset.

Given a sequence of points $P_i = (x_i, y_i, z_i)$ with corresponding accumulated distances d_i for $i = 1, 2, \ldots, n$, and a target distance t_j for $j = 1, 2, \ldots, 17$, we want to find the interpolated point $P'_j = (x'_j, y'_j, z'j)$ at the target distance t_j. We can obtain P'_j using linear interpolation between two consecutive points P_k and P_{k+1}, where

$d_k \leq t_j \leq d_{k+1}$:

$$P'_j = P_k + \frac{t_j - d_k}{d_{k+1} - d_k}(P_{k+1} - P_k) = (x'_j, y'_j, z'_j) \tag{3.2}$$

In this equation, P'_j represents the interpolated point at the target distance t_j, and P_k and P_{k+1} are the two consecutive points in the original sequence of points that have accumulated distances d_k and d_{k+1}, respectively. The interpolation is performed separately for each coordinate (x, y, z).

3.3.3 Data processing

In this subsection, we delineate the supplementary processing and modification procedures applied to the dataset in order to adapt it to the task at hand. Specifically, the task involves training a model to predict the vehicle's future state or trajectory successfully. Our primary dataset for this purpose is the Zenseact Open Dataset, each frame of which is associated with GNSS/IMU data that we leverage.

To ensure the accuracy of our predictions, we adhere to certain mathematical prerequisites. Given the ordered set provided in eq. (3.1), eq. (3.2) mandates the existence of a t_j such that $d_k \leq t_j \leq d_{k+1}$. Thus, it is required that $\exists d_i, \exists d_j : d_i \leq 0$ and $d_j \geq 165$. This requirement implies that we cannot interpolate a single point and, in situations where interpolation is unnecessary, the accumulated distance must be present within the target. This necessitates the elimination of certain samples from the experiment which do not meet these criteria. From the Zenseact Open Dataset, we identified and removed 17705 such samples.

Another critical aspect of our data processing involves ensuring balance within the dataset. To achieve this, we have devised an automated method to categorise vehicle turns into three distinct classes: left turn, straight, and right turn. The vehicle's trajectory is analysed with its initial position set at the origin (the frame's position). We then project two planes, one on each side of the vehicle along the y-axis at $\pm 10m$, and extend them along the x-axis. By counting the number of points appearing in each volumetric segment (left, middle, right), we can classify the vehicle's motion. A threshold of 5 points outside the middle volume is used to determine the categorisation. Our findings suggest that the train split has the following categorical distribution (11733, 47426, 13098).

In order to address the imbalance in the categorical distribution, a resampling strategy was employed to ensure equal representation from each category within our dataset. This strategy involved selecting n samples from each category, with n being equivalent to the size of the smallest category in the dataset as

$$n = \min(|\text{left-turns}|, |\text{straights}|, |\text{right-turns}|). \tag{3.3}$$

The result of this resampling strategy is a balanced dataset with uniform representation across categories, as illustrated by the distribution (11733, 11733, 11733). Consequently, the balanced training dataset now comprises 35199 samples in total.

In an additional step to refine the dataset, we created an alternative balanced training set by subtracting 3296 samples. These excluded samples represent the intersection between a subset of Zenseact's proprietary data used by Kilichenko and the Zenseact open dataset that also followed the distance criteria. Even though this intersection accounts for only 13% of Kilichenko's original validation set, it was deemed essential to exclude these validation frames from the dataset before the resampling process. By doing so, we ensure that our new balanced dataset is devoid of any overlap with our subset of Kilichenko's validation set. This strategic decision allows us to leverage this subset for validation purposes, thereby facilitating a more straightforward comparison of our findings with prior work.

In conclusion, the data processing phase plays a pivotal role in shaping the effectiveness of our trajectory prediction model. The modifications implemented, including the removal of non-interpolable samples and the balancing of the training dataset, are integral in refining the input data and ensuring it is conducive to accurate and reliable trajectory predictions. The balanced dataset, in particular, allows for equitable representation of various driving scenarios, left turns, straight paths, and right turns, thereby providing a more holistic portrayal of diverse traffic conditions. This balanced representation minimises inherent biases in the data, thus promoting a model that is not only robust but also highly generalisable across a range of real-world scenarios. Moreover, the careful removal of overlapping validation samples from Kilichenko's dataset further ensures the integrity and independence of our training and validation sets, thus providing a more reliable benchmark for comparison. Overall, these carefully designed enhancements to the data processing stage are meticulously crafted to bolster the model's predictive capabilities and its applicability in real-world autonomous driving systems.

3.3.4 Model architecture

To maintain consistency with Kilichenko's work, the same model architecture, MobileNet-V2, is used for single ego-trajectory prediction. MobileNet-V2 is a highly efficient convolutional neural network (CNN) architecture, designed with the specific aim of mobile and embedded vision applications in mind. The MobileNet-V2 CNN is pre-trained with weights from ImageNet, and the last linear layer is removed. The remaining network is referred to as the *ProjectionLayer*, which projects the high-dimensional feature space into 17 3D Cartesian coordinates.

The network input consists of a 256x256 RGB image, rescaled from an original size of 3848x2168 and normalised using ZOD's mean and standard deviation.

The images are processed by the convolutional backbone and subsequently downsampled to a 1280x6x20 feature map, which is then passed to the ProjectionLayer. A global average pooling generates a feature vector of size 1280, followed by a 0.2 Dropout layer and a linear layer. This linear layer projects the feature vector to the 51 output values, representing the 17 points in each trajectory with 3 coordinates per point, see fig. 3.1.

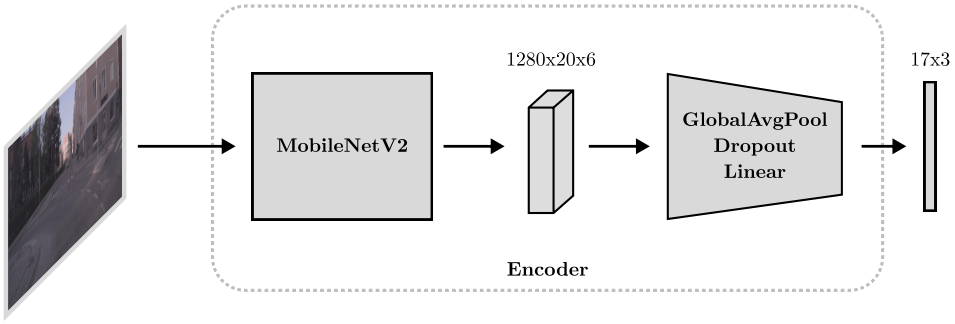


Figure 3.1: Illustration of the network architecture.

3.3.5 Training & evaluation

Our study involves conducting and analysing three types of learning scenarios: centralised learning, federated learning, and isolated centralised learning. The aim of this tripartite experimentation is to gain insights into how each learning paradigm performs under defined conditions and parameters.

In the centralised learning scenario, our model utilises the entirety of our selected training dataset. To enhance generalisation and mitigate any risk of overfitting, we specifically chose our training dataset to be a randomly sampled 5% subset which is 1635 of the balanced frames dataset. We removed Kilichenko's validation frames from the training set prior to this sampling and rebalanced the dataset. The purpose of such segregation is to uphold the integrity of the model evaluation phase, ensuring the validation set is entirely unseen during training. The validation set consists of 3453 images.

The learning process uses a batch size of 8 and employs the Adam optimiser with a learning rate of 1e-4. The model produces a 17x3 vector which is then compared with the ground truth using the L1 loss function

$$L_1(y, \hat{y}) = \frac{1}{n} \sum_{i=1}^{n} |y_i - \hat{y}_i|. \tag{3.4}$$

This function is a robust and commonly used metric in regression problems, owing to its focus on absolute differences between predicted and actual values. Our computational requirements are well met by two Quadro RTX 5000 GPUs, but the setup is also compatible with a single GPU.

The federated learning scenario utilises the Federated Averaging (FedAvg) algorithm (algorithm 1) to aggregate model updates from multiple clients. This scenario comprises five clients ($m = 5$), each holding an equally-sized subset of the centralised training set, specifically 20% of the initial 5%. All clients are involved in each global round of learning, i.e., $C = 1$. The local training on each client follows the same parameters as the centralised learning method in terms of batch size and learning rate.

Over the course of the experiment, we perform 10 global rounds ($t \in \{0, \ldots, 10\}$), providing each local model with 50 epochs of exposure to its individual data subset.

The isolated centralised learning clients have the same datasets as the federated clients and the same learning parameters. The key difference between the methods is that there is no model aggregation and we run the experiment for 100 epochs. Thus, each client performs centralised learning (in isolation) on their individual datasets. We utilise these results as baselines as we expect the federated learning results to exist between the baseline results and the centralised results. We will hereafter refer to this learning setting and any results derived from it as the Baseline.

3.4 Results

Our investigation into the performance of federated learning for autonomous driving tasks unfolds through a series of results and visual data. The first crucial observation can be drawn from fig. 3.2. It represents the federated training and validation loss for clients one to five. One noteworthy characteristic is the emergence of loss spikes at every tenth epoch (10, 20, 30, ..., 100), which correspond to the moments of model aggregation. Such fluctuations suggest that the aggregated model might perform sub-optimally on individual client datasets. Additionally, signs of overfitting become apparent as the validation and training loss curves intersect and diverge. This pattern signifies that the model may be excessively tailored to the training data.

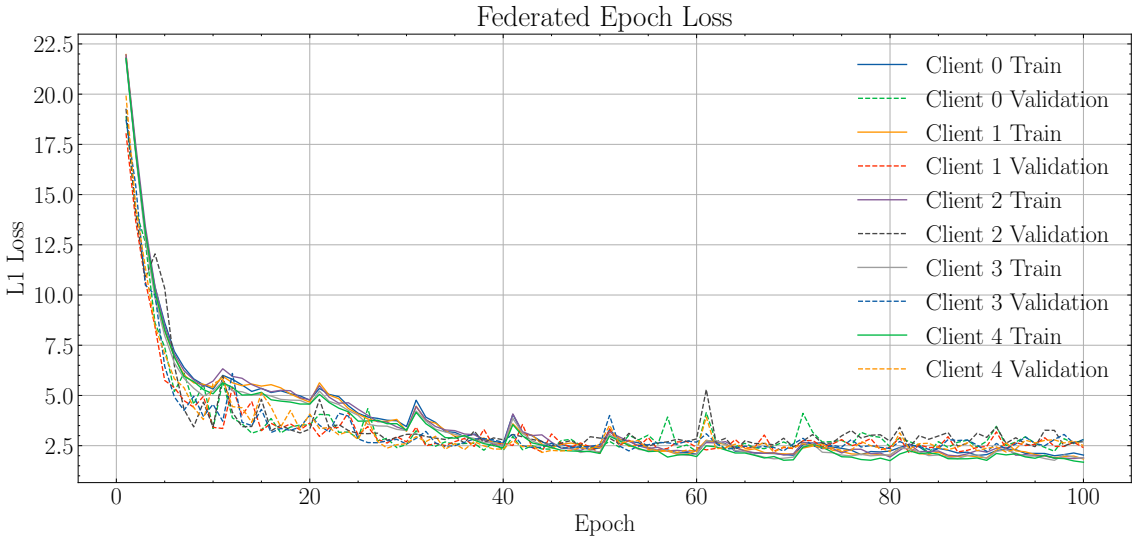


Figure 3.2: Federated training and validation loss for clients one to five. The spikes in loss at every tenth epoch (10, 20, 30, ..., 100), reflect the points of model aggregation. These instances hint that the aggregated model tends to perform sub-optimally on individual client datasets. Additionally, a visible sign of overfitting emerges as the validation and training loss curves intersect and continue diverging, indicating that the model may be excessively tailored to the training data.

The dataset used for our study embraces a variety of environmental contexts. An example of a relatively straightforward scenario is depicted in fig. 3.3. It is characterised by high visual contrast and minimal environmental noise, factors that contribute to its classification as an *easier* example.

In this figure, we provide a multi-perspective visualisation of this scenario, including a bird's-eye view and an elevation view. The bird's-eye view provides a comprehensive depiction of the spatial layout and potential challenges present in the scene. In this case, the road appears largely unobstructed and clearly delineated, further illustrating the less challenging nature of this particular frame.

