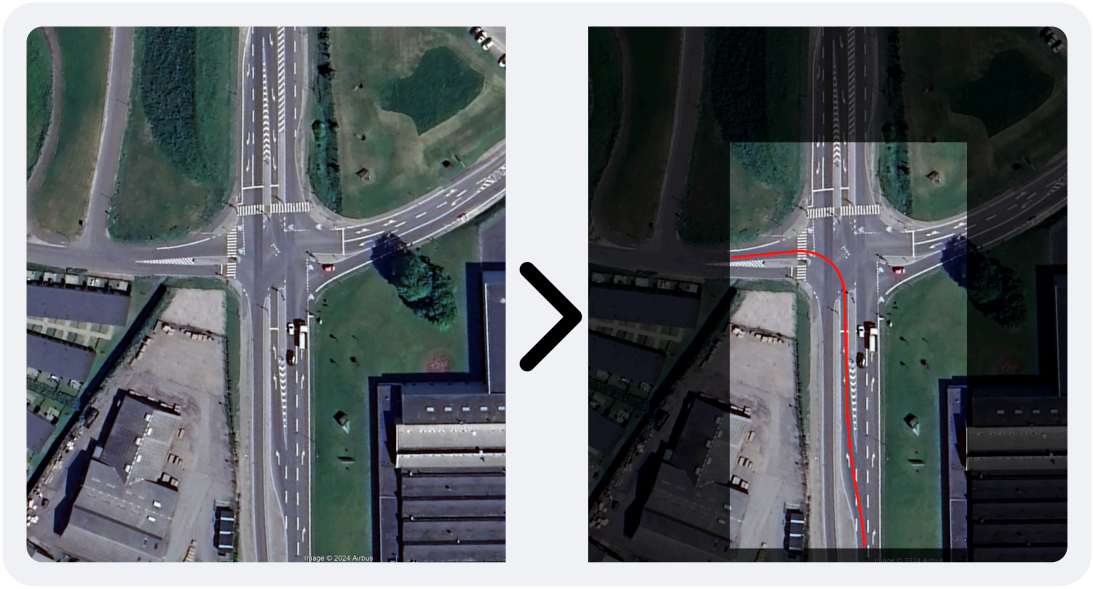
Pre-planning Intersection Traversal for Autonomous Vehicles

Master's Thesis in Computer Engineering

Ian Dahl Oliver

Department of Electrical and Computer Engineering
Aarhus University
Aarhus, Denmark
ian.oliver@post.au.dk

10-05-2025



Supervisor: Lukas Esterle
lukas.esterle@ece.au.dk

AARHUS UNIVERSITY

Preface

This master thesis is titled *"Pre-planning Intersection Traversal for Autonomous Vehicles"* and is devised by Ian Dahl Oliver. The author is a student at Aarhus University, Department of Electrical and Computer Engineering, enrolled in the Computer Engineering Master's programme. The author has completed a Bachelor's degree in Computer Engineering under the same conditions.

The thesis has been conducted in the period from 27-01-2025 to 05-06-2025, and supervised by Associate Professor Lukas Esterle. I would like to express my gratitude to my supervisor for his support and advice throughout the project.

An additional thanks goes to Associate Professor at AU, Kaare Mikkelsen, for his guidance in the early stages of this project.

Enjoy reading,
Ian Dahl Oliver

Abstract

hello Robot Operating System 2 (ROS2)

Nomenclature

Some terminology and type setting used in this thesis may not be familiar to the reader, and are explained here for clarity.

`monospace`	Inline monospace text is used for code function names, variables, or parameters.
`a.b`	In inline monospace text, a period `.` is used to denote a method or property of an object. Can also be used outside of monospace text.
`listing:<int>`	A reference to a specific listing, where `<int>` represents a line number.
`listing:<int>-<int>`	Reference to a range of lines within a listing.
`file.ext`	A reference to a specific file of given file type.
`file.ext:<func>`	A reference to a specific function within a file.
Stack	Also called a suite, a stack is a collection of related functions.

Acronyms Index

Acronym	Definition
AI	Artificial Intelligence
AIM	Autonomous Intersection Management
ANN	Artificial Neural Network
APF	Artificial Potential Field
API	Application Programming Interface
Application Programming Interfaces	ASPP
Atrous Spatial Pyramid Pooling	AUV
Autonomous Underwater Vehicle	AV
Autonomous Vehicle	Autonomous Vehicles
BCE	Binary Cross-Entropy
BEV	Bird's Eye View
CE	Cross-Entropy
CNN	Convolutional Neural Network
Convolutional Neural Networks	CPU
Central Processing Unit	CV
Computer Vision	DL
Deep Learning	DOF
Degrees of Freedom	EV
Electric Vehicle	FCN

Acronym	Definition
PH	Persistent Homology
PNG	Portable Network Graphics
PRNG	Pseudo-Random Number Generator
RL	Reinforcement Learning
RMSE	Root Mean Squared Error
RNG	Random Number Generator
ROS2	Robot Operating System 2
RRT	Rapidly-exploring Random Tree
TOML	Tom's Obvious Minimal Language
UI	User Interface
UX	User Experience
V2X	Vehicle-to-Everything
ViT	Vision Transformer
YAML	YAML Ain't Markup Language

Acronym	Definition
Fully Convolutional Network	FL
Fuzzy Logic	GA
Genetic Algorithm	GAN
Generative Adversarial Network	GIMP
GNU Image Manipulation Program	GNN
Graph Neural Network	GUI
Graphical User Interface	ID
Identifier	JSON
JavaScript Object Notation	MLP
Multilayer Perceptron	Multilayer Perceptrons
NN	Neural Network

Contents

1 Introduction ✓

The introduction to the thesis will be structured as follows. First, Section 1.1 presents the motivation for this thesis with the problem statement following in Section 1.2. With the motivation and problem statement in place, the research questions will be presented in Section 1.3.

1.1 Motivation ✓

Since the dawn of time, humans have made strides in automating any and all systems that surround them. In the days before technology, humans used animals to help them with their daily tasks. As technology advanced and exploded during the Industrial Revolution, humans replaced animals with machinery with the intent of automating production tasks. As this mentality continued to grow and technologies improved at a rapid pace, automation spread to other areas of life, such as transportation. The first Autonomous Vehicles (AVs) were developed in the 1980s, by both Americans, at Carnegie Mellon University [1], [2], and Europeans at Mercedes-Benz and Bundeswehr University Munich [3]. Since then, the development has spread to the individual car manufacturers instead of universities, making it more of a competitive field than a cooperative one. Still, the development in recent years with the rise of more powerful computers and the invention of machine learning, has led the field to become a fiercely researched area with many companies doing their part to create reliable, efficient, and safe AVs.

Despite this rapid development, AVs still encounter many challenges in their deployment, chief amongst which is their ability to handle intersections [4]. Unlike motorway driving — where lane following and obstacle detection are relatively well-defined tasks with few obstacles — intersections introduce a lot of complexity. Challenges arise from many different factors, such as unpredictable driver behaviour of other drivers, a huge variety of intersection types and configurations, and the state of the intersection with regard to faded or obstructed lane markings. Current solutions rely heavily on on-board sensors for perception and reactive decision-making, which can struggle in certain situations. Other solutions to intersections rely heavily on infrastructure support, such as V2X communication, which is not yet widely deployed. V2X communication is a technology that allows vehicles to communicate with each other and with the surrounding infrastructure, such as traffic lights and road signs. This technology can provide real-time information about traffic conditions, road hazards, and other important data that can help improve safety and efficiency on the roads.

A potential alternative to purely perception-based or infrastructure-dependent approaches is pre-planned path traversal, where an AV generates an optimal path through an intersection before reaching it. By leveraging Deep Learning (DL) models trained on annotated satellite imagery, AVs can gain a significantly increased amount of understanding about the intersection it is about to enter. This presents many potential improvements to the capability and efficiency of AVs. Firstly, it can reduce the reliance on on-board sensors by taking away the need for real-time perception and decision-making. Secondly, it can reduce the reliance on infrastructure support by allowing the AV to make decisions based on the pre-planned path. Finally, it can increase the safety of the AV by reducing the number of unpredictable situations it may encounter and give it a fair chance when it then does encounter foreign situations.

Beyond improving the performance of individual AVs, better intersection handling has even broader implications to the public as a whole. Optimized intersection traversal can lead to smoother traffic flow, reduced congestion, and improved urban mobility. By generating more efficient paths, AVs can reduce waiting times, minimize unnecessary stops, and create a more predictable and coordinated traffic environment. As the proportion of autonomous vehicles on the road increases, these optimizations scale exponentially, leading to fewer bottlenecks and a decrease in fuel consumption and emissions.

Furthermore, improving the performance of AVs in other areas can increase productivity as well. These areas include warehouse robotics, automated delivery systems, and smart logistics solutions that enhance efficiency and reduce operational costs. Even in the field of agriculture, AVs can be used to optimize crop management and harvesting processes, specifically in the context of precision agriculture. Racing is another area where AVs can be used to optimize performance and improve safety. By leveraging DL models trained on annotated satellite imagery, AVs can gain a better understanding of the track and its surroundings, allowing them to make more informed decisions and improve their performance, or give detailed information to the driver and crew.

1.2 Problem Statement ✓

Advancements in AV technologies have been at the forefront of tech innovations in the 21st century. A key challenge in the development of fully autonomous vehicles, is their ability to handle intersections. Intersections pose a wide variety of challenges to AVs: from those posed by complex structures, to those posed by the unpredictability of human drivers, to faded lines that make it difficult for the on-board computer vision system to clearly identify lanes or paths. All of these hinder AVs from reaching their full potential and being able to navigate intersections safely and efficiently.

Current existing solutions are very infrastructure-dependent. The Car2X system by Volkswagen, for example, relies on a network of sensors and communication devices installed in the infrastructure to spread information to vehicles on the road [5]. Autonomous Intersection Management (AIM) also relies on infrastructure to provide vehicles with information regarding intersections, with an orchestrator monitoring and managing individual

intersections [6], [7], [8], with active development moving towards a more decentralized and distributed approach [9]. Furthermore, reliance on camera-based vision is susceptible to environmental limitations, such as adverse weather, that reduce system reliability.

The challenges posed by intersections cause major problems for AV developers who want to push fully autonomous driving. AVs' inability to properly react to and handle intersections, leads to significant delays in real-world deployment as a consequence of the unreliability experienced by regulators and the public. If AVs want to enter the market with full self-driving capabilities, full autonomy is a key challenge to be tackled, as it is an essential task experienced when driving.

This project aims to develop a solution that will help AVs to better handle intersections. With the use of DL and Computer Vision (CV) technologies, trained on and utilizing satellite imagery, this project aims to train a small set of models that can accurately identify the proper path for an AV to travel through an intersection. The system is not meant to replace current systems deployed in AVs, but rather assist the existing systems make better decisions when in self-driving mode, approaching an arbitrary intersection.

1.3 Research Questions ✓

The following research questions have been formulated to address key challenges in AV path planning at intersections. The questions are designed to explore the effectiveness of different approaches and models in generating accurate and efficient paths for autonomous vehicles. The research questions are as follows:

RQ-1 How can pixel-subset-based deep learning approaches be optimized to improve accuracy and efficiency in path planning for autonomous vehicles at intersections? How do convolution-based and transformer-based models compare in this context?

RQ-2 Is it possible to design a loss function that effectively captures the similarity between generated and desired paths for autonomous vehicles without forcing exact matches?

RQ-3 Is it possible to create a dataset that allows for the training of a model, such that the data is not too stringent to a singular path?

2 Background ✓

This section outlines the theory relevant to the thesis. It begins with an introduction to AVs to provide a thorough understanding of the context in which the work is situated. The tool with which this project's problem statement will be tackled is presented in Section 2.2, where the fundamentals of DL are presented. Within this section, the focus is on CV and its applications in AVs, as well as the importance of datasets in training DL models. Section 2.3 presents underlying theory on satellite imagery and its applications in AVs. Each section and subsection includes clear examples to aid comprehension.

2.1 Autonomous Vehicles ✓

Initially released in 2014 by the Society of Automotive Engineers (SAE), the J3016 standard [10] defines six levels of driving automation, ranging from Level 0 to Level 5, with the latest revision released in 2021 [11]. These levels are visualized in Figure 1. These levels are further split into two separate categories based on the environment observer; the first three levels are concerned with the human driver being the environment observer, and the latter three levels refer to scenarios where the environment is monitored by the vehicle itself, through features collectively referred to as the Automated Driving System (ADS).

To understand how SAE defines the levels, it is important to gain a high-level overview of how these levels are defined. The document starts by defining the scope of the standard, clearly stating that it "describes [motor] vehicle driving automation systems that perform part or all of the dynamic driving task (DDT) on a sustained basis" [11, p. 4]. This definition excludes any momentary actor systems in place in a car, such as electronic stability control, automatic emergency braking, or lane keeping assistance (LKA). These systems are not considered to be part of the DDT, as they do not perform the driving task on a sustained basis. Finally, three primary actors are defined: the human user, the driving automation system, and other vehicle systems and components.

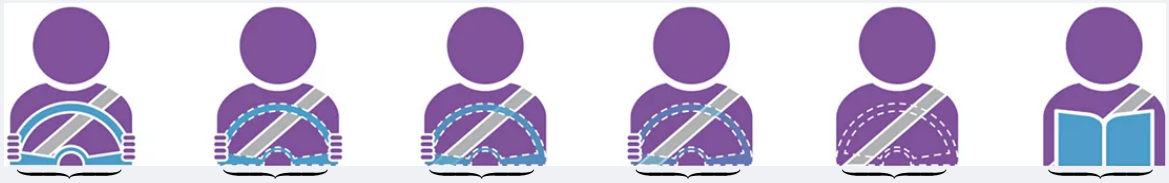


Figure 1. SAE 6 levels of autonomy. Image source: Synopsys [12]

Of course, many different manufacturers are actively developing their own systems, slowly climbing through the levels of automation, with some being further along than others. A comprehensive overview of the levels will be presented the following along with clear examples of each level. The levels are defined as follows:

- **Level 0**: No Automation. *Manual control. The human performs all driving tasks (steering, acceleration, braking, etc)*[1].

Level 0 autonomy may encompass more cars than one would imagine. This is largely due to the fact, that, as stated, momentary systems are not included as giving a vehicle any form of autonomy. Shown with the visualization in Figure 1, this level of autonomy requires the user to be in full control at all times during the DDT. That means keeping both hands on the wheel at all times and staying aware of the environment for the duration of the trip. Common systems like emergency braking and lane keeping assistance do not push vehicles with these features any higher than this level. This is due to their unsustained nature.

Today, while exact numbers are unknown, it is believed that the majority of vehicles on the road today are at this level. Statistics by statista [13], [14], hints that very few vehicles sold before 2014 had any form of autonomy. This is shown by their 2015 statistics, showcasing that 51.3% of vehicles sold that year were Level 0. By 2018 this number had dropped to 24% and by 2023 it was down to just 7%, with a massive shift towards Level 1 coming in at 71%.

- **Level 1**: Driver Assistance. *The vehicle features a single automated system (e.g. it monitors speed through cruise control).*

The first commercially available adaptive cruise control (ACC) vehicles came from Chrysler in 1958 following the invention of the SpeedoStat [15]. They named this system the "Auto-Pilot" and vehicles equipped with it were some of the first vehicles in history to achieve what would later be called Level 1 autonomy. Soon after, Cadillac adopted the technology for their own vehicles and dubbed it "Cruise Control", which since became the default term for the ACC technology, which is still used today, though usually with the term "adaptive" prepended.

As shown in the visualization in Figure 1, the steering wheel has a dashed-out look, and if the pedal were shown, they would be dashed out as well. This is to show that the human driver is not in control of every aspect of the DDT. Per the definition, Level 1 autonomy is defined as an autonomous system working on either the lateral or the longitudinal vehicle motion, which means that it can either control the steering or the acceleration of the vehicle. The human driver is still in control of the other aspect of the DDT. ACC is one of the most common systems elevating vehicles into this level. ACC can in this level not work with other systems, such as automatic steering. For vehicles that do both end up in Level 2.

- **Level 2**: Partial Automation. *Advanced driver-assistance system (ADAS). The vehicle can perform steering and acceleration. The human still monitors all tasks and can take control at any time.*

Level 2 autonomy is the first level where the vehicle can perform both steering and acceleration. This is done through a combination of systems, such as adaptive cruise control

[1] These italicised definitions are from the infographic shown on the Synopsys blog covering the topic [12].

(ACC) and lane-keeping assistance (LKA). The human driver is still in control of the vehicle and must monitor the environment at all times. This level is the most common level with Electric Vehicles (EVs), as all have ACC and LKA systems to a smaller or greater extent and capability. As visualized, the driver must always have at least one hand on the wheel at all times and be aware of the surrounding environment. This is the level where most of the current systems are at, such as Tesla Autopilot, Ford BlueCruise, and GM Super Cruise. Despite only being Level 2, Ford's BlueCruise has been allowed to establish "blue zones" in the EU, where drivers are allowed to remove the hands from the steering wheel when engaged [16].

At this level, the European version of the Tesla Autopilot belongs, where its American relative is climbing for level 4. This is despite the fact that Tesla Full Self-Driving (FSD) is legally classified as Level 2 in both regions, but in the US they operate at a higher level due to more permissive testing environments and less restrictive oversight [17]. Thus, laws and regulations create a functional separation at this level, as this is the level currently allowed in EU countries, with the single exception, as of 2022, being the Mercedes S-Class, allowed to operate in Level 3 under strict conditions [18]. This allowance has since been given to other manufacturers.

- **Level 3**: Conditional Automation. *Environmental detection capabilities. The vehicle can perform most driving tasks, but human override is still required.*

Level 3 autonomy is characterised by their ability to detect and act upon the environment by themselves. This means that the vehicle can perform most DDTs, but the human driver must still be able to take control at any time. This level is the first level where the vehicle can operate without human intervention in certain situations, such as highway driving in cases where overtaking a slow moving vehicle is possible. This is also the first level to rely on automated systems to monitor the environment.

An important note, as the level keeps rising, is what the SAE calls DDT Fallback. This refers to the action of taking over the DDT as a user, in the event of a DDT performance-relevant system failure or upon operational design domain (ODD) exit. ODD essentially means that the vehicle can only operate in certain environments, small and defined, like specific motorways, or broad, like an entire trip. For instance, the ODD for Ford's BlueCruise is when driving in the predefined "blue zones", meaning that the user must take over when leaving them. While Level 3 autonomy has limited spread, some manufacturers are allowed to operate at this level, with Germany laying out the groundwork for the road to Level 4 autonomy [18].

- **Level 4**: High Automation. *The vehicle performs all driving tasks under specific circumstances. Geo-fencing is required. Human override is still an option.*

The main difference between Level 3 and Level 4 is the fact that the vehicle can operate without human intervention in certain situations. This means that the vehicle can perform all driving tasks under specific circumstances, such as highway driving or in urban environments. This is the referred to as geo-fencing, where the vehicle can only operate in certain areas. As visualized in Figure 1, the steering wheel is now only outlined by dashes, meaning that the human driver is not in control of the vehicle at all times, but can take over if needed or required by the vehicle.

There are currently no commercially available vehicles that reach this level of autonomy, but some manufacturers have developed and deployed what are essentially Level 4 systems. For instance, Waymo and Cruise have developed systems that can operate in certain areas, such as San Francisco and Phoenix. These systems are available to customer use, but they are developing vehicle anyone can buy. Both Waymo and Cruise act as a taxi service, where users can request a ride through an app. As mentioned, however, these systems are only allowed to operate in specified, geo-fenced areas, and they are not allowed to operate outside these areas. This is what is keeping these technologies at Level 4 instead of Level 5. They gather massive amounts of detailed data on the cities they operate in and train their systems on this data, meaning they are great in their respective areas, but they are not able to operate outside of them. Furthermore, Tesla unveiled their own robotaxis at the "We, Robot" event, featuring vehicles without steering wheels or pedals. While not commercially available either, they still present a glimpse of the future of Level 4 autonomy, potentially even Level 5.

- **Level 5**: Full Automation. *The vehicle performs all driving tasks under all conditions. Zero human attention or interaction is required.*

Level 5 is the final level of the taxonomy presented by SAE. At this level, the ODD is unlimited, whereas every earlier level has been limited. At this level, both the DDT and DDT Fallback are fully handled by the vehicle. This means that the human user relinquishes the role of driver, and becomes a passenger, purely. The vehicle can operate in any environment, and the human user does not need to be aware of the environment at all. Not only does this mean that the vehicle can operate in any environment, but it also means that the vehicle can operate in all conditions. If at any point during a trip, the system can't figure out how to navigate and for any reason requires human intervention, then it is not Level 5. This is neatly presented in the visualization in Figure 1, where there is no steering wheel in the image and the once-drive-now-passenger is sitting with a book, completely unaware of the environment.

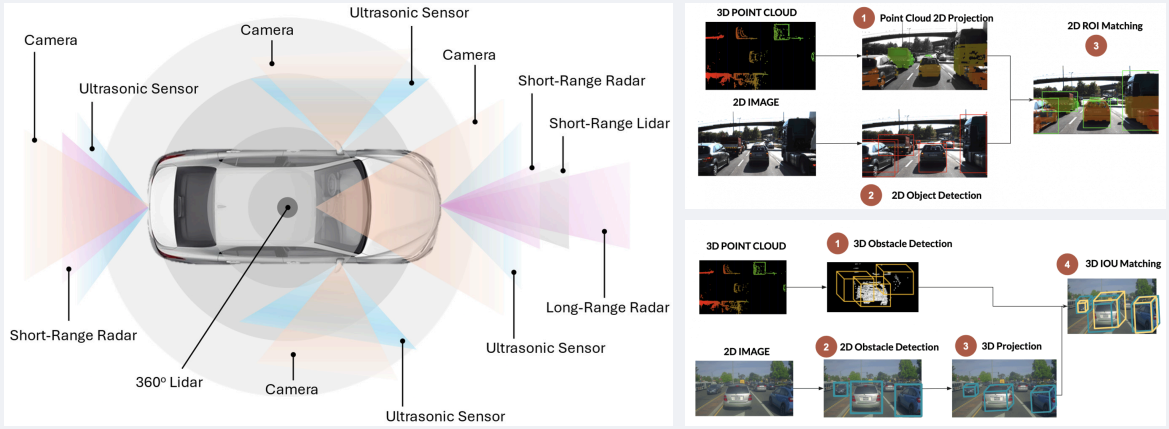
This level is not available anywhere in the world yet, and it is still believed to be at least a decade away before being commercially available to consumers [19]. Level 5 is most commonly the level depicted in futuristic sci-fi movies and games. Movies like "Minority Report" (2002), "Total Recall" (1990), and "Knight Rider" (1982) feature Level 5 AVs used for various tasks, such as autonomous driving, autonomous taxi services, and even fighting crime, respectively. Level 5 vehicles with personalities is a common trait in sci-fi, with the Delamain AI cab service from the game "Cyberpunk 2077" (2020) being a prime example.