The elevation view, on the other hand, allows us to appreciate the undulating nature

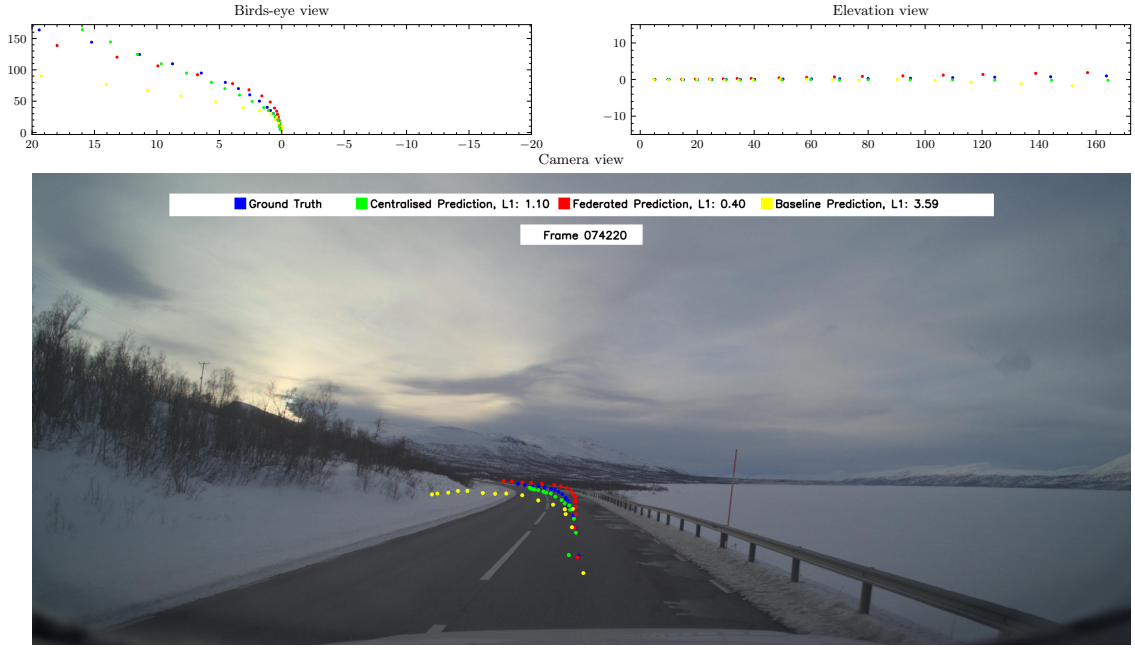


Figure 3.3: This frame presents a relatively straightforward scenario, marked by its high visual contrast and minimal environmental noise. Such conditions contribute to its classification as an *easier* example within the dataset.

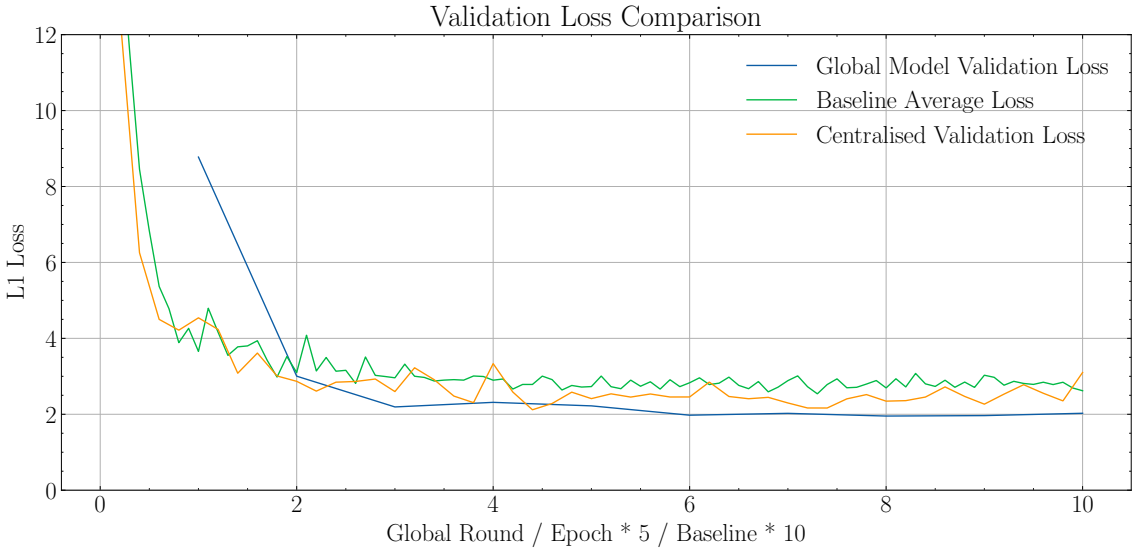


Figure 3.4: The validation losses across different learning settings are depicted on a suitably scaled x-axis to account for the different learning configurations. The federated model's global evaluation commences from the initial global round. Broadly, it appears that the baseline model is the least effective performer. While the minimum validation losses attained by the federated and centralised models are nearly identical, the federated model exhibits a more stable convergence trajectory, suggesting an overall more consistent learning process.

of the road surface, a factor which might introduce complexity in certain scenarios. In the current frame, however, the road appears to be fairly flat, reinforcing the lower complexity of this example.

The prediction results produced by our model for this frame are visually appealing, demonstrating a clear and accurate delineation of the ego-road. This visual appeal is corroborated by the low L1 score for this frame, indicating a high level of agreement between the model's predictions and the ground truth.

Overall, this frame serves as a valuable baseline to assess the performance of our model under less challenging conditions, while also illustrating the multi-faceted nature of our visualisation approach.

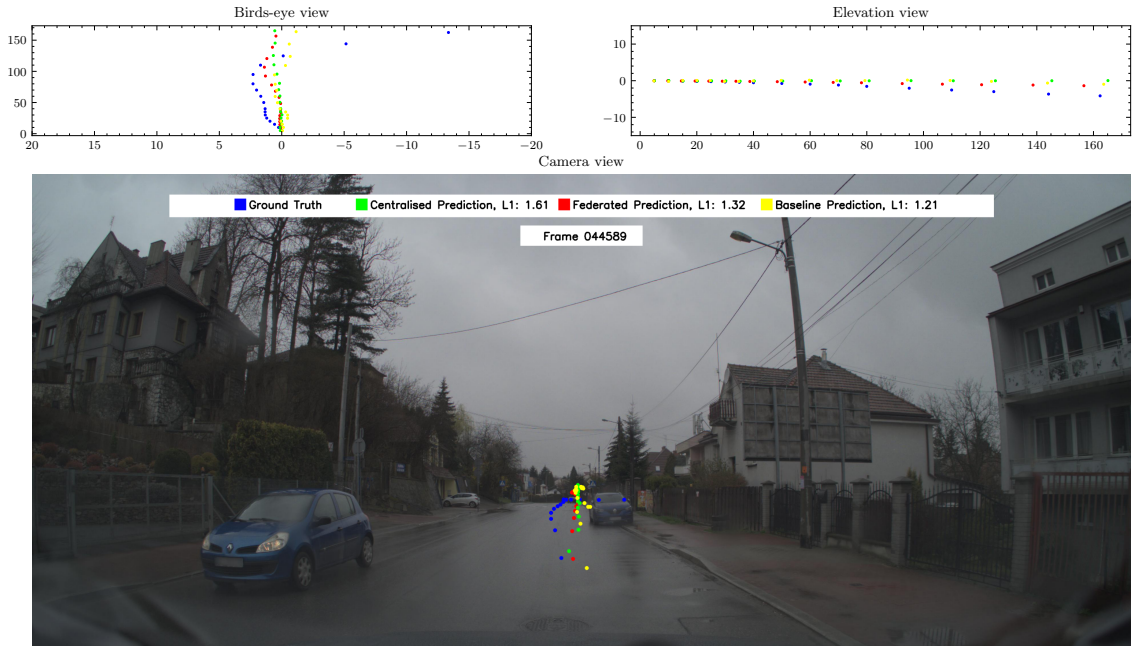


Figure 3.5: This frame showcases one of the more complex scenarios from our dataset. The ground truth points towards an imminent right turn at the end of the road, which is not visually apparent at first glance. Additionally, the presence of a parked car on the right necessitates a slight veer to the left in the driving path - an adjustment that the ground truth also confirms. This depiction underscores the diversity and real-world intricacies that our dataset encapsulates.

We can explore the validation losses across different learning settings in fig. 3.4. The x-axis is suitably scaled to account for the varying learning configurations. Notably, the baseline model appears to be the least effective performer. While the federated and centralised models reach similar minimum validation losses, the federated model exhibits a more stable convergence trajectory. This stability points towards a more consistent learning process, a desirable trait in any learning methodology.

The complexity of real-world conditions is reflected in some frames, such as the one depicted in fig. 3.5. This particular frame showcases an imminent right turn that is not immediately apparent and a parked car on the right that necessitates a slight

veer to the left in the driving path. Both these factors attest to the intricacies the autonomous driving systems need to navigate.

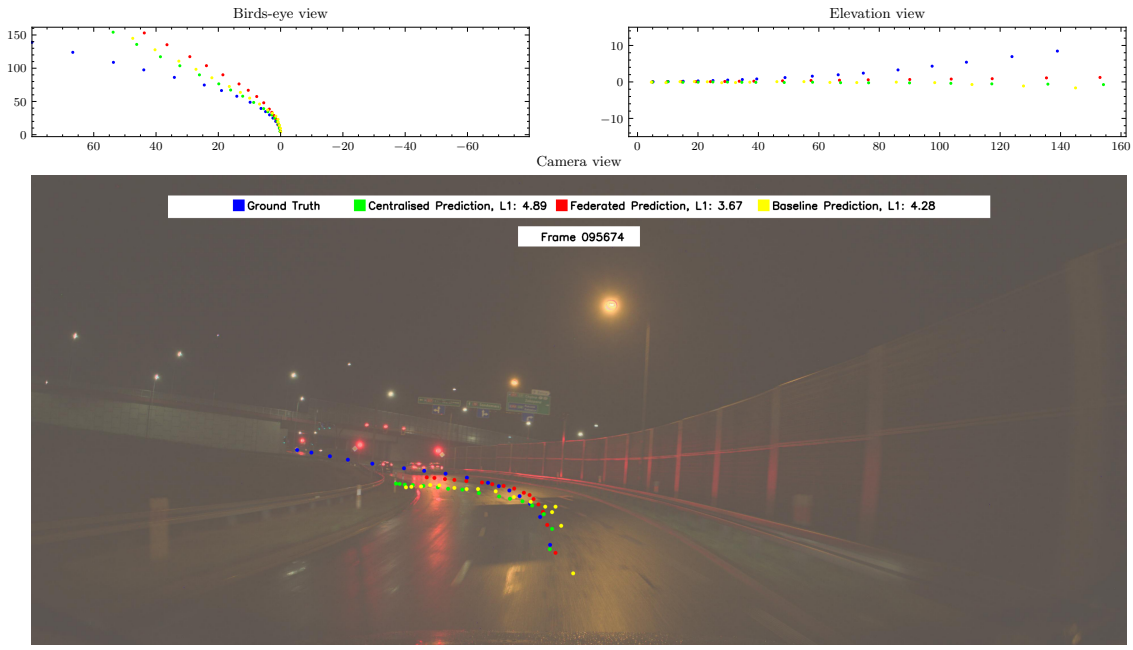


Figure 3.6: At first glance when analysed from a bird's-eye view, the impression is that the centralised method provides a more accurate prediction. However, when considering the elevation (z-axis), the rationale behind the federated method yielding a lower calculated loss than the centralised method becomes clear.

The holistic picture of model performance is not always evident when analysed from a singular perspective. An example is presented in fig. 3.6, where, from a bird's-eye view, the centralised method appears to provide a more accurate prediction. However, the examination of the elevation (z-axis) clarifies why the federated method resulted in a lower calculated loss than the centralised method.

The evaluation metrics of the models we have employed in this study are presented in table 3.1. This table reflects the minimum L1 loss achieved by each model: Centralised Traditional Machine Learning (ML), Federated Learning (Global Model), and Local Centralised Machine Learning (Baseline).

From this tabulation, it's evident that the Federated Learning (Global Model) outperforms the other models with a minimum L1 loss of 1.95. This is a noteworthy result given that the federated learning model had access to less labelled data in comparison to the Centralised Traditional ML model, yet it managed to achieve superior performance.

The Centralised Traditional ML model registers a slightly higher minimum L1 loss at 2.12, suggesting that, while effective, it does not leverage the benefits of a federated learning structure in terms of improving prediction accuracy.

Lastly, the Local Centralised ML (Baseline) model achieves a minimum L1 loss of

2.54. This figure is calculated as the mean value of the lowest validation losses recorded for each individual client. It serves as our baseline metric, representing the performance we could expect from a local machine learning model that doesn't benefit from federated learning or from a centralised aggregation of data.

This tabular presentation of results provides a quantitative comparison of the models' performances, enabling us to better appreciate the benefits of a federated learning approach in this context.

Table 3.1: Model Evaluation Metrics

Model	**min L1**
Centralised Traditional Machine Learning (ML)	2.12
Federated Learning (Global Model)	**1.95**
Local Centralised Machine Learning (Baseline)*	2.54

*This refers to the mean value of the lowest validation losses recorded for each individual client.

3.5 Discussion

3.5.1 The problem design

The single-trajectory prediction model we employ in our study is a simplification of the complex and diverse set of scenarios an autonomous vehicle encounters on the road. By assuming that a single dominant trajectory exists for each image frame, we simplify the problem space to allow for a focused investigation of our methodology's effectiveness. While this approach provides valuable insights and enables the implementation of self-supervised learning, it may overlook certain complexities inherent to real-world driving situations.

3.5.2 Complexity of real-world scenarios

Real-world driving scenarios often involve a multiplicity of potential trajectories, each associated with different probabilities. For instance, at a four-way intersection, a vehicle could proceed straight, turn left, turn right, or even make a U-turn, depending on traffic rules, current traffic conditions, and the driver's intent. Such scenarios defy the single-trajectory assumption and necessitate a multi-modal prediction model capable of predicting a distribution of possible future trajectories.

Moreover, the single-trajectory predictor may struggle to handle emergency situations effectively. In the event of sudden obstacles or abrupt changes in the surrounding environment, a vehicle might need to execute evasive manoeuvres that deviate significantly from the dominant trajectory. An autonomous driving system must be able to anticipate and respond to such contingencies promptly and effectively, a requirement that may not be fully addressed by a single-trajectory prediction model.

Additionally, the presence and behaviour of other road users add another layer of complexity to the trajectory prediction problem. Vehicles do not operate in isolation; instead, they interact continuously with other road users and must adapt their trajectories in response to their actions. The prediction model must therefore consider the trajectories of other vehicles, pedestrians, cyclists, and any other road users, adding to the complexity of the problem.

3.5.3 Limitations and potential transference of findings

While our single-trajectory prediction model simplifies the complexities of the trajectory prediction problem, it is critical to acknowledge its limitations. Our current model lacks the ability to consider the sequential and temporal dynamics inherent in the process of driving. The trajectory prediction task in our model is performed independently for each frame, without carrying forward information from prior frames. This is a significant simplification of the real-world driving scenario, where decisions about the trajectory are heavily influenced by the sequence of events leading up to the current moment.

However, the overarching goal of this work is not to develop a comprehensive solution to the trajectory prediction problem. Rather, we aim to demonstrate the feasibility of federated learning in the context of autonomous driving. In this respect, the single-trajectory prediction task serves as a suitable use case to test and evaluate the application of FL in such a context.

Importantly, we believe that our findings regarding the use of FL have wider implications beyond the simplified task we have chosen to focus on. The lessons learned about the performance, privacy benefits, and implementation challenges of FL in the context of the single-trajectory prediction problem could be potentially transferred to more complex, multi-modal trajectory prediction tasks. Our argument for this transference is based on the understanding that the core challenges of implementing FL, such as dealing with uneven data distribution, privacy preservation, and communication efficiency, are common across different tasks in autonomous driving. These challenges are not unique to the single-trajectory prediction task, and the solutions we develop can inform the application of FL to other problems within the field.

In conclusion, while the single-trajectory prediction task has its limitations, it provides a meaningful context for exploring the potential of federated learning in autonomous driving. The insights gained in this work could inform and guide future efforts to implement FL in more complex and realistic driving scenarios.

3.5.4 The impact of dataset size on the efficacy of federated learning

Our results highlight the potential of federated learning in improving the accuracy of trajectory prediction in autonomous driving systems. Specifically, we found that FL outperforms isolated centralised learning, that is when the models are trained in iso-

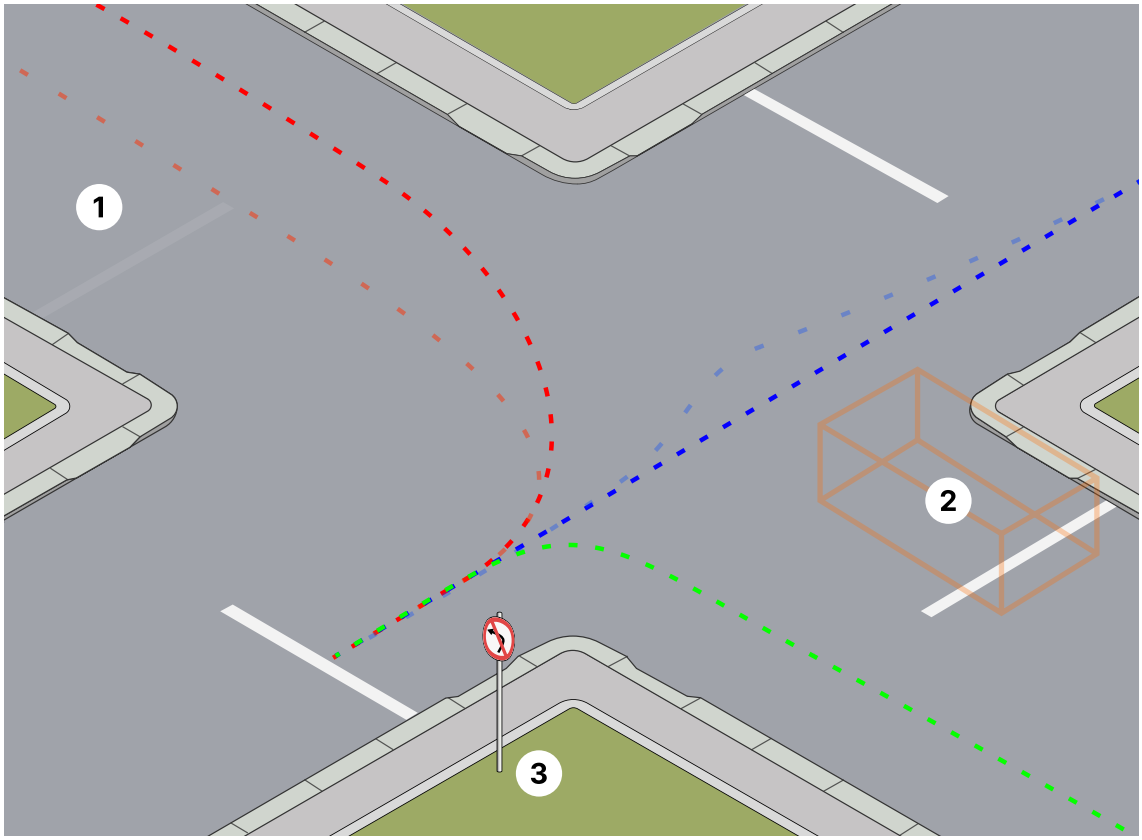


Figure 3.7: Illustration of trajectory ambiguity and complexity at a four-way intersection. Three potential trajectories are outlined: proceeding straight (blue), turning left (red), and turning right (green). Each trajectory corresponds to a specific scenario that highlights the complexities in trajectory prediction. **Scenario 1** presents the issue of a worn-out stop line, which could potentially mislead a prediction model into presuming the road to be one-way. This implies the necessity for comprehensive models, which integrate various contextual cues such as ground signs and road markings, to make accurate predictions. **Scenario 2** encapsulates a challenge inherent in our single-frame prediction setup. The bounding box represents a vehicle emerging from the right, which may not be entirely captured within a single image frame used for prediction. In reality, the trajectory could deviate slightly (faded blue line) to avoid a potential collision with the emerging vehicle, making the straight-line trajectory (blue line) less plausible despite it seeming dominant in the given frame. This scenario underscores the limitation of our setup, where changes between frames are not considered, and emphasises the need for models to account for dynamic elements in the environment. **Scenario 3** signifies the importance of traffic regulations in trajectory prediction. The 'No Left Turns' sign is a crucial piece of information that should be incorporated by comprehensive models to generate accurate and safe predictions. This figure encapsulates the inherent complexities and ambiguities in trajectory prediction, emphasising the need for models that can integrate a wide range of information and context.

lation on the clients' datasets and no model aggregation is performed. This signifies the value of aggregating and sharing knowledge across a federation of autonomous

vehicles, pointing towards a promising avenue for enhancing the performance of onboard prediction systems.

However, it is also pertinent to understand that the effectiveness of FL, like any machine learning paradigm, is influenced by the size of the client dataset and the complexity of the prediction task. Our findings suggest that while FL provides substantial benefits in our current setting with relatively limited client datasets, these gains may diminish if the clients were to have access to larger individual datasets.

This is due to the inherent trade-off in FL between the benefits of shared learning and the computational and communication costs associated with it. As individual datasets grow in size, the unique insights that can be gleaned from them may begin to outweigh the benefits of federated aggregation. Additionally, the results acquired might turn out different for differently sampled subsets of the data, an ablation study or cross validation would be ideal to further strengthen the results of this study.

Moreover, the complexity of the task plays a significant role. For relatively simpler prediction tasks, even smaller datasets might be sufficient to train a high-performing model, reducing the relative advantage of FL. Conversely, for more complex tasks, the collaborative learning approach of FL might continue to provide significant benefits, even as the size of individual datasets increases.

3.5.5 Interpretation of results

Our exploration of the results unveils a substantial variance, likely a mirror reflecting the diverse conditions and scenarios inherent within the frames. Despite this variance, our models exhibit a satisfactory capability to handle both the more accessible and challenging examples effectively. This performance encapsulates our primary goal: to illuminate the potential of federated learning (FL) within this context, as opposed to merely striving for unparalleled performance.

Intriguingly, an overfitting trend surfaced at various junctures during the training process across all models. Although not explicitly highlighted in our results, this phenomenon, coupled with insights derived from other studies, intimates the potential for improved performance. The implication is clear: federated learning could offer a compelling, alternative approach to this task, particularly if the findings from this study can be effectively transferred to future work.

A noteworthy characteristic of the federated model is its stable convergence trajectory, see fig. 3.4, possibly indicative of FL's inherent regularisation properties [49]. However, our models present a limitation in the accurate estimation of the z-axis depth, which could be contributing to some observed inconsistencies. This shortcoming becomes evident when analysing the birds-eye view (x and y axes), where we find paths aligning more closely with the ground truth despite the higher loss. Such inconsistencies, also visible within the frame, become accentuated in downhill

scenarios. Here, minor alterations in degree can exert a significant impact on the loss of points distant from the origin. We posit that the observed z-axis depth issue could arise from our decision to reduce image size or a potential imbalance within the dataset's distribution.

Our approach did not enforce specific balancing in the dataset concerning elevation. Moving forward, it appears plausible that the optimal resolution to this task would benefit from a meticulously curated dataset, balanced across key scenarios. This dataset might encompass diverse parameters such as time of day, road scenarios, occlusion, and more, laying the groundwork for enhanced model performance.

4

Road Segmentation

In this chapter, we will explore the road segmentation problem and its significance in the context of autonomous driving. Road segmentation is an essential task in the perception module of an autonomous driving system, focusing on accurately identifying and segmenting the drivable area from the surrounding environment. This segmentation enables the vehicle to navigate safely and efficiently in complex traffic scenarios by understanding the road's geometry and boundaries.

To provide a comprehensive understanding of the road segmentation problem, this chapter will delve into the relevant theoretical concepts and techniques employed in solving the problem. Additionally, our approach to addressing the road segmentation problem will also be presented in this chapter. We will outline the framework and methodology used in our experiments, including the pre-processing techniques, model selection, and evaluation metrics. A detailed description of the dataset used in our experiments and the experiments' configuration will be provided, and we will outline our approach to solving the problem using federated learning.

Moreover, we will discuss the experimental results obtained using our approach for road segmentation. The results will be analysed and compared with relevant techniques, providing insights into the effectiveness of our approach and its potential for real-world implementation. We will also explore the potential benefits and limitations of employing federated learning in the context of road segmentation.

In conclusion, this chapter will provide a comprehensive overview of the road segmentation problem, emphasising its relevance in the autonomous driving domain. By presenting relevant theoretical concepts, our approach to solving the problem, and the experimental results obtained, we aim to contribute to the development of innovative techniques for enhancing the privacy of autonomous driving systems and the implementation of federated learning in real-world applications.

4.1 The problem

The road segmentation problem in autonomous driving involves accurately identifying and segmenting the drivable area from the surrounding environment using input from various sensors, such as cameras, lidar, and radar. The primary goal is to pro-

vide the vehicle with a clear understanding of the road's geometry and boundaries, enabling it to make informed decisions and take appropriate actions in real time.

The motivation for road segmentation stems from the need to enable autonomous driving systems to operate safely and efficiently in diverse driving scenarios, such as navigating urban environments, winding roads, and various weather conditions. Accurate and reliable road segmentation is necessary for several critical functions, including lane detection, trajectory planning, and obstacle detection [50].

The challenges in road segmentation arise from the complexity and variability of real-world environments. The segmentation process must account for factors such as varying road textures, lighting conditions, and occlusions. Moreover, the segmentation must be robust, accurate, and timely to ensure the safety of the vehicle and its passengers.

Therefore, developing effective solutions to the road segmentation problem is crucial for the advancement of safe and efficient autonomous driving systems.

In the framework of this study, we operate under the principles of federated learning, emphasising conditions where access to abundant labelled data is limited or where privacy concerns impede the conventional data gathering. Although our current problem context benefits from a surplus of labels, we use this situation as a backdrop to methodically explore and evaluate existing self-supervised, unsupervised, and few-shot learning methods. Our aim is to assess and delineate the prerequisites for successful deployment of federated learning schemes, thereby contributing to the broader understanding of these techniques in real-world scenarios where data privacy is paramount.

4.2 Relevant theory & related work

The field of semantic segmentation, and specifically road-segmentation is exhaustive. Long et al. [51] proposed one of the first deep learning works for semantic image segmentation, using a fully convolutional network (FCN). There are additional contributions in the field of semantic segmentation using FCNs. However, more commonly today is to use some kind of encoder-decoder based model [52]. Some key contributions include Noh et al. [53] who introduced an encoder-decoder network using a VGG-16 and deconvolution (a.k.a. transposed convolution). SegNet was proposed by Badrinarayanan et al. [54] with novelty methods in the manner in which the decoder upsamples its lower resolution input feature map.