* * *

To achieve Level 3 autonomy and higher, the vehicle must be able to understand the surrounding environment. This is done by various means and is done largely the same across the industry, with a few outliers. The most common way to understand the environment is through the use of sensors, such as cameras, radar, and lidar. Cameras are a fairly recent addition to the AV technology stack. They are used to detect and understand the environment around the vehicle, such as detecting pedestrians, cyclists, and other vehicles [20]. This requires a fair amount of computing power and efficiency, as the vehicle must be able to process the data from the cameras in real-time. The technology required is relatively

new, which is why it is only in recent years that cameras have become a mainstay in AVs, at least in the context of autonomous driving. Drivers have had rearview and Bird's Eye View (BEV) cameras for many years, however, these were only to assist said driver. Before cameras, it was common for vehicles to be equipped with radar and lidar systems. These systems offer a much higher resilience to adverse conditions, such as rain and fog, but they are also much more expensive. Lastly, ultrasonic sensors are used to detect objects in close proximity to the vehicle, such as when parking. These sensors are not used for understanding the broader environment, but they are used for detecting objects around the vehicle. A typical configuration of these sensors is shown in Figure 2a.

These systems are often used in conjunction with one another to create a single, coherent picture of the environment. This is what is known as sensor fusion, where the deluge of data from the various sensors is combined. Not only does this create a coherent picture, but it also draws on the strengths and weaknesses of each of the aforementioned technologies. For example, a camera has great spatial resolution and little noise, but is poor at estimating velocity and distance where radar and lidar shine, respectively [21]. There are 3 types of sensor fusion classifications: Low-level, mid-level, and high-level. Low-level is considered early fusion, and the others late fusion.



(a) (b)

Figure 2. (a) shows a typical sensor configuration, showcasing the large amount of different sensors on a single vehicle. (b) shows the difference between Early (top) and Late (bottom) fusion. Image source: (a) Yeong *et al.* [22], (b) Think Autonomous [23].

Low-level fusion is considered early fusion, as it combines the raw data streams, i.e. camera pixels, lidar points, etc., from the sensors before any processing is done. While this method retains the most information possible from the sensors, it is also very computationally heavy [22]. Mid-level fusion is considered late fusion, as it combines the intermediate representations. For instance, the vehicle's camera and radar might recognize a vehicle, then it recognizes those two representations as representing the same object. Finally, high-level fusing is also considered late fusion, as they combine the mid-level fusions with positional tracking. This is often through the use of probabilistic filters.

The most popular of these is the Kalman filter [24]. It is a powerful recursive algorithm used in sensor fusion to estimate the state of a dynamic system by combining measurements from multiple, potentially noisy sensors over time. It consists of two main steps: the predic-

tion step and the update step. In the prediction step, the Kalman filter uses a mathematical model of the system to predict the next state based on the current state and control inputs:

$$\hat{x}_{k|k-1} = F_k \hat{x}_{k-1|k-1} + B_k u_k \tag{1}$$

where $\hat{x}$ is the state estimate, F_k is the state transition model, B_k is the control input model, and u_k is the control input. Then, in the update step, the filter combines the predicted state with the new measurement to produce an updated estimate:

$$\hat{x}_{k|k} = \hat{x}_{k|k-1} + K_k\left(z_k - H_k \hat{x}_{k|k-1}\right) \tag{2}$$

where K_k is the Kalman gain, z_k is the measurement, and H_k is the observation model. By iteratively applying these steps, the Kalman filter fuses information from different sensors, accounting for their noise characteristics, to produce a statistically optimal estimate of the system's state. This sensor fusion is typically more accurate than relying on a single sensor, as it combines the strengths of each sensor type. This helps the vehicle understand its immediate environment, and with the help of technologies like Vehicle-to-Everything (V2X), it can further understand the world around it. With all of these technologies in a vehicle's stack, it can make informed decisions and navigate dynamically complex environments effectively.

With this robust and detailed understanding of the world, many different control methodologies have been developed to control vehicles and robots, designated as the control problem. The most common and simple method is the Proportional-Integral-Derivative (PID) controller. Briefly, PID controller always tries to minimize the error between the measured state and a desired state. This error is then subject to three different operations: proportional, integral, and derivative, each multiplied by some scalar. Each property influences things like speed, overshoot, and stability. A more complex method is the Model Predictive Control (MPC) method. This method uses a model of the system to predict the future states and optimize the control inputs over a finite time horizon. It is particularly useful for systems with constraints, such as steering angle, acceleration, and velocity. MPC is often used in conjunction with path planning, where the vehicle must follow a specific path while avoiding obstacles.

While path-planning is covered in Section 3, how AVs achieve this will briefly be covered, for with this robust representation of the environment and the control methodologies, the vehicle can make informed decisions and navigate dynamically complex environments effectively. Today, this is largely done with the use of DL, but more manual methods exist. These methods are less generalizable than DL ones, but are robust in their own ways. One such method is built up around the concept of a state machine, known as rule-based methods. These are scenario specific rules that are applied at predefined objectives, such as slowing down automatically when seeing a red light. For DL approaches, data-driven approaches are common. Behavioural cloning and imitation learning are common when you have massive amounts of driving data. These methods are trained on the data, and learn to mimic the behaviour of the human driver. This is done by training a Neural Network (NN) to predict the control inputs based on the sensor data. DL approaches have

witnessed massive growth in the last decade, and are now the most common methodology for AVs.

2.2 Deep Learning ✓

The groundwork for the modern Artificial Intelligence (AI) we see today, has been under development since the 1940s. In 1943, Warren McCulloch and Walter Pitts proposed the artificial neuron [25]. Their goal was to create a model that acted like the human brain, in that brains are made up of constantly firing neurons. They proposed NNs can be modelled using a logical calculus based on the "all-or-none" firing principle of neurons. They demonstrate that neural activations correspond to logical statements. This provides a systematic way to analyse and predict neural behaviour, but also lays the theoretical groundwork for later advances in computational neuroscience and AI. Already in 1951, Minsky and Edmonds presented the Stochastic Neural Analog Reinforcement Calculator (SNARC), which consisted of 40 artificial neurons [26]. As will become the norm, this network was trained by adjusting the strengths of the connections between neurons based on the outcomes of the previous trials.

Another concept, introduced early, was what would come to be known as reinforcement learning. Arthur Samuel, often called the father of machine learning, introduced the Samuel Checkers-Playing program in 1952, which learned the game of checkers via self-learning, improving its skills over time by playing many more games than a human ever could [27]. The term AI was coined in 1956, but it was not until 1958 Frank Rosenblatt developed the first Artificial Neural Network (ANN), which he named the perceptron, which was a single-layer NN that could learn to classify patterns [28]. Building on the early breakthroughs, the following decades saw an expansion of ideas that sought to overcome the limitations of single-layer NNs. Researchers began experimenting with multi-layered architectures and early variants of backpropagation [29]. In general, this period laid important groundwork for the following decades in the research and development of AI and, subsequently, DL.

One of the most important breakthroughs came in the form of the rediscovery and refinement of backpropagation [30]. This sparked the renaissance in NN research, as interest in learning from data was renewed. This breakthrough enabled the development of DL architectures, which could learn complex patterns and representations from large datasets. The following decades saw the integration of statistical learning methods with NN architectures that further enhanced their accuracy and robustness. What is considered the first point of contact with AI, came in 2012 with the release of AlexNet [31]. This model was capable of classifying images with a high degree of accuracy, paving the way for modern learning algorithms and DL architectures.

Looking at the highlights of the previous paragraphs, the most important takeaways are as follows. McCulloch and Pitts stated that "At any instant a neuron has some threshold, which excitation must exceed to initiate an impulse." which, as will be presented, is exactly how modern machines are taught to think and learn. SNARC used a method of

adjusting the strengths of the connections between neurons based on the outcomes of the previous trials. Samuel's Checkers-Playing program used a method of self-learning, which is the basis of modern reinforcement learning. Rosenblatt's perceptron was the first ANN, introducing the concept of layers in a NN, later expanded to the modern Multilayer Perceptron (MLP).

The MLP is the most common and simple type of NN used in DL. It consists of an input layer, one or more hidden layers, and an output layer. Each layer consists of a number of neurons, which are connected to the neurons in the previous and next layers. The connections between the neurons are weighted, and these weights are adjusted during training to minimize the error between the predicted output and the actual output. The training process is done using backpropagation, which is a method for calculating the gradients of the weights with respect to the error. This is done by applying the chain rule of calculus to calculate the gradients of each weight in the network.

This process will now be broken down into its components, with the intent of creating a clear understanding of how MLPs work, and how they are trained. This will act as a springboard for understanding the more complex architectures and methods used in this work.

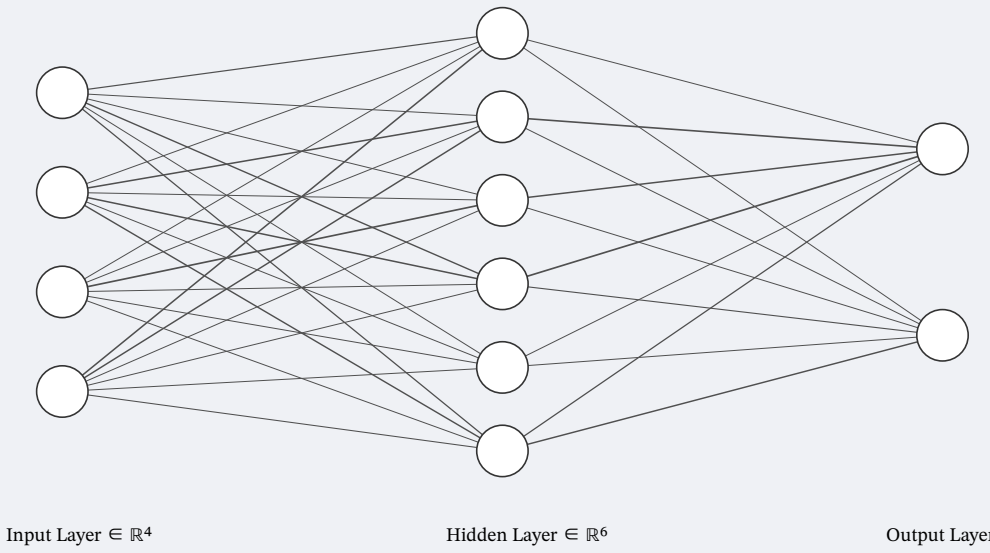


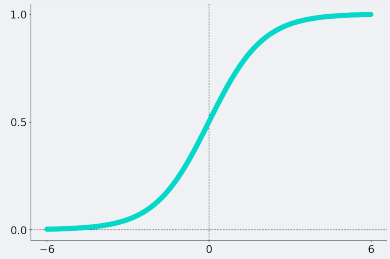
Figure 3. Multi-Layer Perceptron (MLP).