There are a variety of milestones after SegNet in the field of semantic segmentation [52], while we will not go through the full history of semantic segmentation, we discuss one such milestone. The introduction of U-Net [55], an encoder-decoder architecture proposed by Ronneberger et al. Its unique feature lies in its use of skip connections to preserves spatial information lost during down-sampling. It is particularly useful in situations where there are few examples and was developed

specifically for medical imaging.

The cityscapes dataset [56] released in 2015, published in CVPR in 2016, spawned a series of studies that attempted semantic segmentation for the urban city scene (road included). Some noteable examples: the DeepLab series [57] and Zhao et al.'s Pyramid Scene Parsing Network (PSPNet) [58]. They were evaluated on cityscapes.

In 2017 Chen proposed RBNet for road and road boundary detection [59]. In 2018, Teichmann et al. proposed MultiNet, a multi-task model for segmenting and classifying road types [60]. Teichmann used a variety of ResNet and VGG decoders. The different models showed similar segmentation performance. In 2019, Wang et al. proposed adding a contour map to improve performance [61]. Other types of models attemp at fusing different sensor data, for instance using depth maps [62] (2022) *Fast Road Segmentation via Uncertainty-aware Symmetric Network*, or using LiDAR point clouds together with camera images [63].

In recent years, semi-supervised and unsupervised and self-supervised learning techniques have garnered significant attention as per the advances in the natural language processing field (NLP). Unsupervised representation learning has worked well as shown by GPT [64], [65] and BERT [66] in NLP and notable findings such as SimCLR [67] and MoCo [68] build the foundation of unsupervised visual representation learning. Google subsequently introduced BYOL with online and moving-averaged target networks [69]. These studies showed great promise, however, they were limited to the classification task and used convnets as backbone. In 2021 Facebook published a key paper, Emerging Properties in Self-Supervised Vision Transformers [70] where they introduced DINO. They showed that vision transformer (ViT) features contain explicit information about the semantic segmentation of an image, which does not emerge as clearly with supervised ViTs, nor with convnets.

Recent advances in semantic segmentation often use ViTs as backbone, CutLER [71], a zero-shot unsupervised object detector and instance segmentor showed impressive results in segmenting and detecting prevelant objects in a scene. In our study, we will attempt semi-supervised semantic segmentation using FixMatchSeg [72]. FixMatchSeg extends the FixMatch algorithm [73], originally designed for classification tasks, to semantic segmentation. It makes use of a small labelled dataset and a larger unlabelled dataset. By generating pseudo labels for the unlabelled data and encouraging consistency between the predictions for augmented and non-augmented versions of the same image, FixMatchSeg can effectively learn from both labelled and unlabelled data.

FixMatch incorporates two semi-supervised learning (SSL) methodologies: consistency regularisation and pseudo-labelling. In pseudo-labelling, an initial training phase utilises a small labelled dataset, which subsequently enables the model to predict pseudo-labels for unlabelled data. These pseudo-labels are then incorporated into the training dataset as if they were actual labels, thus enhancing the model's learning scope through iterative retraining on this expanded dataset. Consequently, the model capitalises on the latent information within the unlabelled data, increas-

ing its performance especially when labelled data are sparse. This technique offers significant advantages in scenarios such as autonomous driving, where data labelling proves to be a costly and time-intensive process [72], [73].

In the context of FixMatch, pseudo-labelling is synergistically combined with consistency regularisation. This process entails the creation of two versions of each unlabelled image: a weakly-augmented version and a strongly-augmented counterpart. The weakly-augmented image facilitates the generation of a pseudo-label, whereas the strongly-augmented image serves to calculate the loss relative to this pseudo-label. Through this mechanism, the model is encouraged to maintain consistent predictions across varying augmentations of the same image, thereby enhancing the robustness of the model's predictions against input variations [72], [73].

4.3 Methodology

4.3.1 Dataset

In conducting this experiment, we leveraged the Zenseact's Open Dataset [31]. More specifically, we relied on the frames category of the dataset, which comprises $100k$ frames. The frames selected for this particular experiment were those with associated road labels. Following some additional processing, our dataset comprised of 52199 training frames and 5763 validation frames. In our experiments, we use a randomly selected subset of 5100 training frames and 2500 validation frames. We utilise the same subset of training and validation data for all experiments. For a more in-depth understanding of the ZOD dataset, readers are referred to section 2.5 or [31], [48].

4.3.2 Ground truth

The ground truth employed in this study is the road semantic segmentation mask associated with the ZOD frame. Our target task is a binary pixel-wise classification. Hence, we transform the 2D polygons into a binary segmentation mask that matches the size of the RGB image, effectively categorising each pixel as either road (1) or not road (0).

In contrast to classification tasks, semantic segmentation does not exhibit invariance under geometric transformations such as flips, or affine and elastic distortions that alter the shape or position of objects in the image. Consequently, any transformations applied to the dataset must also be reflected in the ground truth.

The transformations of the ground truths are conducted in separate groups due to the use of the FixMatchSeg algorithm. This algorithm necessitates different types of data, including labelled, unlabelled, and pseudo-labelled data, which all undergo distinct transformations and are generated in varying ways.

Figure 4.1: A ZOD frame with the road segmentation mask overlaid (in red).

4.3.3 FixMatchSeg

In this experiment, the FixMatchSeg algorithm, an extension of the FixMatch algorithm specifically designed for semantic segmentation tasks [72], forms the basis of our methodological approach. This choice is driven by its ability to enhance model performance under conditions of scarce labelled data, while exploiting the abundance of unlabelled data.

To maintain consistency with FixMatch [73] and FixMatchSeg [72], we utilise the same notational convention. A labelled data set of size B is defined as $X = \{(x_b, p_b) : b \in (1, ..., B)\}$, with each x_b representing an image and p_b denoting the corresponding ground truth segmentation mask. An unlabelled batch is characterised as $U = \{u_b : b \in (1, ..., \mu B)\}$, wherein μ is a hyperparameter determining the ratio of unlabelled to labelled data. Consequently, we have B labelled examples and μB unlabelled examples in the training data set.

We denote the predicted class distribution image for the input image x as $p_m(y|x)$. In terms of semantic segmentation loss, we resort to the well-established soft dice loss DL, along with the boundary loss BL. Each of these losses contributes equally ($\lambda = 1$) to the final loss computation.

In congruence with FixMatchSeg and FixMatch, our study employs two variations of image augmentation: weak augmentation α and strong augmentation $\mathcal{A}$. In the context of semantic segmentation, it is critical to note that the output target is not invariant under geometric transformations that alter the image's shape or location. Therefore, we ensure that the same transformation is applied to both the input image and the mask label during geometric transformation-based augmentation.

For weak augmentation α, our selection includes random horizontal flipping applied universally across images, complemented by random affine transformations, which encompass rotation within -10 to 10 degrees, shifting up to 10% of the image size in both horizontal and vertical directions, and scaling between 90% and 110% of the original image size.

For the stronger counterpart, $\mathcal{A}$, we incorporate a variety of techniques. These include a Gaussian blur with a kernel size of 13 and a sigma randomly chosen within the range 0.01 to 2.0, random image sharpness adjustment with a factor of 10 (applied at a 50% probability), random solarisation with a threshold value of 0.5 (also applied with a 50% probability), colour jittering with adjustments of 20% brightness, 70% contrast, 40% saturation, and 50% hue. Additionally, we perform a random inversion of all pixel values in the image, which is applied with a 50% probability. Of note, we deliberately avoid any geometric or shape-changing transformation in the strong augmentation to maintain the identical geometrical shape in both weakly and strongly augmented images.

Our loss computation is twofold: supervised loss l_s and unsupervised loss l_u. The supervised loss is calculated using the soft dice loss DL and boundary loss BL against the labelled images. This calculation is given by:

$$l_s = \frac{1}{B}\sum_{b=1}^{B}(DL(p_b, p_m(y|x_b)) + BL(p_b, p_m(y|x_b))) \tag{4.1}$$

Here, p_b is the ground truth label, while p_m is the predicted mask. A pivotal aspect of both FixMatch and FixMatchSeg is the fusion of consistency regularisation and pseudo-labelling within a single unsupervised loss framework, employing both weak and strong augmentations.

We derive an artificial label from the weakly augmented version $\alpha(u_b)$ for each unlabelled input image u_b. First, we compute $p_m(y|\alpha(u_b))$, which signifies the model's predicted class distribution map for the weakly augmented unlabelled image. Given that we are addressing a binary segmentation problem with $L = 2$ classes, we extract the pixel-wise max of this image to yield q_b, and compute pixel-wise argmax to obtain the pseudo-label, $\hat{q}_b = \text{argmax}[p_m(y|\alpha(u_b))]$. This pseudo-label, $\hat{q}_b$, represents the model's predicted segmentation output for the weakly augmented unlabelled image. To ascertain the inclusion of $\hat{q}_b$ as a pseudo-label, we calculate the prediction confidence score as the mean of the pixel values of q_b, symbolised as $\overline{q_b}$; this reflects the model's average maximum confidence in predicting different classes over the entire image. If this average $\overline{q_b}$ surpasses the confidence threshold τ, it is treated as a pseudo-label and subsequently deemed as ground truth for the strongly augmented unlabelled image $\mathcal{A}(\alpha(u_b))$. See fig. 4.2 for an illustration of how the pusedolabels are selected.

Consequently, the unsupervised loss is defined as:

$$l_u = \frac{1}{\mu B} \sum_{b=1}^{\mu B} \mathbb{1}(\overline{q_b} \geq \tau)(DL(\hat{q}_b, p_m(y|\mathcal{A}(\alpha(u_b))) + BL(\hat{q}_b, p_m(y|\mathcal{A}(\alpha(u_b))))) \quad (4.2)$$

Ultimately, the total loss for our FixMatchSeg implementation, l, is a sum of the supervised loss and the unsupervised loss, weighted by λ_u (i.e., $l = l_s + \lambda_u l_u$).

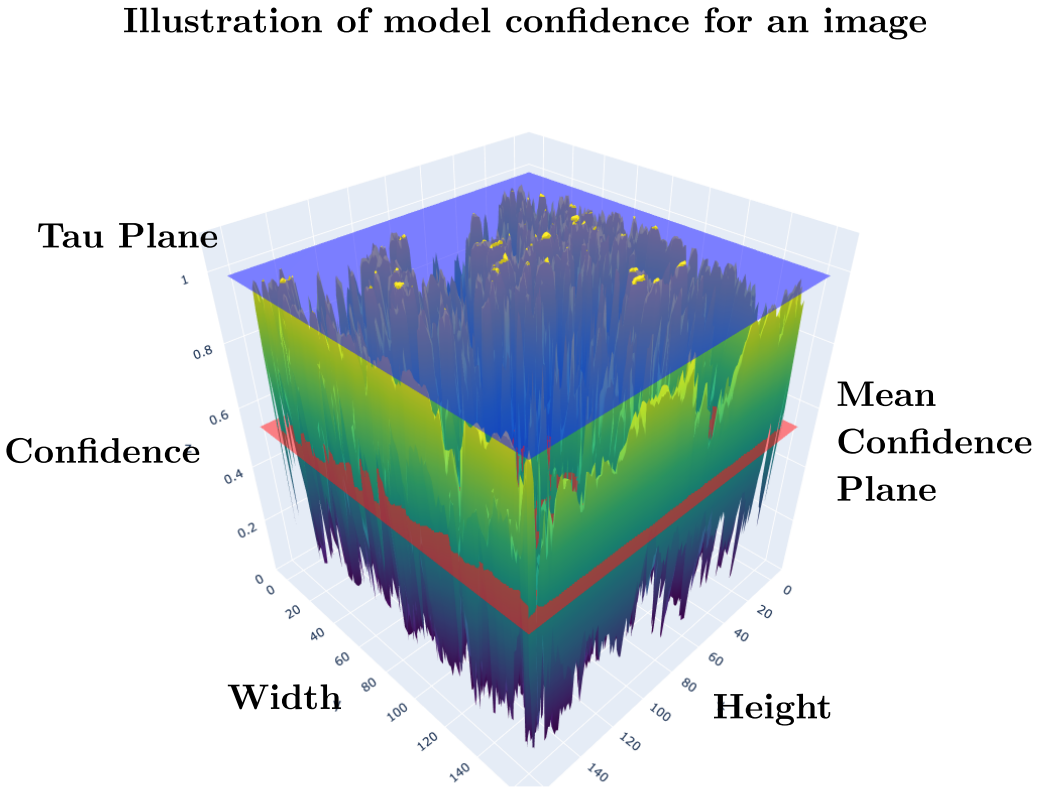


Figure 4.2: Illustration of model confidence for an image. The output of the model is transformed to show the confidence for its predicted class. 1.0 is very confidence in the predicted class, 0 is not confident. What class is not shown here. The τ plane shows our confidence threshold and the mean confidence plane is $\overline{q_b}$.

4.3.4 Training & evaluation

In this study, the model architecture adopted is U-Net, with an EfficientNet-B0 encoder pre-trained on ImageNet*. The model, which accommodates input images with three channels (RGB), produces a one-channel output image that depicts the segmentation predictions for the road.