MLPs are composed of a number of layers, each consisting of a number of neurons. Each neuron is connected to the neurons in the previous and next layers. The connections between the neurons are weighted (as shown by the different opacities of the connections in Figure 3). MLPs are often what is called "fully connected", meaning that each neuron in a layer is connected to every neuron in the next layer. This is done to allow for the maximum amount of information to be passed between the layers.

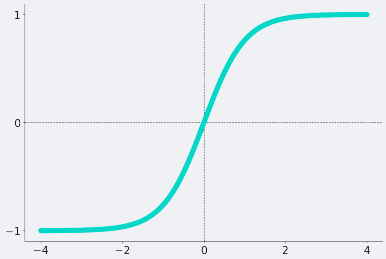
To start from the smallest possible unit, the neuron, it is important to understand how they work. During a forward pass through a NN (giving it some input feature), the value of each neuron is calculated as follows:

$$z = \boldsymbol{w}^T \boldsymbol{x} + b = \sum_i \boldsymbol{w}_i x_i + b \tag{3}$$

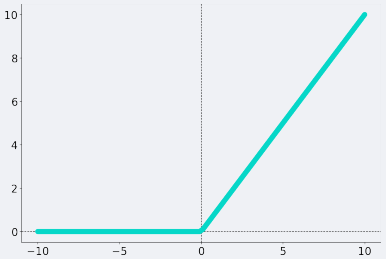
where $\boldsymbol{w}$ is the weight vector, $\boldsymbol{x}$ is the input vector, and b is the bias term. The bias term is a constant that is added to the weighted sum of the inputs. This allows the neuron to learn a threshold value, which is important for learning complex patterns. The output of the neuron a is then calculated using an activation function $a = \varphi(z)$, which introduces non-linearity into the model. Three of the most common activation functions φ are shown below:



$\sigma(z) = \frac{1}{1+e^{-z}}$



$\tanh(z) = \frac{e^z - e^{-z}}{e^z + e^{-z}}$



$\mathrm{ReLU}(z) = \max(0, z)$

The first of the three is the sigmoid function. It maps the input to a value between 0 and 1, making it particularly useful for binary classification tasks. tanh scales the output to a value between -1 and 1, often leading to faster convergence. The last one, ReLU, is the most common activation function used in DL today. It is defined as $\mathrm{ReLU}(z) = \max(0, z)$, meaning that it outputs the input directly if it is positive, and 0 otherwise. This function is computationally efficient and helps mitigate the vanishing gradient problem, which can occur with sigmoid and tanh functions. The vanishing gradient problem occurs when the gradients of the weights become very small, making it difficult for the model to learn.

Another important activation function is the softmax function, which is used in the output layer of a MLP for multi-class classification tasks. It converts the raw output logits into probabilities by exponentiating each logit and normalizing them:

$$\mathrm{softmax}(z_i) = \frac{e^{z_i}}{\sum_j e^{z_j}} \tag{4}$$

which ensures that the sum of the probabilities is equal to 1. Now, here is an example of a forward pass through the MLP. First, the input features $\boldsymbol{x}$ is passed through the hidden layer:

$$\begin{aligned} \boldsymbol{z}^{(1)} &= \boldsymbol{W}^{(1)} \boldsymbol{x} + b^{(1)} \\ \boldsymbol{a}^{(1)} &= \varphi\big(\boldsymbol{z}^{(1)}\big) \end{aligned} \tag{5}$$

where $\boldsymbol{W}^{(1)}$ is the weight matrix for the first layer, $\boldsymbol{z}^{(1)}$ is the weighted sum of the inputs, and $\boldsymbol{a}^{(1)}$ is the output of the first layer. The output of the first layer is then passed through the output layer:

$$\begin{aligned} \boldsymbol{z}^{(2)} &= \boldsymbol{W}^{(2)}\boldsymbol{a}^{(1)} + b^{(2)} \\ \hat{y} &= \text{softmax}\big(\boldsymbol{z}^{(2)}\big) \end{aligned} \tag{6}$$

where, in this example, the output $\hat{y}$ is put through a softmax activation to get probabilities on the output. Depending on the task of the NN, different loss functions are used. Loss functions are used to measure the difference between the predicted output and the actual output. For the task of regression (predicting a continuous value), the most common loss function is the Mean Squared Error (MSE) loss, which is defined as:

$$L(\hat{\boldsymbol{y}}, \boldsymbol{y}) = \frac{1}{N}\sum_{i}(\hat{y}_i - y_i)^2 \tag{7}$$

where N is the number of samples, $\hat{y}$ is the predicted output, and y is the actual output. For classification tasks, the most common loss function is the Cross-Entropy loss, which is defined as:

$$L(\hat{\boldsymbol{y}}, \boldsymbol{y}) = -\sum_{i} y_i \log(\hat{y}_i) \tag{8}$$

which calculates the dissimilarity between the predicted distribution $\hat{\boldsymbol{y}}$ and actual distributions $\boldsymbol{y}$. The binary version, called Binary Cross-Entropy (BCE), is also commonly used for binary classification tasks, and is defined as:

$$L(\hat{\boldsymbol{y}}, \boldsymbol{y}) = -\sum_{i}(y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i)) \tag{9}$$

where y_i is the actual label, and $\hat{y}_i$ is the predicted probability of the positive class. Once the forward pass is complete and the loss L is calculated using a function like MSE or Cross-Entropy, the goal of training the neural network is to minimize this loss. This done through what is called gradient descent. Gradient descent is a method for mathematical optimization, meaning, by doing certain operations, a network's parameters can reach minima, where the error, or loss, is the smallest it can be. This is done by calculating the gradients of the loss with respect to the parameters, and updating the parameters in the opposite direction of the gradients. The goal is to reach what is called local minima, where the loss is minimized. These minima symbolize the best (smallest) loss that a network can achieve, depending on its weights and biases.

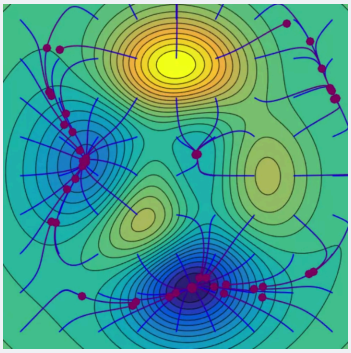


Figure 4. Gradient descent. Image source: WikiPedia [32]

Formally, this is known as backpropagation, which is a method for calculating the gradients of the loss with respect to the parameters. This is done by applying the chain rule of calculus to calculate the gradients of each weight in the network. So, the way to find the gradient for a weight w_{ij} connecting neuron i to neuron j in the next layer is to calculate the partial derivative of the loss with respect to that weight:

$$\frac{\partial L}{\partial w_{ij}} = \frac{\partial L}{\partial z_j}\frac{\partial z_j}{\partial w_{ij}} \tag{10}$$

where

$$z_j = \sum_i w_{ij}x_i + b_j \quad \Rightarrow \quad \frac{\partial z_j}{\partial w_{ij}} = x_i \tag{11}$$

meaning the gradient is

$$\frac{\partial L}{\partial w_{ij}} = \delta_j x_i \tag{12}$$

where δ_j is the error term for neuron j, calculated using one of the aforementioned loss functions. With this, it is now time to optimize the weights. This is done by using one of the optimization algorithms available. Two of the most commonly used optimizers are Stochastic Gradient Descent (SGD) and Adam. SGD is a simple and effective optimization algorithm that updates the weights using the gradients calculated during backpropagation. The update rule for SGD is:

$$w_{ij}^{(t+1)} = w_{ij}^{(t)} - \eta\frac{\partial L}{\partial w_{ij}} \tag{13}$$

where η is the learning rate, which controls the step size of the update. The learning rate is often a very small value, as to help NNs converge to local minima slowly. A fair bit more involved and extremely influential is the Adam optimizer. Adam means Adaptive Moment Estimation, and is an adaptive learning rate optimization algorithm. It combines the advantages of two other extensions of SGD: momentum and RMSProp. It maintains two moving averages for each parameter: one for the gradients (first moment) and one for the squared gradients (second moment). It also includes a bias-correction mechanism to counteract the initialization bias of the first moment estimates. The update rules for Adam at time t are as follows:

1. Compute the gradients: $g_t = \frac{\partial L}{\partial w_t}$
2. Update the biased first moment estimate: $m_t = \beta_1 m_{t-1} + (1-\beta_1)g_t$
3. Update the biased second moment estimate: $v_t = \beta_2 v_{t-1} + (1-\beta_2)g_t^2$
4. Bias-correct the moment estimates: $\hat{m}_t = \frac{m_t}{1-\beta_1^t}$, $\hat{v}_t = \frac{v_t}{1-\beta_2^t}$
5. Update the parameters: $w^{(t+1)} = w^{(t)} - \eta\frac{\hat{m}_t}{\sqrt{\hat{v}_t}+\varepsilon}$

where β_1 and β_2 are the decay rates for the first and second moment estimates, respectively. These are typically very close to 1, commonly set to 0.9 and 0.999. ε is a small constant added to prevent division by zero, and is typically set to 10^{-8}. Both m_0 and v_0 are initialized as 0. While a constant learning rate η is often good enough, too small of a constant η may result in the models being stuck in the closest local minima, which is likely not the best. A higher η is not exactly the way to combat this, as the training may become highly unstable, if the changes in weights are too large. This is where learning rate scheduling comes in. Learning rate scheduling is a technique used to adjust the learning rate during training.

There are several different types of learning rate schedules, such as step decay, exponential decay, and cosine annealing.

Step decay is the simplest, where the learning rate is reduced by a factor every few epochs. Exponential decay is largely the same, but it is constantly decreasing the learning rate each epoch. First, the step decay defines the learning rate as:

$$\eta^{(t)} = \eta_0 \cdot \gamma^{\left\lfloor \frac{t}{\text{step-size}} \right\rfloor} \tag{14}$$

where t is the epoch, η_0 is the initial learning rate, and γ is the decay factor. $\lfloor \cdot \rfloor$ denotes the floor function. Step decay is simple and can help a model settle into local minima, but it is not very flexible. The non-smooth nature of the floor function used in the step decay can lead to abrupt changes in the learning rate, which can cause instability. Furthermore, it requires careful tuning of the step size and decay factor. Exponential decay combat some of these disadvantages by using a continuous decay function. The learning rate is defined as:

$$\eta^{(t)} = \eta_0 \cdot e^{-k \cdot t} \tag{15}$$

where k is the decay constant. This allows for a smoother and more gradual decrease in the learning rate, which can help the model converge more effectively. However, it still requires careful tuning of the decay constant k. It is, however, still a smooth uniform decay, which means it doesn't help the model potentially escape poor local minima. This is where cosine annealing comes in. Cosine annealing is a more advanced learning rate schedule that uses a cosine function to adjust the learning rate. The learning rate is defined as:

$$\eta^{(t)} = \eta_{\min} + \frac{1}{2}(\eta_{\max} - \eta_{\min})\left(1 + \cos\left(\frac{\pi t}{T}\right)\right) \tag{16}$$

where $\eta_{\min}$ is the minimum learning rate, $\eta_{\max}$ is the maximum learning rate, and T is the total number of epochs for the annealing cycle. This cosine function provides a periodic transition between high and low learning rates. When coupled with warm restarts (occasionally setting $\eta^{(t)} = \eta_{\max}$), this scheduler helps the optimizer escape local minima and thereby increase performance. The downsides are that it is more complex than the other schedulers, and it requires careful tuning of more parameters than both step and exponential decay.

Finally, some other common concepts w.r.t training are these:

- Regularization: A technique used to prevent overfitting by adding a penalty term to the loss function. The most common form of regularization is L2 regularization, which adds a term to the loss function that is proportional to the square of the weights. This encourages the model to learn smaller weights, which can help prevent overfitting. L2, or weight decay, is defined as: $L_{\text{total}} = L_{\text{data}} + \lambda \left\| w \right\|_2^2$.
- Dropout: A technique used to prevent overfitting by randomly dropping out a fraction of the neurons during training. This forces the model to learn more robust features and prevents it from relying too heavily on any one neuron.
- Batch Normalization: A technique used to normalize the inputs to each layer in the network. First, the batch mean and variance are found:

$$\begin{aligned} \mu_B &= \frac{1}{m}\sum_{i=1}^{m} x_i, \\ \sigma_B^2 &= \frac{1}{m}\sum_{i=1}^{m} (x_i - \mu_B)^2 \end{aligned} \tag{17}$$

where m is the batch size. Then, the inputs are normalized:

$$\hat{x}_i = \frac{x_i - \mu_B}{\sqrt{\sigma_B^2 + \varepsilon}} \tag{18}$$

where ε is a small constant added to prevent division by zero. Finally, the normalized inputs are scaled and shifted:

$$y_i = \gamma \hat{x}_i + \beta \tag{19}$$

where γ and β are learnable parameters that allow the network to scale and shift the normalized output.

In summary, the architecture and training process of the MLP has been presented in detail. By starting with creating an understanding of how each neuron worked, to how they are connected, and how the entire MLP is trained, a solid understanding of the MLP has been created, and generally how to train NNs. The forward and backward propagation was presented, as well as the loss functions and optimizers. The most common activation functions were presented, as well as the most common learning rate schedulers. With all of these methods, it is possible to train our MLP at a task we desire. However, the MLP is too simple of an architecture to learn any really complex tasks. This is where the more complex architectures and methods come in.

* * *

Before moving on to the complexities introduced in the subfield of computer vision, it is important to note that DL consists of three main learning paradigms. The first, and most common, of these is supervised learning. Supervised learning is concerned with training models using labelled data. This means that the model is trained on a dataset where the input features are paired with the corresponding output labels, i.e. the dataset consists of pairs (x, y) where x is the input features and y is the corresponding labels, often referred to as the ground truth labels. With each of these pairings, the training process involves minimizing a loss function, as described earlier, between the input features x and the ground truth labels y. Therefore, this learning paradigm is often used for image classification tasks, speech recognition, and natural language processing tasks. In other words, supervised learning is for when you know how you want the outcome to look by coming as close to the ground truth as possible.

Alternatively, you might not know exactly how your output should look. This is where unsupervised learning comes in. Unsupervised learning is concerned with training models using unlabelled data. This means that the model is trained on a dataset where the input features are not paired with any output labels. The goal of unsupervised learning is to learn the underlying structure of the data, such as clustering similar data points together or reducing the dimensionality of the data. This learning paradigm is often used for tasks

such as clustering, anomaly detection, and dimensionality reduction. Furthermore, this paradigm is typically used when training generative models. Generative models are models that learn to generate new data points that are similar to the training data. This is done by learning the underlying distribution of the data, and then sampling from that distribution to generate new data points. This task is most often seen in image generators, such as Generative Adversarial Networks (GANs).

Finally, Reinforcement Learning (RL) is a learning paradigm where a model learns through interactions with the environment. The model learns to take actions in an environment to maximize a reward signal. This is done by learning a policy, which is a mapping from states to actions, and a value function, which is a mapping from states to expected rewards. The goal of reinforcement learning is to learn a policy that maximizes the expected cumulative reward over time. This learning paradigm is often used for tasks such as game playing, robotics, and autonomous driving. Popular examples include AlphaGo [33] beating the world's greatest Go player, and OpenAI Five [34] beating the world champions in Dota 2.

These three paradigms broadly cover the methods used to train different models for different tasks. Each offers their strengths and weaknesses. Labelling data can be a time-consuming task, but create faster more targeted training, where unlabelled data and environmental interaction may reach beyond the performance initials goals. Each paradigm has their own objectives when training, from minimizing some prediction error to discovering hidden patterns to maximizing some cumulative reward. This also means that each paradigm lend themselves to achieve specific goals, such as classification, anomaly detection, and decision-making, respectively. Computationally, they also offer different complexities: supervised learning can vary a lot depending on the task, but is typically moderately complex; unsupervised learning is often more complex, as the model must learn the underlying structure of the data; and reinforcement learning is often the most complex, as it requires the model to learn through iterative interactions with the environment.

These paradigms form the backbone of machine learning techniques used across various domains, including computer vision. Specifically, supervised learning has driven significant advancements in tasks like image classification, where labelled datasets guide models to recognize intricate visual patterns. One landmark achievement highlighting the potential of supervised methods in computer vision is AlexNet, which is widely regarded as the starting point of modern artificial intelligence.

2.2.1 Computer Vision ✓

AlexNet is widely considered to be the first point of contact with what we classify as AI today [31]. It was an image classification model that was trained on the ImageNet dataset, and was capable of classifying images with a high degree of accuracy. This was the first NN to clearly pass other machine learning methods, such as kernel regression and support vector machines, in image classification tasks. This was a major breakthrough in the field of computer vision, opening the field to more researchers by proving the capabilities of NNs. Since its release, the interest in computer vision tasks has exploded, and is thus deployed in

a wide range of applications, such as image classification, object detection, image segmentation, pose estimation, and image generation.

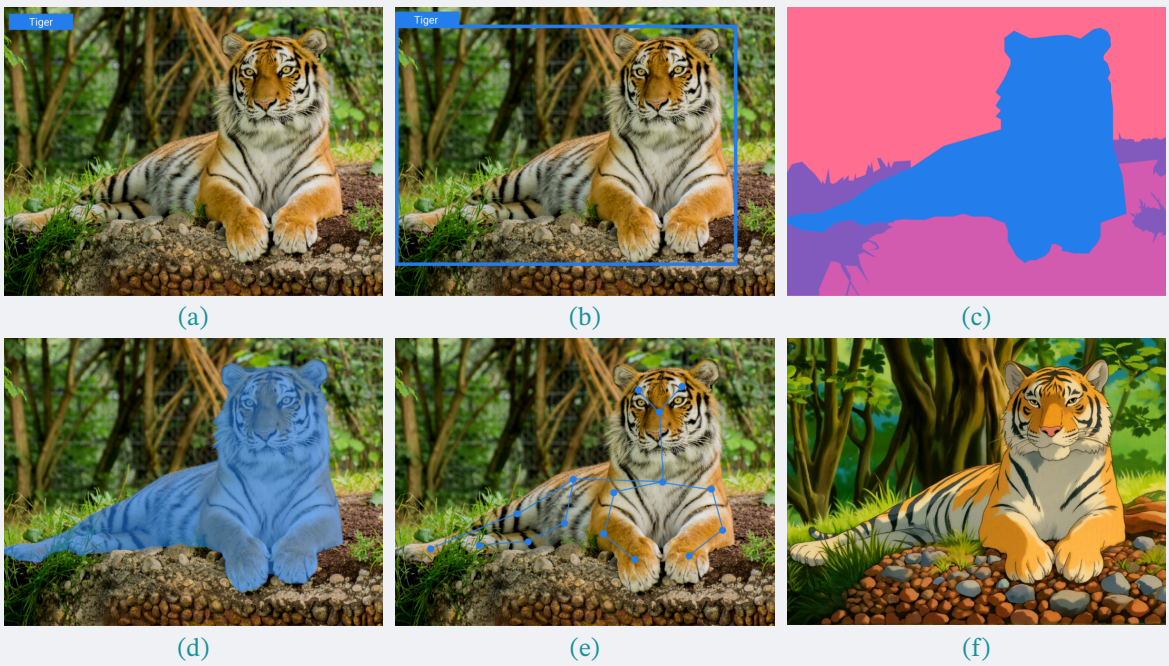


Figure 5. (a) Image classification. (b) Object detection. (c) Semantic segmentation. (d) Instance segmentation. (e) Pose estimation. (f) Image generation. Image source: (a-e) viso.ai [35] (f) Altered with OpenAI 4o [36].

The image classification task is the simplest of the computer vision tasks. It is concerned with classifying an image into one of a number of classes. This task, which was AlexNet's main task, is shown in Figure 5a. As shown, it is concerned with labelling the entire image as one class. It should be noted "class" means the name of a group of similar things, like the example shows the class of the image is "tiger". This idea of classes, becomes important when moving on to the other CV tasks. The next task is object detection, which is concerned with detecting and localizing objects in an image. This task is shown in Figure 5b. It is concerned with not only classifying the image, but also localizing the objects in the image. This is done by drawing bounding boxes around the objects in the image, and classifying them. Note that there may be multiple instances of the same objects in the image, each requiring their own bounding box for correct classification.

The next task is semantic segmentation, which is concerned with classifying each pixel in the image. This task is shown in Figure 5c. It is concerned with classifying each pixel in the image into one of a number of classes. For this task, the model needs to be trained to not just identify the objects in the image, but also to identify the boundaries of the objects. This is done by creating a class label or class mask in the training data. A class mask is an image where each pixel is assigned a class label, and as mentioned prior, this class mask is the ground truth for said image. Closely related is the task of instance segmentation, which is concerned with classifying each pixel in the image, but also differentiating between different instances of the same object. This task is shown in Figure 5d. The main difference between these two segmentation tasks, is the fact that semantic segmentation does not differentiate between different instances of the same object. This means that if there are two objects of the same class in the image, they will be classified as the same object. This is

not the case for instance segmentation, where each instance of the same object is classified as a different object.

The next task is pose estimation, which is concerned with estimating the pose of an object in the image. It can be split into two separate tasks of its own: estimation of the 3D position of an object within an image and estimating the pose of the isolated object. The latter of these is shown in Figure 5e, where a rigging skeleton is drawn over the object to indicate its pose. This is done by estimating the position of the joints in the image, and drawing lines between them. The former task is concerned with estimating the 3D position of an object in the image, often achieved through a cascade of methods, such as 3D reconstruction, depth estimation, and multi-view stereo techniques.

The final task is image generation, which is concerned with generating new images from a given input. This task is shown in Figure 5f, where the image of the tiger has been altered to mimic a specific art style. This is a very common usage of image generators, but they can also create images from the ground up. In GANs, this is done by training the model on a dataset of images, and then generating new images that are similar to the training data. Other methods like diffusion iterate over noisy images, slowly refining them to create a new image based on some description. This is done by training the model to learn the underlying distribution of the data, and then sampling from that distribution to generate new images.