*The Segmentation Models for PyTorch (SMP) library is employed `https://github.com/qubvel/segmentation_models.pytorch`

For our analysis, a random subset of 5100 training frames and 2500 validation frames is employed. These quantities are arbitrarily selected and are utilised in the execution of various benchmarks.

The **Centralised Supervised Learning** benchmark utilises all 5100 labelled images from the training subset to train the segmentation model. It operates independently of any specific properties of the FixMatch or FixMatchSeg algorithms. The performance of this benchmark serves as the lower bound, where a lower score implies superior performance.

The **Federated FixMatchSeg Learning** benchmark uses 100 labelled images and 5000 unlabelled images. The unlabelled images are randomly and evenly distributed among five clients, granting each client access to 1000 unlabelled images. All clients, however, can access the same set of 100 labelled images.

The **Centralised FixMatchSeg Learning** benchmark uses 100 labelled images and 5000 unlabelled images. This benchmark is anticipated to establish a lower bound (where lower is better) on the potential of FixMatchSeg in this setting.

The **Centralised Local Learning** benchmark employs only 100 labelled images. This benchmark is expected to provide the upper bound (where lower is better), as it does not benefit from any federated aggregation and does not utilise any unlabelled images for improvement.

All the aforementioned experiments operate on the same subset of 5100 images, encompassing the same set of unlabelled and labelled images. Consequently, the 5000 unlabelled images employed in one experiment are identical to those used in the other experiments, and the same applies to the labelled images. For the Centralised Supervised Learning experiment, the 5000 unlabelled images used in the other experiments are also used, but with labels attached. We rescale the images to a size of 160x160 pixels.

$$D = \frac{2 \cdot \sum_{i=1}^{N} y_i \cdot \hat{y}_i}{\sum_{i=1}^{N} y_i + \sum_{i=1}^{N} \hat{y}_i}, \quad \text{Dice Loss} = 1 - D \tag{4.3}$$

Unlike the original FixMatchSeg, we opted to use only the DiceLoss function in our study, see eq. (4.3). This choice was made arbitrarily. The Adam optimiser is applied with a learning rate of $0.5e-3$. A key parameter in FixMatch is μ and τ. In our FixMatchSeg experiments, we set $\mu = 8$, as suggested by a variable batch size of [2,16], where 2 represents the batch size for labelled images, and 16 is the maximum for the unlabelled images. Note that the batch size for unlabelled images is variable and depends on the parameter τ and the confidence level of the model. Non-FixMatch experiments utilise a batch size of 8. For all experiments, the confidence threshold parameter $\tau = 0.99$. The DiceLoss function is used for validation as well.

4.4 Results

The investigation into the road segmentation problem reveals a modest level of complexity that, as per visual assessment, can be resolved satisfactorily. A compelling insight is the convergence of all methodologies, inclusive of the centralised local approach, which leverages a mere set of 100 labelled images, towards similar outcomes. The comparison with other techniques, most prominently the supervised method involving 5100 images, manifests in subtle shifts in their optimum validation performances as indicated by the dashed lines in fig. 4.3 and enumerated in table 4.1. Notably, both Centralised FixMatchSeg and Federated FixMatchSeg demonstrate comparable performances, yielding similar loss levels.

Table 4.1: Model Evaluation

Model (# labelled images, # unlabelled images)	**min DiceLoss**
Centralised Supervised Learning (5100, 0)	**0.136**
Federated FixMatchSeg Learning (100, 5000)	0.156
Centralised FixMatchSeg Learning (100, 5000)	0.151
Centralised Local Learning (100, 0)	0.179

The Centralised FixMatchSeg model, interestingly, harnesses all available unlabelled data, as evident from the rising confidence levels over time, depicted in fig. 4.4. This process culminates in the introduction of more pseudo-labels/samples into the learning phase, leading to a reduction in shared loss. By the final stages, the quantity of unlabelled samples employed approaches 16, equalling our stipulated maximum batch size for unlabelled images (fig. 4.4).

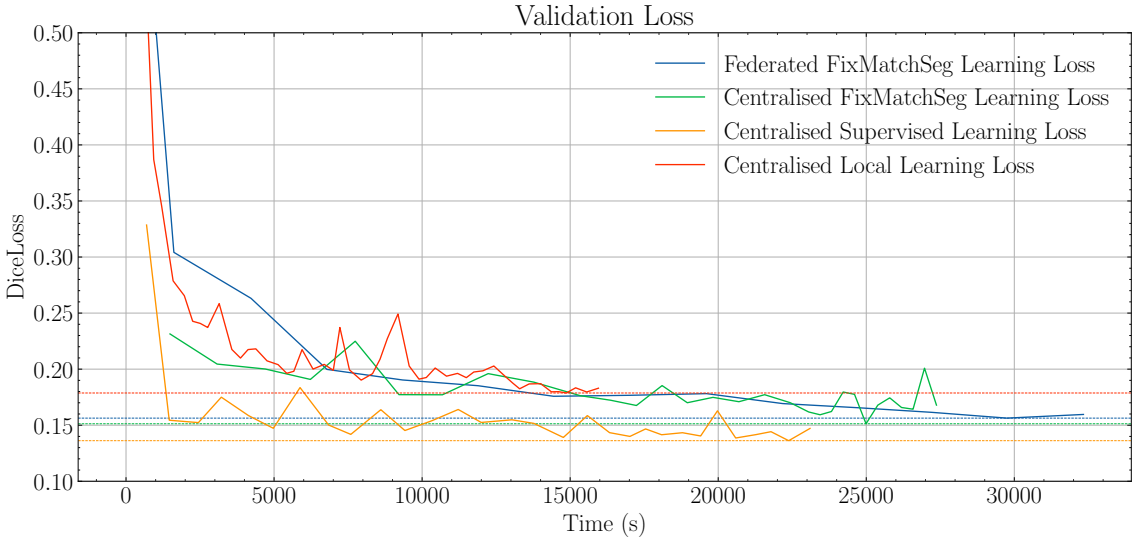


Figure 4.3: Validation losses for the different experiment setups over elapsed time in seconds. The dashed lines show the minimum loss achieved.

The utilisation of unlabelled data, as illustrated in fig. 4.4, is a testament to the effectiveness of this resource. When scenarios with a common set of 100 labelled

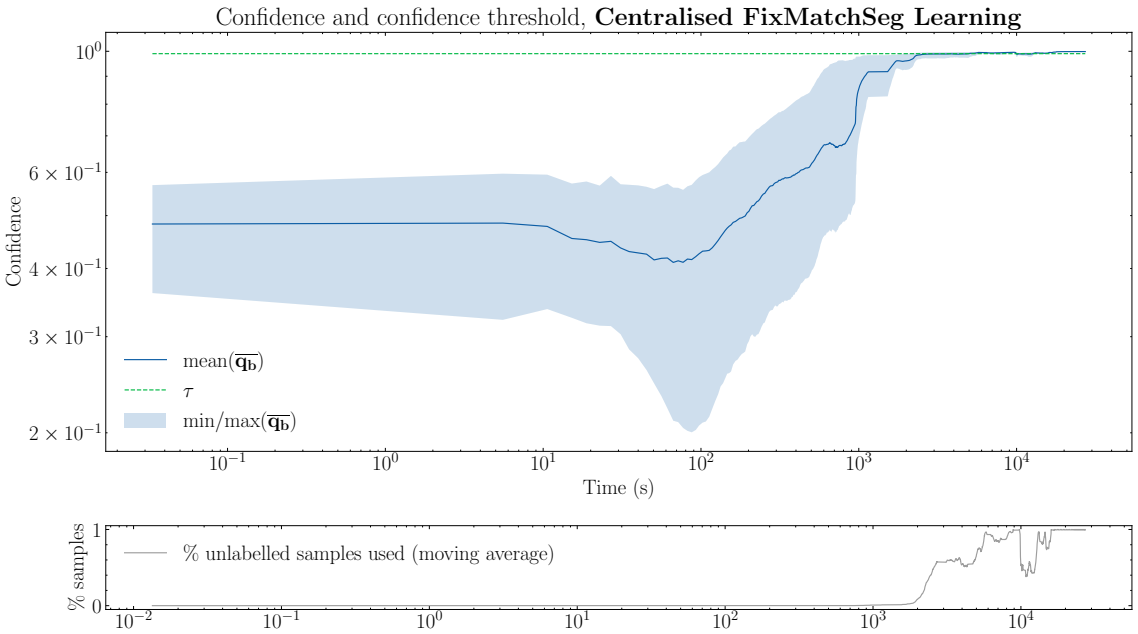


Figure 4.4: Confidence levels for labelled and unlabelled data. The bottom plot shows the percentage of pseudo-labels or unlabelled images that are used in the FixMatchSeg algorithm for an iteration over time. At the end, the batch is filled with unlabelled data at max size 16.

images and a multitude of unlabelled images are compared, namely, Centralised Local Learning versus Centralised FixMatchSeg Learning, it becomes evident that the FixMatchSeg algorithm and the unlabelled images collectively contribute to the reduction in validation loss.

Despite these differences, visual evaluation suggests minor disparities and shared challenges across all models. Referencing figs. 4.5 to 4.8, it is clear that all models struggle with accurately capturing the left turn in image 1 (top left) and 8 (bottom right). Nonetheless, the models have effectively grasped the semantic representation of roads.

An additional observation is the presence of inaccuracies in the ground truth. For example, image 5 incorrectly classifies a highway railing as a road, and image 6 mislabels a stone wall to the right as a road.

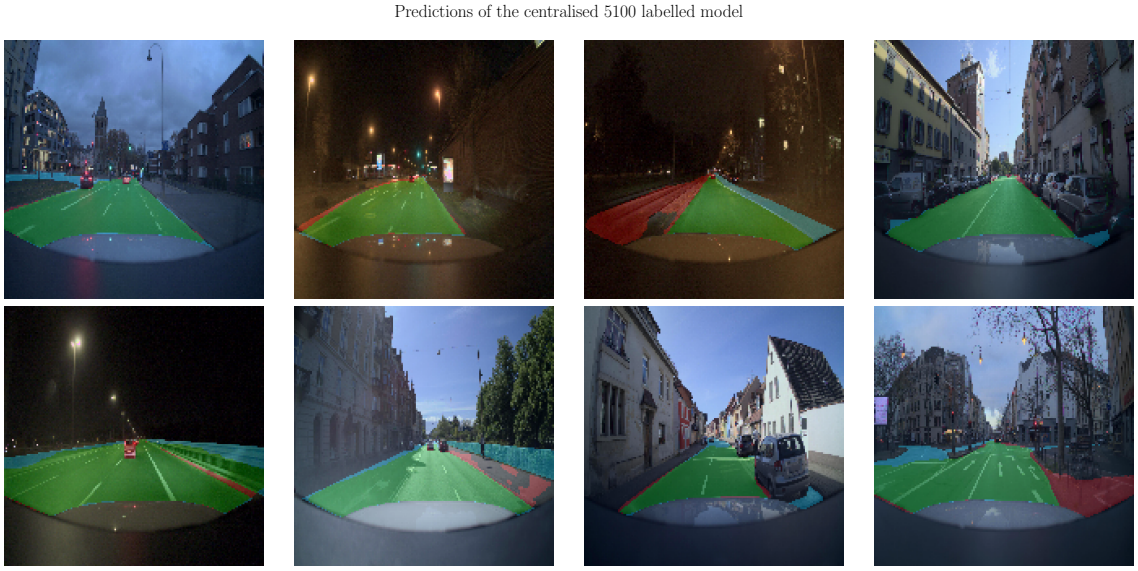


Figure 4.5: Predictions with overlay. Green: TP, Red: FP, Blue: FN

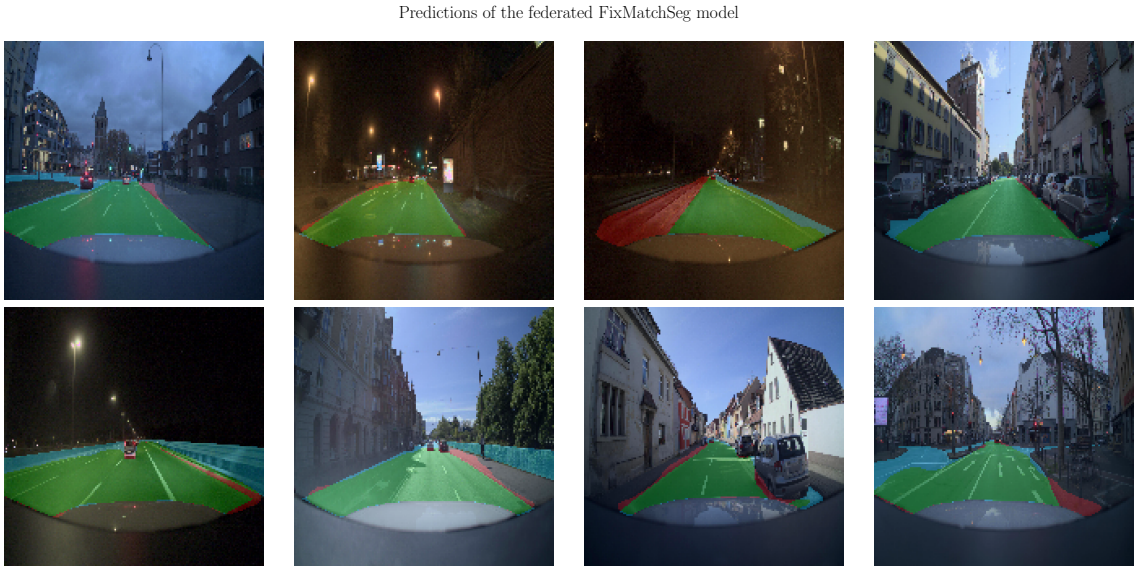


Figure 4.6: Predictions with overlay. Green: TP, Red: FP, Blue: FN

4.5 Discussion

4.5.1 The problem design

The design of our problem is a pivotal aspect that critically influences the results and inferences derived from our study. This design embraces our selection of models, employed learning methods, and steps taken in data processing, each of which exerts a significant influence on model performance and interpretability.

In our research, we employed models that have a well-established reputation for efficacy in tasks associated with image segmentation. However, an intriguing ob-

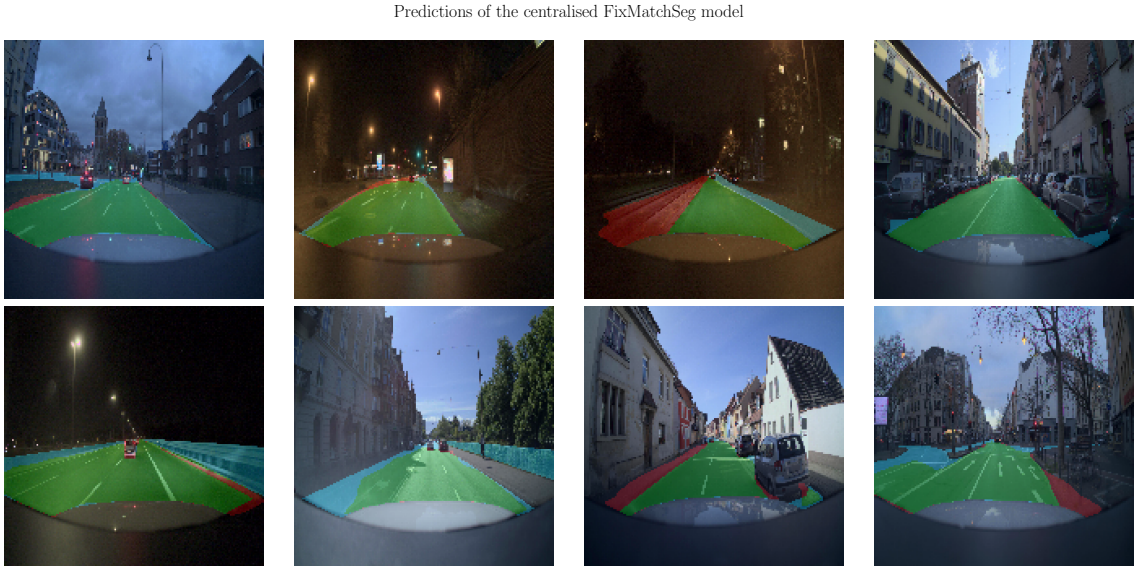


Figure 4.7: Predictions with overlay. Green: TP, Red: FP, Blue: FN

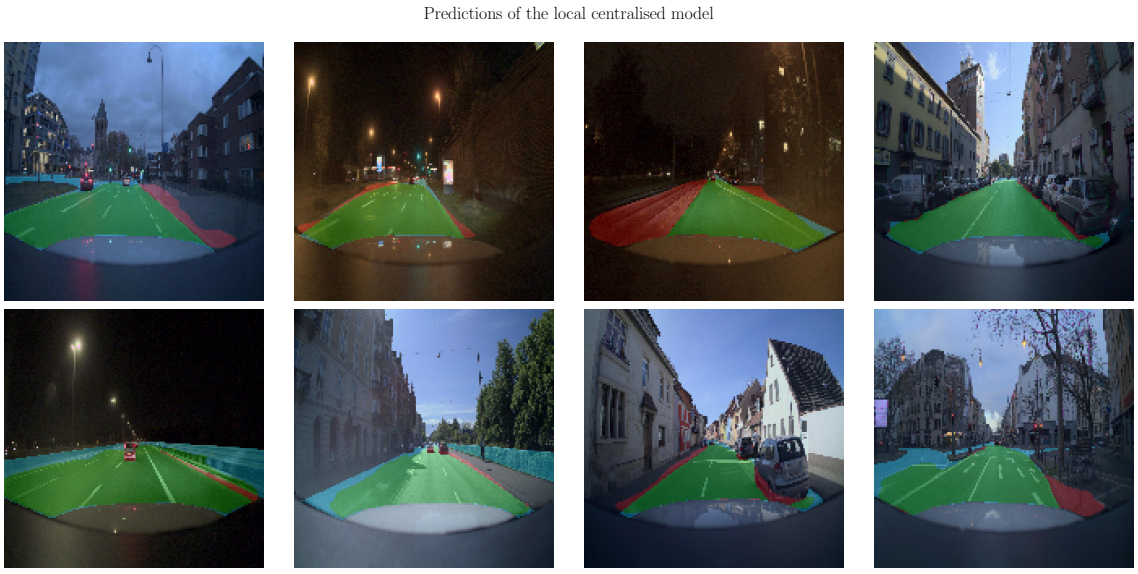


Figure 4.8: Predictions with overlay. Green: TP, Red: FP, Blue: FN

servation was the uniformity of losses across all models, irrespective of the learning method implemented. This could potentially indicate inherent constraints or limitations in our chosen models. It's plausible that certain intrinsic properties of these models induce comparable behaviours, manifesting in similar loss profiles. This observation underscores the potential requirement to expand our exploration to encompass a wider variety of model architectures, which might prove more adept at capturing the nuances of the road segmentation problem.

An additional critical aspect of our problem design lies in the resolution of the images used in our study. Set at 160x160 pixels, this resolution may inherently limit the level of detail in the images, potentially failing to provide a comprehensive

representation of the complexities associated with real-world scenarios. An evident manifestation of this limitation is the models' inability to correctly segment certain road features, such as specific turns further away from the ego-car. This raises pertinent questions regarding the potential advantages of employing higher-resolution images and whether an enhanced model performance, potentially stemming from such a change, would offset the associated computational costs and the relative performance between the different models.

4.5.2 Road segmentation as a task

The task of road segmentation inherently presents a unique set of challenges stemming from the intricacy of real-world scenarios. As a binary problem, it simplifies the complex reality of roads and their environments into two distinct categories: road and not-road. While this simplification aids in establishing a manageable task, it could concurrently limit the models' ability to wholly capture the elaborate nuances associated with the diversity of real-world conditions.

Our models need to not only identify and accurately segment roads under a plethora of conditions and environments but also distinguish roads from non-road features that may exhibit visually similar characteristics. This complexity demands high precision and a robust modelling approach to effectively navigate the intricate task landscape.

An intriguing observation from our study was the similarity in challenges encountered by all models, despite the diversity in their learning mechanisms. This homogeneity in struggle points could potentially suggest that the task of road segmentation might necessitate more sophisticated or bespoke models. Such models, designed explicitly with the task's complexity in mind, could provide a more comprehensive account for the extensive variety of real-world conditions and situations.

Another noteworthy observation pertains to mislabelling in the ground truth. The identified instances of erroneous labelling underscore the critical role of high-quality, accurately labelled data in training effective and robust models. Mislabelled data can lead to incorrect learning, reducing the models' ability to generalise well to unseen data, a vital aspect of successful road segmentation.

4.5.3 Usage of unlabelled data

The use of unlabelled data represents an intriguing proposition in semi-supervised learning scenarios, stemming from its relative ease of acquisition and substantial availability compared to labelled data. Labelling data often necessitates human expertise, can be time-consuming, and involves substantial financial implications. Therefore, leveraging unlabelled data is a cost-effective alternative, provided that it is utilised efficiently and effectively.

Our study delved into the realm of semi-supervised learning through the implementation of the FixMatchSeg algorithm, which facilitated the assimilation of unlabelled

data into the learning process. The comparative results, obtained through the exploration of different learning mechanisms, underscore the positive contributions of unlabelled data. In scenarios where unlabelled data was integrated, lower validation losses were noted compared to the model trained exclusively on labelled data.

Nevertheless, while the inclusion of unlabelled data seemingly enhances the learning process, its efficacy hinges heavily on its representativeness. If the unlabelled data fails to encapsulate the breadth and complexity of the task at hand, it may impose limitations on the model's generalisability to unseen instances. This reveals an inherent dichotomy; while unlabelled data can be readily acquired, its successful incorporation requires a judicious approach to ensure it comprehensively represents the task complexity and diversity.

4.5.4 Federated learning for road segmentation

The implementation of federated learning in our study was an attempt to innovate and leverage distributed data sources in the domain of road segmentation. This learning paradigm, by emphasising data privacy and confidentiality, offers a promising avenue to maximise the utility of distributed, on-device data while respecting user privacy concerns.

Considering the strict privacy constraints that prohibit the data from leaving the vehicle, federated learning enables us to incorporate semi-supervised, self-supervised, or unsupervised learning strategies on distributed data sources. Consequently, it widens the scope of data availability and enhances its diversity, potentially contributing to the development of more robust and precise models.

Despite these enticing benefits, the results from our study reveal a contrasting reality; federated learning did not outperform the other learning methodologies in terms of validation loss. This discrepancy signals the challenges innate to federated learning, including but not limited to the quality and diversity of data across different nodes, and the necessity for effective aggregation methods.

In essence, while the use of federated learning unveils a new horizon in the privacy-conscious utilisation of distributed data, it comes with its own set of complexities and constraints. These results underscore the importance of refining and optimising federated learning strategies specific to road segmentation. Future research should focus on the exploration of improved federated learning setups, potentially incorporating advanced aggregation mechanisms and ensuring high-quality data. It is through this focused refinement that federated learning could truly manifest its potential in the application of road segmentation.

5

Federated Learning for autonomous driving

In this chapter we discuss federated learning in general and it's potential application in the domain of autonomous driving. We utilise federated learning literature in autonomous driving applications, the results of this thesis, and other relevant literature regarding self-supervised, semi-supervised, and unsupervised learning as they are prerequisites for federated learning in autonomous driving contexts.

5.1 The state of federated learning

Federated learning, an innovative and transformative form of machine learning, empowers a distributed model of computation across multiple nodes, with each node retaining its local data. This approach aligns perfectly with the global trend of data privacy and security, mitigating the need to exchange raw data, thereby fostering privacy while reducing dependencies on data transfer infrastructure.

We posit that FL is an idea whose time has come. The rapid growth and adoption of digital technologies, along with heightened consciousness around data privacy, provide the ideal ecosystem for the expansion of FL. It's a promising solution that aligns technical advancement with privacy norms, a balance that has been somewhat elusive in the data-driven world.

Moreover, FL's core principle of collaboration and learning from distributed data sources is precisely what makes it a promising framework for diverse sectors, including but not limited to, healthcare, telecommunications, and IoT.

Specifically, in the context of autonomous driving, FL could be a game-changer. Autonomous vehicles generate a vast amount of data, and learning from this data in a distributed yet unified manner can significantly enhance their intelligence. FL allows these vehicles to learn from each other's experiences without compromising data privacy, ensuring each vehicle's local learnings contribute to the collective intelligence.

Yet, it is essential to acknowledge the challenges that exist in the current state of FL.

These include creating efficient communication protocols to minimise computational overhead, dealing with non-IID data distribution, and maintaining the robustness and reliability of the learning process. While these challenges pose certain hurdles, we believe they also offer exciting avenues for research and innovation, encouraging us to refine and redefine FL's possibilities.

The state of FL today is a reflection of the continuous innovation in this field. We see this not as a completed journey, but as a promising start, a roadmap to a future where privacy and technology coexist harmoniously, and fields like autonomous driving are revolutionised by the power of collective, yet private, learning.

5.2 ML Federations for AD

The concept of federated learning in the context of autonomous driving is a relatively recent and increasingly pivotal area of research. By leveraging on-device machine learning capabilities, FL allows for distributed training, which is particularly beneficial given the large amount of data that autonomous vehicles generate and process.

A key question that arises in the application of FL to AD is whether such an approach can indeed surpass traditional, centralised learning methods, especially in tasks related to computer vision, an integral component of AD.

Several authors have already ventured into this territory, each contributing unique methodologies and insights, but also encountering specific challenges. These previous efforts provide invaluable context and basis for our research, as we seek to further the application of FL in AD.

Previous work has often operated under the idealistic assumption that labelled data is plentiful and persistently available on client vehicles. However, in real-world situations, this scenario is not always tenable. Due to privacy constraints and the sheer volume of data produced by AD systems, it is often impractical and undesired to offload raw data from the vehicles. Instead, data needs to be processed and learnt from in a localised, on-vehicle manner.

This leads us to a pivotal question: How can we adapt to situations where there's an acute scarcity of labelled data, or where it's non-existent altogether? Here, we propose to tackle this challenge by employing semi-supervised, self-supervised, or unsupervised learning methodologies within the FL framework.