With these tasks in place, the methods for which either is achieved will now be presented. AlexNet consisted mainly of convolutional layers, with the occasional pooling layer. Convolution is a very important mechanic within the field of computer vision, as it allows for models to gain an understanding of the features within an image.

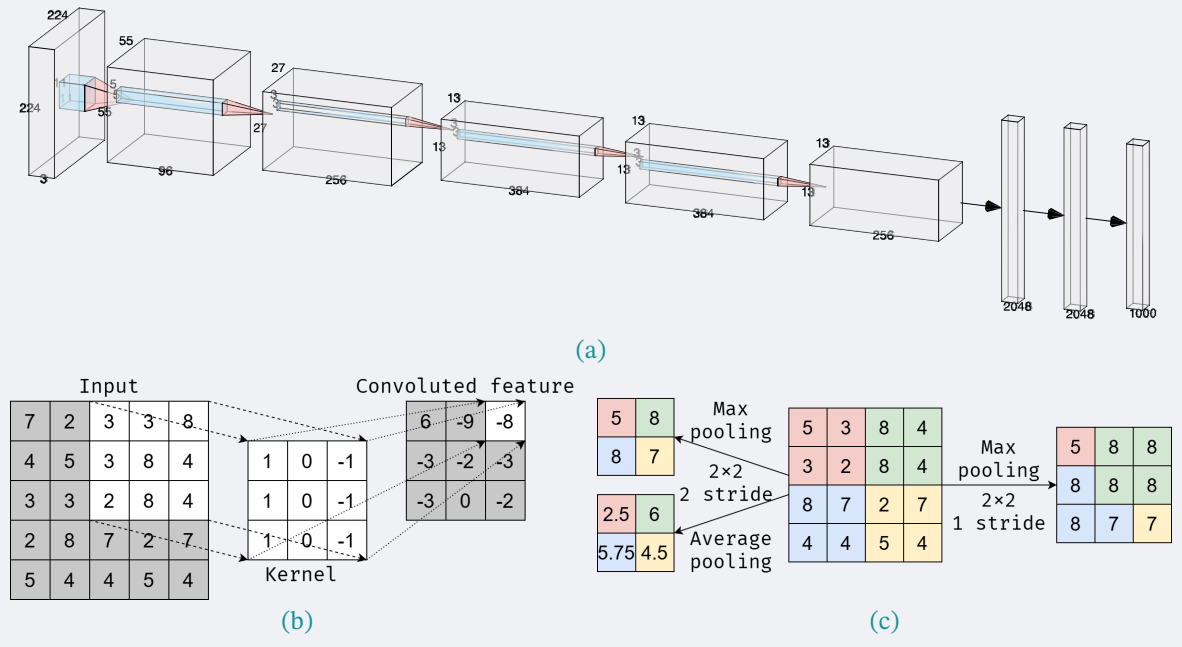


Figure 6. AlexNet architecture.

Convolution is achieved by convolving a kernel over the input image. This is done by sliding the kernel over the image, and at each position, calculating the dot product between the kernel and the image. This is done by multiplying each element in the kernel with

the corresponding element in the image, and then summing the results. The result of this operation is a new image, called a feature map. This map is typically smaller than the original image, as the kernel only stays within the bounds of the image itself. This can be bypassed by padding the input image with zeros, which allows the kernel to convolve along the very outer edge of the image. For input images that often containing three channels (RGB), the kernel is convolved with each channel separately, or, rather, the kernel is three-dimensional. This outputs one channel, which is why many different kernels are used. As seen in Figure 6a, the first layer of AlexNet after the input layer consists of 96 channel, with the next being 256 channels.

Other than padding, other factors come into the usage of convolution kernels. The stride is the number of pixels the kernel is moved at each step. A stride of 1 means that the kernel is moved one pixel at a time, while a stride of 2 means that the kernel is moved two pixels at a time. This can be used to reduce the size of the feature map, as the kernel will skip some pixels in the image. This is also seen in the figure, where the kernel is a large 11×11 kernel with stride 4. This results in the output being significantly smaller than the input image. Next, the kernel size is the size of the kernel itself, and is typically a small odd number, such as 3, 5, or 7. The kernel size is important, as it determines the size of the receptive field of the model. The receptive field is the area of the input image that the kernel is able to see at any given time. A larger kernel size means a larger receptive field, which allows the model to learn more complex features. However, a larger kernel size also means more parameters to train, which can lead to overfitting.

Padding, stride, and kernel size are also relevant to the other operation incorporated into AlexNet: pooling. Pooling is a downsampling operation that reduces the size of the feature map. This is done by taking the maximum or average value of a small region in the feature map, and using that value as the new value for that region. This is done by sliding a pooling kernel over the feature map, and at each position, calculating the maximum or average value of the region covered by the kernel. The result of this operation is a new feature map, which is smaller than the original feature map.

These concepts of stride and pooling are illustrated in Figure 6c. The size of the pooling kernel is often the same as the stride, meaning the kernel does not encompass any overlapping pixels. As shown with the right-hand side example, the stride is 1, which results in overlapping regions. This is a rather undesired effect, so a stride equal to the kernel size is often used. This is shown in the left-hand side example, where the stride is 2, and the pooling kernel is 2×2. This results in a non-overlapping pooling operation, which is often desired. The most common pooling operations are max pooling and average pooling. Max pooling takes the maximum value of the region covered by the kernel, while average pooling takes the average value of the region covered by the kernel. Examples of both are shown in Figure 6c on the left-hand side of the input.

Finally, a few more components of the AlexNet architecture are worth mentioning. After each convolutional layer, a ReLU activation function is applied. This is done to introduce non-linearity into the model, which allows it to learn more complex features. The final layer of AlexNet is a fully connected layer, which is used to classify the image into one of the classes. This is done by flattening the feature map into a vector, and then passing it

through a series of fully connected layers. The final output is then passed through a softmax activation function, which converts the output into probabilities for each class. AlexNet utilizes dropout for regularization, as to prevent overfitting. This is done by randomly dropping out a fraction of the neurons during training, which forces the model to learn more robust features. Furthermore, to prevent overfitting, during training, the input images are put through augmentation. This will be covered in greater detail in Section 4.3.2. Finally, AlexNet also uses Local Response Normalization (LRN) layers. It works by normalizing the activations of neurons in a local region across the channel dimension.

Moving beyond the convolutional NN (CNN) architecture that is AlexNet, many other methods and architectures exist to help machines understand images. Closely related by the fact that it is also considered a CNN, is the U-Net architecture, which is widely used for image segmentation tasks [37]. U-Net is a fully convolutional network that consists of an encoder and a decoder. The encoder is responsible for downsampling the input image, while the decoder is responsible for upsampling the feature map back to the original size. This is done by using skip connections between the encoder and decoder, which allows the model to learn both low-level and high-level features. The upsampling, opposed to downsampling/pooling, is done by using transposed convolutions, which are the reverse of normal convolutions. Whereas normal convolution reduce patches to singular values, transposed convolution does the opposite; it takes a singular value and expands it to a patch using a kernel. This typically doubles the height and width of the feature map. Convolutions in general see heavy usage in CV tasks, as the kernel they use are learnable, meaning they are affected by backpropagation, meaning they can learn.

* * *

In 2022, ChatGPT hit the ground running, exploding into the public conscience and making AI a mainstream tool that everyone suddenly knew about [38]. This leap in Natural Language Processing (NLP) was made possible by the introduction of the Transformer architecture, which was introduced in 2017. In their landmark paper “Attention is All You Need”, Vaswani *et al.* [39] introduced the Transformer architecture, which revolutionized the field of NLP. The transformer architecture is based on the self-attention mechanism, which allows the model to weigh the importance of different words in a sentence when making predictions. To understand self-attention, imagine reading a sentence. When you read a word, you don’t just look at the word itself, but also at the surrounding words. In the sentence “The cat chased the mouse because it was fast”, the word “it” refers to “the mouse”. Self-attention allows a model to do the same computationally. The self-attention mechanism works by calculating how relevant a word in a sentence is to every other word in the same sentence. This allows the model to look at other parts of the input sequence and determine which parts are most important for predicting the next word. This mechanism is heavily utilized in the transformer architecture. Transformers can process all words in a sequence in parallel. This is a major advantage over recurrent NNs (RNNs), which process words sequentially. They typically consist of an encoder and a decoder, where the encoder processes the input sequence and the decoder generates the output sequence. Both use multiple layers stacked with self-attention and feed-forward networks. Thus, transformers helped technologies, like the Generative Pre-trained Transformers (GPTs), gain an incredible understanding of human language and explode into the public conscience.

Inspired by the self-attention mechanism and transformers, Dosovitskiy *et al.* [40] introduced the Vision Transformer (ViT) architecture in 2021. This extended the capabilities of transformers into the field of computer vision. To make transformers applicable to images, the input image is divided into fixed-size patches, which are then flattened and linearly embedded into a sequence of tokens. This is very much akin to the way words in a sentence are processed in NLP. This allows the transformer to treat visual data in a similar way to text data, enabling it to learn complex relationships between different parts of the image through the self-attention mechanism. The ViT architecture consists of a series of transformer blocks, each containing a multi-head self-attention layer and a feed-forward network. The output of the final transformer block is then passed through a classification head to produce the final output. So, the vision architecture typically consists of three main steps:

Patch embedding: An input image is divided into equally-sized non-overlapping patches, typically 16x16 pixels each. These patches are flattened into vectors and transformed into embeddings through a learned linear projection, producing fixed-dimensional patch embeddings. **Transformer encoder**: The embeddings go through multiple stacked layers of transformer encoder blocks, each comprising two sub-layers: a multi-headed self-attention mechanism and a feed-forward neural network. Multi-headed self-attention computes attention scores to weigh the relevance of each patch relative to all others simultaneously. To retain spatial information, positional embeddings are added to the patch embeddings before being input into the transformer layers. **Classification head**: The processed embeddings from the transformer layers include a special classification token (`[class]`), prepended to the sequence of patch embeddings. The `[class]` embedding captures a global representation of the image. This global embedding is then passed through a fully-connected neural network to produce class predictions for image classification tasks.

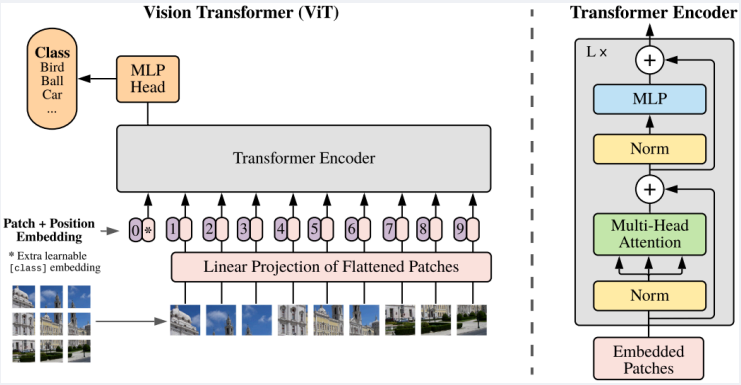


Figure 7. Vision Transformer architecture.
Image source: Dosovitskiy *et al.* [X]

Closely related to the ViT is the Swin Transformer, introduced by Liu *et al.* [41] in 2021. The Swin Transformer is a hierarchical transformer that uses a shifted windowing scheme to reduce the computational cost of self-attention. This is done by dividing the input image into non-overlapping windows, and then applying self-attention within each window. These windows are merged in the deeper layers of the network, achieving great highly accurate segmentation masks.