In our research, we have ventured into this less-charted territory, implementing a semi-supervised learning algorithm for ego-road segmentation, and using imitation learning for trajectory prediction. We believe that our efforts stand as a pioneering attempt to overcome the significant hurdle of data limitation in a federated setting.

To the best of our knowledge, this is the first study of its kind that addresses this challenge head-on, rather than sidestepping it. The traditional assumptions are

reconsidered, and new methodologies are brought into play. Our approach strives to bridge the gap between the theoretical ideal of FL and its practical application in the field of AD.

By shifting the focus from reliance on abundant labelled data to harnessing the potential of self-supervised and semi-supervised learning methodologies, we aim to develop a more realistic and effective FL framework for AD. This not only alleviates the dependency on labelled data but also maximises on-device processing capabilities. In our view, this is a crucial step in the evolution of FL for AD, contributing to both its theoretical understanding and its practical feasibility.

5.3 The Labelling Gap

Federated learning applications in real-world scenarios, particularly in autonomous driving, confront a substantial challenge, which we term the *labelling gap*. In the AD landscape, a wealth of data is readily available; however, due to privacy considerations, logistical hurdles, and the high costs of data transportation and manual annotation, label availability significantly dwindles within FL contexts. With data residing on the device, or more specifically, the vehicle, in a federated setting, obtaining labels becomes a non-trivial task since the data is not permitted to leave the vehicle.

Unsupervised, semi-supervised, and self-supervised learning methodologies emerge as potential solutions to bridge this gap. These methodologies have the capacity to learn from valuable data representations and patterns without heavily relying on the availability of labels.

Foundational work such as SimCLR [67] and MoCo [68] lay the groundwork for unsupervised visual representation learning. Google's introduction of BYOL, with online and moving-averaged target networks [69], marked a significant milestone in this journey. Despite their considerable promise, these methodologies were limited to classification tasks and utilised convnets as a backbone.

A crucial shift occurred in 2021 when Facebook unveiled their research titled *Emerging Properties in Self-Supervised Vision Transformers* [70], introducing DINO. They demonstrated that vision transformer (ViT) features could encapsulate explicit semantic segmentation information of an image, a characteristic not as apparent with supervised ViTs or convnets. Recent progress in semantic segmentation frequently employs ViTs as backbones, as evidenced by CutLER, a zero-shot unsupervised object detector and instance segmentor which exhibited impressive results in detecting and segmenting prevalent objects in a scene.

We may speculate that the future of computer vision tasks will mirror the progression seen in natural language processing, where models are trained without labels, and only a fraction of available labels are employed in a few-shot setting for task-specific fine-tuning.

In conclusion, while the labelling gap presents a significant hurdle in applying FL to AD, it is not an insurmountable obstacle. The consistent advancements in self-supervised and unsupervised learning methodologies offer an exciting avenue to tackle this issue. Incorporating these methods within an FL framework holds the potential to revolutionise our capacity to learn from the vast, yet predominantly unlabelled, data generated by AD systems.

5.4 Our contribution

In this research, we aimed to push the frontiers of federated learning in autonomous driving by focusing on two critical computer vision tasks - ego-trajectory prediction and ego-road segmentation. By simplifying these problems and experimenting with various learning settings, our work explores the influence of labels, federation and aggregation, and local learning on the learning process.

Our methodology introduces FixMatchSeg, a semi-supervised learning algorithm, as an innovative solution to leverage unlabelled data, thereby reducing dependency on labels. While there have been previous studies attempting self-annotation in AD tasks, such as steering angle prediction akin to our trajectory prediction task, our work stands apart in the literature. To our knowledge, we are the first to experiment with unlabelled data in FL for AD, a significant stride towards resolving the "labelling gap" discussed earlier.

The results of our research indicated that, for our simplified tasks, FL performance was at par with traditional learning methods or exhibited only minor differences. This finding is significant as it supports the hypothesis that FL, even with limited labelled data, can effectively address key AD tasks.

However, it's important to note the limitations of our work. While our findings are promising, the simplified nature of our tasks may not perfectly mirror real-world complexity. Additionally, our results were obtained within a controlled experimental setting, making further real-world testing essential.

5.5 Future directions

Our work provides an exciting foundation that offers several potential paths for future research. Here, we propose some avenues of inquiry that could further the understanding and development of Federated Learning within autonomous driving and federated learning.

Incorporating Vision Transformers (ViTs): Given the success of ViTs in self-supervised learning scenarios, particularly in tasks requiring semantic segmentation, it would be worthwhile to explore their potential within a federated learning context. Recent advances like DINO and CutLER suggest that the adoption of ViTs in federated settings might lead to significant breakthroughs.

Exploration of Other Self-Supervised Methods: While we have focused on a semi-supervised learning method (FixMatchSeg) in our work, the landscape of self-supervised learning methods is vast and continuously evolving. Therefore, exploring and integrating other self-supervised methods, such as contrastive learning, into federated learning frameworks might yield further improvements in performance, especially when labelled data is limited.

Understanding Data Distributions: As autonomous vehicles generate vast amounts of data, understanding and effectively managing the data distributions across different clients (vehicles) becomes a critical issue in a federated learning context. Future work could focus on designing novel methods for client selection, sampling, and scheduling, taking into account the inherent non-IID nature of data distributions in this field.

Development of High Quality Datasets: Given the importance of high-quality data for effective machine learning, efforts should be directed towards the creation of robust, diverse, and representative datasets for autonomous driving. These datasets should ideally be designed with privacy-preserving measures in mind, to ensure they can be readily used in federated learning scenarios. These datasets could also be used in few-shot or semi-supervised methods with self-supervised/unsupervised backbones.

6

Conclusion

In this thesis, we delved into the world of federated learning in the context of autonomous driving. We posed the crucial question - can FL surpass traditional centralised learning methods, especially when labelled data is scarce or absent? This is a pressing query that arises when we consider the reality of data privacy constraints and the impracticality of offloading raw data from vehicles.

Our thesis led us to challenge the assumptions of previous work, which largely operated under the idealistic premise that labelled data would be plentiful and persistently available on client vehicles. In contrast, we decided to tackle the problem of scarce labelled data head-on. We adopted and implemented a semi-supervised learning algorithm for ego-road segmentation and utilised imitation learning for trajectory prediction, making this, to the best of our knowledge, one of the first studies to directly address this significant challenge.

However, as we journeyed through our exploration, we acknowledged that the road to effectively applying FL in AD is not without its bumps. While our results demonstrate the potential of FL, even when operating with limited labelled data, the performance was found to be on par with or slightly less effective than traditional learning methods for the simplified tasks we studied.

Moreover, our research has primarily focused on leveraging semi-supervised learning to bridge the 'labelling gap'. There remains a vast landscape of unsupervised and self-supervised learning methodologies that are yet to be thoroughly explored within the FL framework. Given the rapid advances in self-supervised learning and its success in deep localisation tasks, we believe that the integration of such methods in a federated context could lead to significant breakthroughs in the AD domain.

While our work has pushed the boundaries of FL in the AD context, we are only at the beginning of this exploration. The promise that FL holds for the realm of autonomous driving is substantial, but so are the challenges that lie ahead. We hope that our work serves as a stepping stone, prompting further research and encouraging continued innovation in this field. The potential benefits of a future where autonomous vehicles can learn collectively, without compromising privacy and efficiency, certainly make this challenging journey worthwhile.

Bibliography

[1] S. Singh, "Critical reasons for crashes investigated in the national motor vehicle crash causation survey," Tech. Rep., 2015.

[2] T. J. Crayton and B. M. Meier, "Autonomous vehicles: Developing a public health research agenda to frame the future of transportation policy," *Journal of Transport & Health*, vol. 6, pp. 245–252, 2017.

[3] E. Yurtsever, J. Lambert, A. Carballo, and K. Takeda, "A survey of autonomous driving: Common practices and emerging technologies," *CoRR*, vol. abs/1906.05113, 2019. arXiv: 1906.05113. [Online]. Available: http://arxiv.org/abs/1906.05113.

[4] European Commission, Directorate General for Transport, *European Commission, Autonomous Vehicles & Road Safety*, Online, Feb. 2018.

[5] R. Frisoni, A. Dall'Oglio, C. Nelson, *et al.*, "Research for tran committee–self-piloted cars: The future of road transport?," 2016.

[6] A. Ghasemieh and R. Kashef, "3d object detection for autonomous driving: Methods, models, sensors, data, and challenges," *Transportation Engineering*, vol. 8, p. 100 115, 2022, ISSN: 2666-691X. DOI: https://doi.org/10.1016/j.treng.2022.100115. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2666691X22000136.

[7] J. Mao, S. Shi, X. Wang, and H. Li, "3d object detection for autonomous driving: A review and new outlooks," *arXiv preprint arXiv:2206.09474*, 2022.

[8] M. Consortium. "Melloddy: Imi." Accessed: 2023-05-31, The Innovative Medicines Initiative (IMI). (2023), [Online]. Available: https://www.melloddy.eu/ (visited on 05/10/2023).

[9] Q. Yang, Y. Liu, Y. Cheng, Y. Kang, T. Chen, and H. Yu, *Federated Learning* (Synthesis Lectures on Artificial Intelligence and Machine Learning). Morgan & Claypool Publishers, 2019, ISBN: 9781681736983.

[10] T. Li, A. K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, and V. Smith, "Federated learning: Challenges, methods, and future directions," *IEEE Signal Processing Magazine*, vol. 37, no. 3, pp. 50–60, 2020. DOI: 10.1109/MSP.2020.2975749. [Online]. Available: https://arxiv.org/abs/1908.07873.

[11] P. Kairouz, H. B. McMahan, B. Avent, *et al.*, "Advances and open problems in federated learning," *arXiv preprint arXiv:1912.04977*, 2019. [Online]. Available: https://arxiv.org/abs/1912.04977.

[12] I. Goodfellow, Y. Bengio, A. Courville, and Y. Bengio, *Deep learning*. MIT press, 2016. [Online]. Available: http://www.deeplearningbook.org/.

[13] A. Lindholm, N. Wahlström, F. Lindsten, and T. B. Schön, *Machine Learning - A First Course for Engineers and Scientists*. Cambridge University Press, 2022. [Online]. Available: https://smlbook.org.

[14] H. B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," 2017. arXiv: 1602.05629.

[15] K. Martineau, *What is federated learning?* Oct. 2022. [Online]. Available: https://research.ibm.com/blog/what-is-federated-learning.

[16] Q. Yang, Y. Liu, T. Chen, and Y. Tong, "Federated machine learning: Concept and applications," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 10, no. 2, pp. 1–19, 2019.

[17] J. Anderson, N. Kalra, K. Stanley, P. Sorensen, C. Samaras, and O. Oluwatola, *Autonomous vehicle technology: A guide for policymakers*, RAND Corporation, 2014.

[18] W. Shi, J. Cao, Q. Zhang, Y. Li, and L. Xu, "Edge computing: Vision and challenges," *IEEE Internet of Things Journal*, vol. 3, no. 5, pp. 637–646, 2016.

[19] T. S. A. for Privacy Protection (IMY). "Final report on imy's pilot project with regulatory sandbox on data protection." (2023), [Online]. Available: https://www.imy.se/globalassets/dokument/rapporter/slutrapport-om-imys-pilotprojekt-med-regulatorisk-testverksamhet-om-dataskydd-230315.pdf (visited on 06/03/2023).

[20] European Parliament and Council of the European Union, *Regulation (EU) 2016/679 of the European Parliament and of the Council*, of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation), OJ L 119, 4.5.2016, p. 1–88, May 4, 2016. [Online]. Available: https://data.europa.eu/eli/reg/2016/679/oj (visited on 04/13/2023).

[21] M. Abadi, A. Chu, I. Goodfellow, *et al.*, "Deep learning with differential privacy," in *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, 2016, pp. 308–318.

[22] E. Bagdasaryan, A. Veit, Y.-C. Hua, D. Estrin, and V. Shmatikov, "Backdoor attacks on federated learning," in *2020 IEEE Security and Privacy Workshops (SPW)*, IEEE, 2020, pp. 13–19. DOI: 10.1109/SPW51313.2020.00009.

[23] T. Li, A. K. Sahu, A. Talwalkar, and V. Smith, "Fair resource allocation in federated learning," *arXiv preprint arXiv:2005.06622*, 2020.

[24] M. Bojarski, D. Del Testa, D. Dworakowski, *et al.*, "End to end learning for self-driving cars," in *ECCV*, 2016.

[25] C. Chen, A. Seff, A. Kornhauser, and J. Xiao, "Deepdriving: Learning affordance for direct perception in autonomous driving," *ICCV*, 2015.

[26] O. Chapelle, B. Schölkopf, and A. Zien, *Semi-Supervised Learning*. MIT Press, 2010, ISBN: 9780262514125.

[27] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2015.

[28] P. Izmailov, D. Podoprikhin, T. Garipov, D. Vetrov, and A. G. Wilson, "Averaging weights leads to wider optima and better generalization," *arXiv preprint arXiv:1803.05407*, 2018.

[29] T. Litman, "Autonomous vehicle implementation predictions," *Victoria Transport Policy Institute*, 2019.