Now, a term I have not explained yet is “overfitting”. Overfitting is a common problem in machine learning, where the model learns the training data too well, and is unable to generalize to new data. This phenomenon is seen in the training and evaluation (or test or validation) graphs of a model’s training process. Early in the training process, the model’s performance on both the training and test sets improve as the model learns the task. However, as training continues, the model’s performance on the training set continues to improve, while the performance on the test set starts to degrade. This is a sign that the model is overfitting to the training data. This is visualized in Figure 8, where the training loss continues to decrease, while the validation loss starts to increase after training iteration 75.

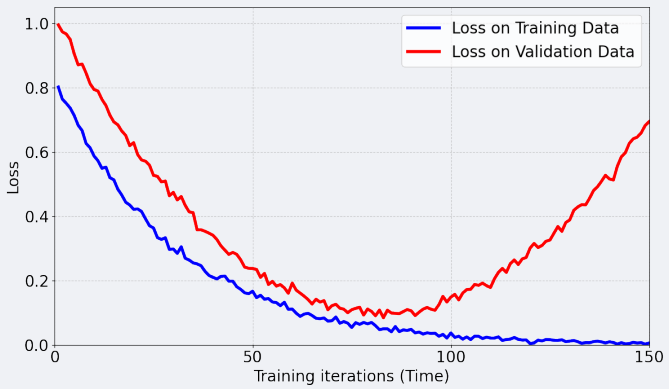


Figure 8. Training and evaluation loss plot.

To combat overfitting, several techniques are used. As presented already, dropout is commonly used to combat overfitting. This is done by randomly dropping out a fraction of the neurons during training, which forces the model to learn more robust features. Another common technique is early stopping, where the training process is stopped when the performance on the validation set starts to degrade. This is done by monitoring the validation loss during training, and stopping the training process when the validation loss starts to increase too significantly. As shown, the validation loss comes across as noisy, so before employing early stopping, it is important to be certain that the validation loss is getting consistently worse. Alternatives to early stopping is simply by using a checkpoint system, where the weights of the model, or its current state, are saved at regular intervals. Regularization is also a common method to combat overfitting. This is done by adding a penalty term to the loss function, which encourages the model to learn smaller weights. The most common form of regularization is L2 regularization, which adds a term to the loss function that is proportional to the square of the weights. This encourages the model to learn smaller weights, which can help prevent overfitting. L1 regularization is also used, which adds a term to the loss function that is proportional to the absolute value of the weights.

Finally, data augmentation comes in many shapes and sizes, literally. Data augmentation is a technique used to artificially increase the size of the training dataset by applying various transformations to the input data. A selection of these is shown in Figure 9. These data augmentation techniques are used to create new training samples by applying various transformations to the original image. This is done to increase the diversity of the training data, which can help improve the performance of the model. The shown examples are very common in computer vision tasks. Finally, a technique called cross-validation is often used. This method splits the dataset into multiple subsets, or folds. During training, one of the folds is used for validation, while the others are used for training. This is done to ensure that the model is not overfitting to a specific subset of the data.

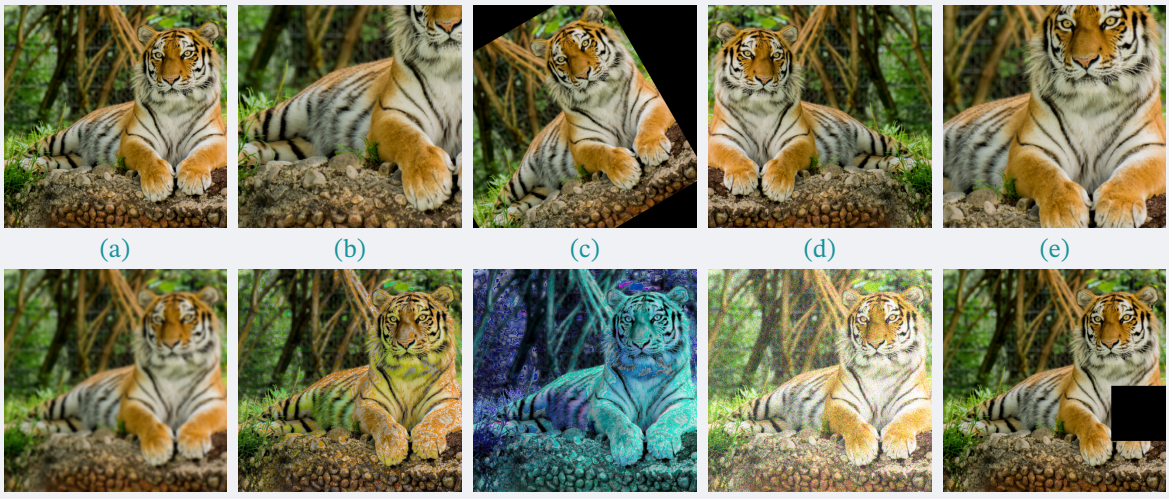


Figure 9. Data augmentation techniques. (a) Original image. Applied augmentation are (b) cropping, (c) rotation, (d) flipping, (e) zooming, (f) blurring, (g) saturation adjustment, (h) hue adjustment, (i) noise addition, and (j) occlusion. Image source: viso.ai [35].

The architectures, methods, and techniques present in modern deep learning and computer vision are vast and complex. However, these methods are only as effective as the data that fuels them. Selecting, generating, or collecting the right data for any given task is as crucial for the success of the model as the model itself.

2.2.2 Datasets ✓

Datasets are the lifeblood of any machine learning model. Entries flow through a model's structure, helping it learn to perform a specific task. In datasets, both the quantity and quality of the data are important. The more data a model has to learn from, the better it will perform. However, the quality and variety are also tantamount to the model's performance. If the data is not representative of the task, the model will not be able to learn to perform the task. If too little variety is present, it will not learn to generalize and only be able to perform the task on scenarios closely related to that which makes up its dataset.

There exists many different datasets for very different purposes. As presented earlier, the ImageNet dataset was key to the success of AlexNet. This dataset consists of millions of images with many thousands of classes. For the task of image classification, this dataset was very sufficient in providing the model with enough data to learn and generalize from. However, the base ImageNet dataset is insufficient for related CV tasks, such as object detection and segmentation. For these tasks, the dataset needs to be annotated with bounding boxes and class labels for each object in the image. This is a very time-consuming task to do from the ground up, which is why researchers often rely on pre-annotated datasets that align closely with the specific requirements of their task.

Creating datasets of any significant size can be a rather difficult task, where some tasks are simpler than others. For image classification, each dataset entry only requires a single image and a class label. This is a fairly simple task to do, as the most difficult part comes

from collecting the images, with labelling taking very little time per entry. For other tasks like object detection and segmentation, the task of labelling becomes much more difficult. This is because each entry requires a bounding box or segmentation mask for each object in the image. This is a very time-consuming thing to do, and the quality of the labels may be of worse quality than desired, meaning there also needs to be some form of quality control. This means that the creation will take even longer, as labels that are poorly made, might heavily affect the model's performance.

To mitigate errors in the labelling process, some move to simply make the dataset with synthetic data. This is done by using a simulator of some kind, ranging from a simple game engine to a full-fledged simulator. This is done by creating a virtual world, where the objects in the world are simulated. This allows for the creation of a dataset with perfect labels, as the simulator knows exactly where each object is in the world. Furthermore, making synthetic data can help increase the diversity in a dataset for areas where it might be difficult to obtain through real-world means. The most common problem with generating synthetic data, especially imagery of real-world scenarios, is the fact that the generated data does not look like real-world data. This can cause problems for models trained on the data, as they might not be able to generalize to real-world data, as textures and shadows are less likely to look realistic.

For AVs, the datasets are often more complex than just images. They often consist of multiple sensors, such as LiDAR, radar, and cameras. This balloons the size of the dataset, resulting in singular entries being multiple gigabytes in size. The Waymo Open Dataset [42] is a prime example of this. It consists of 1950 entries, each consisting of 20 seconds of video at 10Hz. Each contains the data from a mid-range LiDAR, 4 short-range LiDARs, and 5 cameras. The entries contain labelled data for 4 object classes: vehicles, pedestrians, cyclists, and signs. Figure 10e shows an example of some LiDAR readings overlaid on the corresponding camera image. Waymo is a very detailed dataset, with a lot of data to learn from. However, it is also a very large dataset, which makes it difficult to work with on consumer-level computers. Competitors do exist, such as the nuScenes [43] dataset and the Argoverse [44] dataset.

Closely related are the lane-level datasets. These datasets are concerned with detecting and classifying lanes in images. This is a very important task for AVs, as it allows them to understand the road topology and navigate accordingly. The TuSimple dataset example shown in Figure 10d illustrates this task well, providing annotated lane markings for training and evaluation. For systems like AIM that manage the vehicle going through an intersection, datasets like the INTERACTION [45] dataset are needed. This dataset consists of overhead views of intersections, with annotated vehicles moving through them. This dataset is very useful for training models to understand the interactions between vehicles at intersections.

Finally, some satellite image-based datasets also exists for different purposes. The SpaceNet [46] dataset is concerned with the labelling of buildings and roads etc. in satellite images. This dataset consists of high-resolution satellite images and different annotations for the images, such as individual houses, both through bounding boxes and segmentation masks. For a bit wider scale, the DeepGlobe [47] dataset offers semantic masks of roads, building, and land cover.

(a) SpaceNet [46] (b) INTERACTION [45]

(c) DeepGlobe [47] (d) TuSimple [48] (e) Waymo [42]

Figure 10. Examples from datasets for AVs.

Common for all datasets, is the fact that they require a dataloader in order to be used in the program training the model. A dataloader is a class that is responsible for loading the data from the dataset and preparing it for training. This means that the dataset needs to be structured in a consistent way, such that the dataloader can work as seamlessly with it as possible. A dataset typically consists of a number of entries, each with their own attributes, such as ground truth labels, images, LiDAR and radar data, and other sensor data. Maintaining a consistent structure across all entries is important, otherwise a dataloader might get needlessly complex with unnecessary fail-safes.

Dataloaders often come with the ability to split a dataset into training and validation sets. This is done by randomly splitting the dataset into two parts, where one part is used for training and the other part is used for validation. This is commonly done by splitting it, with a certain percentage going to either set. Alternatively, the dataset can be split beforehand, meaning it consists of predefined training and validation sets. This can be a good choice when the model could become prone to overfitting, which is likely to happen when the dataset entries are similar looking, such as with intersections. If techniques like cross-validation are used, the model is at some point trained on every entry, which for long training sessions will eventually lead to overfitting. This can, however, be mitigated with the use of data augmentation.

2.3 Satellite Imagery ✓

Datasets can consist of many different views. Many datasets for AVs consist of images taken from a car driving around, typically consisting of many different cameras and sensors. Others, like is the focus in this project, consist of satellite images. Satellite imagery provides

a bird's eye view of the world. This is a very useful perspective for many tasks, such as road extraction and lane detection. Many different sources of satellite imagery exist, such as Google Maps, Azure Maps, and Sentinel, each offering their own capabilities through Application Programming Interfaces (APIs). The most important among these is the ability to get high-resolution static images of a specific location, preferably as close in dates as possible. This not only increases the likelihood of images being of a usable resolution, but also reflecting of the current state of locations.