[30] J. Beinke, S. Tomforde, H. Schmeck, and J. Branke, *Swarm learning: Decentralised learning for decentralised systems*, 2021.

[31] M. Alibeigi, W. Ljungbergh, A. Tonderski, *et al.*, "Zenseact open dataset: A large-scale and diverse multimodal dataset for autonomous driving," *arXiv preprint arXiv:2305.02008*, 2023.

[32] Z. Du, C. Wu, T. Yoshinaga, K.-L. A. Yau, Y. Ji, and J. Li, "Federated learning for vehicular internet of things: Recent advances and open issues," *IEEE Open Journal of the Computer Society*, vol. 1, pp. 45–61, 2020. DOI: 10.1109/OJCS.2020.2992630.

[33] A. Nguyen, T. Do, M. Tran, *et al.*, "Deep federated learning for autonomous driving," in *2022 IEEE Intelligent Vehicles Symposium (IV)*, 2022, pp. 1824–1830. DOI: 10.1109/IV51971.2022.9827020.

[34] H. Zhang, J. Bosch, and H. H. Olsson, "End-to-end federated learning for autonomous driving vehicles," in *2021 International Joint Conference on Neural Networks (IJCNN)*, 2021, pp. 1–8. DOI: 10.1109/IJCNN52387.2021.9533808.

[35] L. Fantauzzo, E. Fanì, D. Caldarola, *et al.*, "Feddrive: Generalizing federated learning to semantic segmentation in autonomous driving," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2022, pp. 11 504–11 511. DOI: 10.1109/IROS47612.2022.9981098.

[36] A. M. Elbir, B. Soner, S. Çöleri, D. Gündüz, and M. Bennis, "Federated learning in vehicular networks," in *2022 IEEE International Mediterranean Conference on Communications and Networking (MeditCom)*, 2022, pp. 72–77. DOI: 10.1109/MeditCom55741.2022.9928621.

[37] Z. Cui, R. Nishihara, F. Nie, L. Wilcox, and M. I. Jordan, "Multimodal trajectory predictions for autonomous driving using deep convolutional networks," in *2019 International Conference on Robotics and Automation (ICRA)*, 2019, pp. 2090–2096. DOI: 10.1109/ICRA.2019.8793938.

[38] W. Schwarting, J. Alonso-Mora, and D. Rus, "Planning and decision-making for autonomous vehicles," *Annual Review of Control, Robotics, and Autonomous Systems*, vol. 1, May 2018. DOI: 10.1146/annurev-control-060117-105157.

[39] A. Kendall, J. Hawke, D. Janz, *et al.*, "Learning to drive in a day," *CoRR*, vol. abs/1807.00412, 2018. arXiv: 1807.00412. [Online]. Available: http://arxiv.org/abs/1807.00412.

[40] F. Codevilla, M. Miiller, A. López, V. Koltun, and A. Dosovitskiy, "End-to-end driving via conditional imitation learning," *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 1–9, 2018. DOI: 10.1109/ICRA.2018.8463184.

[41] F. Leon and M. Gavrilescu, "A review of tracking and trajectory prediction methods for autonomous driving," *Mathematics*, vol. 9, no. 6, 2021, ISSN: 2227-7390. DOI: 10.3390/math9060660. [Online]. Available: https://www.mdpi.com/2227-7390/9/6/660.

[42] M. Bergqvist and O. Rödholm, *Deep path planning using images and object data*, 2018.

[43] P. Cai, Y. Sun, H. Wang, and M. Liu, "Vtgnet: A vision-based trajectory generation network for autonomous vehicles in urban environments," *CoRR*, vol. abs/2004.12591, 2020. arXiv: 2004.12591. [Online]. Available: https://arxiv.org/abs/2004.12591.

[44] P. Cai, Y. Sun, Y. Chen, and M. Liu, "Vision-based trajectory planning via imitation learning for autonomous vehicles," in *2019 IEEE Intelligent Transportation Systems Conference (ITSC)*, 2019, pp. 2736–2742. DOI: 10.1109/ITSC.2019.8917149.

[45] O. Khakhlyuk, "Using recurrency for ego-lane trajectory prediction from a single monocular camera, Verwenden von rekurrenz für vorhersage der ego-lane-trajektorie mit einer monokularen kamera," Master's Thesis, Technical University of Munich, Munich, 2023.

[46] H. Kilichenko, "Multimodal trajectory prediction for self-driving vehicles using a single monocular camera, Multimodale trajektorienvorhersage für selbstfahrende fahrzeuge mit einer einzigen monokularen kamera," Master's Thesis, Technical University of Munich, Munich, 2023.

[47] O. T. Solutions, *Oxts rt3000 inertial and gnss navigation system*, https://www.oxts.com/products/rt3000/, Accessed: 2023-05-05, 2023.

[48] Zenseact. "Zenseact open dataset, devkit," GitHub. (2021), [Online]. Available: https://github.com/zenseact/zod (visited on 05/10/2023).

[49] V. Smith, C.-K. Chiang, M. Sanjabi, and A. S. Talwalkar, "Federated learning: Strategies for improving communication efficiency," in *Advances in Neural Information Processing Systems*, 2017, pp. 3397–3407.

[50] G. Rizzoli, F. Barbato, and P. Zanuttigh, "Multimodal semantic segmentation in autonomous driving: A review of current approaches and future perspectives," *Technologies*, vol. 10, p. 90, Jul. 2022. DOI: 10.3390/technologies10040090.

[51] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," *CoRR*, vol. abs/1411.4038, 2014. arXiv: 1411.4038. [Online]. Available: http://arxiv.org/abs/1411.4038.

[52] S. Minaee, Y. Boykov, F. Porikli, A. Plaza, N. Kehtarnavaz, and D. Terzopoulos, "Image segmentation using deep learning: A survey," *CoRR*, vol. abs/2001.05566, 2020. arXiv: 2001.05566. [Online]. Available: https://arxiv.org/abs/2001.05566.

[53] H. Noh, S. Hong, and B. Han, "Learning deconvolution network for semantic segmentation," *CoRR*, vol. abs/1505.04366, 2015. arXiv: 1505.04366. [Online]. Available: http://arxiv.org/abs/1505.04366.

[54] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *CoRR*, vol. abs/1511.00561, 2015. arXiv: 1511.00561. [Online]. Available: http://arxiv.org/abs/1511.00561.

[55] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," *CoRR*, vol. abs/1505.04597, 2015. arXiv: 1505.04597. [Online]. Available: http://arxiv.org/abs/1505.04597.

[56] M. Cordts, M. Omran, S. Ramos, *et al.*, "The cityscapes dataset for semantic urban scene understanding," *CoRR*, vol. abs/1604.01685, 2016. arXiv: 1604.01685. [Online]. Available: http://arxiv.org/abs/1604.01685.

[57] L. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," *CoRR*, vol. abs/1706.05587, 2017. arXiv: 1706.05587. [Online]. Available: http://arxiv.org/abs/1706.05587.

[58] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," *CoRR*, vol. abs/1612.01105, 2016. arXiv: 1612.01105. [Online]. Available: http://arxiv.org/abs/1612.01105.

[59] Z. Chen and Z. Chen, "Rbnet: A deep neural network for unified road and road boundary detection," in *Neural Information Processing: 24th International Conference, ICONIP 2017, Guangzhou, China, November 14-18, 2017, Proceedings, Part I 24*, Springer, 2017, pp. 677–687.

[60] M. Teichmann, M. Weber, J. M. Zöllner, R. Cipolla, and R. Urtasun, "Multinet: Real-time joint semantic reasoning for autonomous driving," *CoRR*, vol. abs/1612.07695, 2016. arXiv: 1612.07695. [Online]. Available: http://arxiv.org/abs/1612.07695.

[61] Q. Wang, J. Gao, and Y. Yuan, "Embedding structured contour and location prior in siamesed fully convolutional networks for road detection," *CoRR*, vol. abs/1905.01575, 2019. arXiv: 1905.01575. [Online]. Available: http://arxiv.org/abs/1905.01575.

[62] Y. Chang, F. Xue, F. Sheng, W. Liang, and A. Ming, "Fast road segmentation via uncertainty-aware symmetric network," *arXiv preprint arXiv:2203.04537*, Mar. 2022. DOI: 10.48550/arXiv.2203.04537. arXiv: 2203.04537 [cs.CV]. [Online]. Available: https://arxiv.org/abs/2203.04537.

[63] K. Huang, B. Shi, X. Li, X. Li, S. Huang, and Y. Li, "Multi-modal sensor fusion for auto driving perception: A survey," *CoRR*, vol. abs/2202.02703, 2022. arXiv: 2202.02703. [Online]. Available: https://arxiv.org/abs/2202.02703.

[64] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever, *et al.*, "Improving language understanding by generative pre-training," 2018.

[65] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, *et al.*, "Language models are unsupervised multitask learners," *OpenAI blog*, vol. 1, no. 8, p. 9, 2019.

[66] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[67] T. Chen, S. Kornblith, M. Norouzi, and G. E. Hinton, "A simple framework for contrastive learning of visual representations," *CoRR*, vol. abs/2002.05709, 2020. arXiv: 2002.05709. [Online]. Available: https://arxiv.org/abs/2002.05709.

[68] K. He, H. Fan, Y. Wu, S. Xie, and R. B. Girshick, "Momentum contrast for unsupervised visual representation learning," *CoRR*, vol. abs/1911.05722, 2019. arXiv: 1911.05722. [Online]. Available: http://arxiv.org/abs/1911.05722.

[69] J. Grill, F. Strub, F. Altché, *et al.*, "Bootstrap your own latent: A new approach to self-supervised learning," *CoRR*, vol. abs/2006.07733, 2020. arXiv: 2006.07733. [Online]. Available: https://arxiv.org/abs/2006.07733.

[70] M. Caron, H. Touvron, I. Misra, *et al.*, "Emerging properties in self-supervised vision transformers," *CoRR*, vol. abs/2104.14294, 2021. arXiv: 2104.14294. [Online]. Available: https://arxiv.org/abs/2104.14294.

[71] X. Wang, R. Girdhar, S. X. Yu, and I. Misra, "Cut and learn for unsupervised object detection and instance segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 3124–3134.

[72] P. Upretee and B. Khanal, "Fixmatchseg: Fixing fixmatch for semi-supervised semantic segmentation," *arXiv preprint arXiv:2208.00400*, 2022.

[73] K. Sohn, D. Berthelot, C. Li, *et al.*, "Fixmatch: Simplifying semi-supervised learning with consistency and confidence," *CoRR*, vol. abs/2001.07685, 2020. arXiv: 2001.07685. [Online]. Available: https://arxiv.org/abs/2001.07685.

A

Appendix 1

A.0.1 Using EMA

In a trajectory planning experiment featuring a pre-trained MobileNet model fine-tuned via a centralised approach on a balanced dataset of 34,000 images, validated against Kilichenko's 3500-image dataset [46], we found that the Exponential Moving Average (EMA) played a pivotal role in enhancing the outcomes. Specifically, EMA significantly contributed to a reduction in loss, thereby demonstrating its efficacy in improving model performance. Both training and validation loss follow a downward trend without overfitting, demonstrating the model's efficacy in learning the data's inherent patterns. Notably, the EMA validation loss reached 1.396 after only 8 epochs and while the validation loss of the original model reaches this value after 33 epochs. This shows how EMA helps with faster convergence. Furthermore, the application of EMA provides a smoothed curve that eliminates abrupt changes, highlighting the underlying trend more clearly, see fig. A.1 .

In essence, using EMA for validation and inference facilitates more accurate evaluations and predictions by mitigating the impact of sudden loss spikes. This smoothens the loss evolution curve, assisting in identifying the genuine learning progression, and thereby enhancing the model's ability to generalise from the training to the validation set. This shows EMA's positive impact on model performance and reliability.

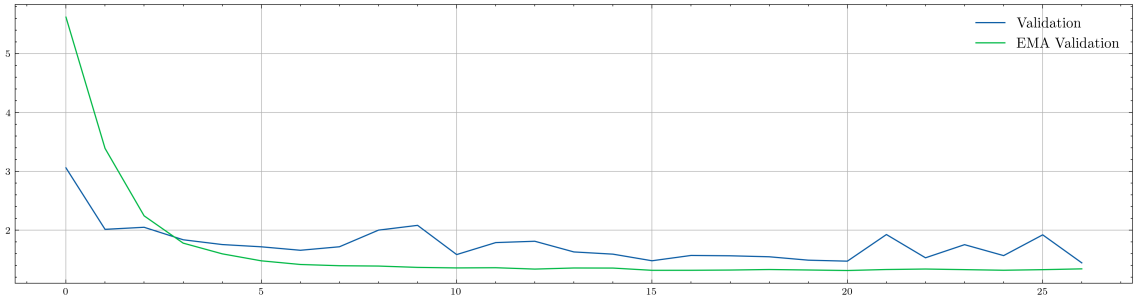


Figure A.1: The impact of EMA in reducing and stabilising the validation loss.

DEPARTMENT OF ELECTRICAL ENGINEERING
CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden
www.chalmers.se

CHALMERS
UNIVERSITY OF TECHNOLOGY