The resolution of satellite images refers to the size of the pixels in the image. The less area a pixel covers, the more information and detail about the location is captured. Google Maps Static API offers a sub-meter spatial resolution, meaning each pixel represents about 30-50cm on the ground [49]. At this level of detail, things like road markings and lane boundaries can be detected. In contrast, if the resolution is at a meter-level spatial resolution, then these fine details are at best blurred and at worst completely lost since they represent potentially a fraction of the land making up said pixel. At this level, only the general structure may be inferred from the resulting satellite image. Thus, having a decently high-resolution image is very important for giving NNs a chance at understanding intersections. For example, it needs to be able to see road-marking that make up the lanes, especially important when certain lanes are for going certain ways, often marked by arrows; you would not want a vehicle go straight through an intersection when it has placed itself in a lane that is only for turning right.

Satellite images provide strong advantages to AVs. Firstly, they are not reliant on live images, meaning they can be used even when it is cloudy. This means that they can provide a consistent and reliable source of information about the environment, regardless of weather conditions. Secondly, they can help AVs understand the road topology of any intersection before arriving at it. This means that a vehicle can place itself in the appropriate lane beforehand, delivering a smoother experience to the driver and passengers. This is especially important for AVs, as they need to be able to navigate complex intersections without human intervention. Finally, satellite images can be used to create high-definition maps of the environment. These maps can be used to help AVs navigate and understand their surroundings, potentially even help them find paths that offer smoother rides or other desired traits of travelling to some destination.

In summary, this chapter has thoroughly established the theoretical foundation necessary for the work presented in this thesis. By exploring the taxonomy of autonomous vehicles and the technologies enabling their development, it has provided context for the broader application landscape. The in-depth overview of deep learning — from its historical roots to modern architectures — lays the groundwork for understanding the computational methods used in this project. Furthermore, the exploration of datasets and the unique role of satellite imagery highlights the importance of data quality and perspective in training effective models. With this background in place, the thesis now transitions into more closely related works before moving onto the more specific methodologies and experiments that form the core of this work.

3 Related Work

This brief chapter will present the work in areas most closely related to the work presented in this thesis, namely path-planning and intersection management. Path-planning has been a long-standing area of research in the field of robotics and autonomous navigation, with many algorithms and techniques developed over the years. Intersection management is a more recent area of research that focuses on the challenges of managing multiple vehicles at intersections, particularly in urban environments.

3.1 Path-planning ✓

Path-planning is the task of having a certain amount of knowledge about the environment and finding a path from a starting point to a goal. This task is one of the most fundamental tasks in the field of robotics and autonomous navigation, and has thus a long history of improvement and evolution.

One of the first algorithms to be used for path-planning is the Dijkstra algorithm from 1959 [50], which is a graph search algorithm that finds the shortest path between nodes in a graph. The A* algorithm is another popular algorithm from 1968 that is used for path-planning, and is a combination of Dijkstra's algorithm and a heuristic function that estimates the cost of the cheapest path from a node to the goal node [51]. Some years later, the D* algorithm was introduced in 1994, which is an incremental search algorithm that finds the shortest path between nodes in a graph, and is an improvement over the A* algorithm [52]. D* has since become a very popular algorithm for path-planning in robotics, with improved alternatives like Focused D* the year after [53] and D* Lite from 2005 [54] proving use in real-world applications.

The concept of Artificial Potential Fields (APFs) was introduced in 1986 [55]. It assumes some repulsive field around obstacles to avoid and a pulling force towards the goal, resulting in autonomous robots navigating towards the goal while avoiding obstacles. This method is particularly effective in dynamic environments, where the obstacles are moving. It is, however, very prone to local minima and situations where it might get trapped [56], significantly reducing its effectiveness in complex environments, such as a tight hallway where the robot might fit, but the calculated repulsive force is too great or if it encounters a dead-end or U-shaped obstacle, leading it to loop infinitely. Combined with other global path-planning algorithms, it has shown considerable success, especially in the field of swarm robotics [57], [58].

The Rapidly-exploring Random Tree (RRT) algorithm was introduced in 1998 [59], and is a popular algorithm for path-planning in robotics. It is a randomized algorithm that builds a tree of possible paths from the starting point to the goal point, and is particularly useful in high-dimensional spaces. A node is randomly chosen from the initial point. The intermediate node is then determined based on the movement direction and maximum section length. If obstacles are detected, the route in that direction is ignored. Otherwise, a new random point is selected. The RRT* algorithm was introduced in 2011 [60], improving on the original with two small but significant modifications: a cost function that takes into account the distance between nodes and a re-wiring step that allows the tree to be restructured to find a better path. It has shown great usage in real-world applications regarding Autonomous Underwater Vehicles (AUVs) [61], [62], despite challenges regarding the need for information about large areas [63].

Other areas of research in path-planning include Genetic Algorithms (GAs) and Fuzzy Logic (FL). GA [64] is inspired by the process of natural selection, where only the fittest organisms survive. Generally, the algorithm works by generating a random population of solutions, and then selecting the most efficient ones by using some cost function. Then these selected solutions go through the crossover process where they are combined and mutated to generate new solutions. FL is another old method from 1965 used for path-planning [65], [66]. It depends on functions used in fuzzification, inference, and defuzzification. These functions are based on a descriptive classification of the input data, such as low, medium, or high collision risk. Based on the defuzzification process, the robot decides on the best path to take.

NNs are also finding their usage in the field. NNs are made to imitate the human brain's innate ability to learn. They are trained on data and learn how to react to it. They are used in the field, not necessarily for path-planning explicitly, but more in conjunction with other algorithms that use their output as input. I.e. a NN might be able to tell the controller where some obstacle is, meaning it is giving a helping hand to algorithms like APF. Akin to APF, RL models are taught to react to their surroundings, driving towards a goal and being rewarded and penalized for the actions that it takes, like how APF is moving towards a goal and avoiding obstacles due to the repulsive forces.

In summary, the evolution of path-planning – from early graph search methods like Dijkstra and A* to more adaptive techniques such as D*, RRT, and learning-based models – illustrates a steady push toward efficiency and robustness. Approaches like APFs, GAs, and FL add further flexibility, each with its own trade-offs. Together, these methods highlight the ongoing effort to balance computational efficiency with real-world challenges.

3.2 Intersection Management

4 Methodology

This section covers the methodology and work produced as part of this thesis. The first section to be detailed is the retrieval of satellite images. This includes the API usage and method used for signing URLs. Following this, the loss functions used and created for this project are detailed. It encompasses the three loss functions used in the project, including cross entropy loss, a custom coldmap based loss, and a topology based loss. Then the dataset created and used in this project is presented. It includes the creation of the dataset and the data augmentation techniques used. Finally, the models used to answer Research Question **RQ-1** are presented, followed by the training strategy used to train the models with the various loss functions. ✓

4.1 Satellite Imagery ✓

Satellite imagery is a key component of this thesis project. The imagery will be used for both training and testing the DL models, by creating a dataset detailed in Section 4.3, and as input to said models during inference. This section covers the acquisition of satellite imagery, the process of signing URLs as required by the API, and the code created for these purposes.

This project utilizes Google Maps Static API as provided by Google Cloud Platform. The API allows for the retrieval of static map images at a given resolution and zoom level. This API was chosen due to its ease of use, the quality of the retrieved images, and the fact that it is free to use for a limited number of requests. The API is used to retrieve satellite imagery of a given location.

4.1.1 Image Acquisition ✓

Google Maps Static API can retrieve images by forming requests with specific parameters that define the center, zoom level, size, and additional options for the map. For this project, images of type `satellite` are used, as they provide the highest level of detail for each retrieved image. Other types like `roadmap` or `terrain` do not provide enough detail to create a path that would realistically help navigate any kind of intersection as things like line markings are abstracted away.

To request an image, a URL is generated dynamically for the API, incorporating the required parameters. The parameters of the API request are as follows:

- `center`: The latitude and longitude of the center of the map (e.g. `41.30392`, `-81.90169`).
- `zoom`: The zoom level of the map. 1 is the lowest zoom level, showing the entire Earth, and 21 is the highest zoom level, showing individual buildings.
- `size`: The dimensions of the image to be retrieved, specified in pixels (e.g., `400x400`).
- `maptype`: Specifies the type of map to be retrieved. Options include `roadmap`, `satellite`, `terrain`, and `hybrid`.
- `key`: The API key used to authenticate the request.
- `signature`: Secret signing signature given by Google Cloud Platform through their interface.

Furthermore, the API allows for markers to be placed on the map, which can be used to highlight specific points of interest. This is, however, not relevant to this project.

4.1.1.1 URL Signing ✓

While requests to the API can be made using only the API key, the usage is severely limited without URL signing. URL signing is a security measure that ensures that requests to the API are made by the intended user. The signature is generated using the API key and a secret key provided by Google Cloud Platform. The URL signing algorithm is shown in Algorithm 1 and is provided by Google [67].

Algorithm 1: URL Signing Algorithm (`sign_url`)

Input: URL, secret

url ← urlparse(URL)
secret_decoded ← base64_decode(secret)
signature ← HMAC_SHA1(secret_decoded, url.path + '?' + url.query)
signature ← base64_encode(signature)
URL_signed ← URL + '&signature=' + signature

Output: URL_signed

As input is the URL with filled parameters and the secret key. The algorithm generates a signature using the HMAC-SHA1 algorithm with the secret key and the URL to be signed. The signature is then base64 encoded and appended to the URL as a query parameter. The signed URL can then be used to make requests to the API.

4.1.2 Implementation ✓

The main functionality of satellite imagery retrieval can be seen in Listing 1. An example of the output of the functionality can be seen in Figure 11.

```python
def get_sat_image(lat: float, lon: float, zoom: int = 18, secret: str = None):

    req_url = f"https://maps.googleapis.com/maps/api/staticmap?center={lat},{lon}&zoom={zoom}&size=400x400&maptype=satellite&key={API_key}"

    signed_url = sign_url(req_url, secret)

    response = requests.get(signed_url)
    return response


def save_sat_image(response: requests.Response, filename: str = "map.png") -> None:
    if response.status_code != 200:
        raise Exception(f"Failed to get image, got status code {response.status_code}")

    with open(filename, "wb") as f:
        f.write(response.content)
```

Listing 1. Python functions used to retrieve and save satellite imagery (`get_sat_image` and `save_sat_image`)

Listing 1 shows two functions, `get_sat_image` and `save_sat_image`, that are used to retrieve and save satellite imagery, respectively. The `get_sat_image` function constructs a URL for the Google Maps Static API request and signs it using the `sign_url` function detailed in Algorithm 1. The signed URL is then used to make a request to the API, and the response containing the image is returned. This response can then be passed to the `save_sat_image` function, which saves the image to a file with the specified filename.

A small `rotate_image` function was also created to rotate the retrieved image by some degrees, as the orientation of the satellite images can vary. The code can be seen in Listing 2. This is meant to help simplify the task performed by the model, as it alleviates the need to handle poorly angled images.

```python
def rotate_image(image_path, angle) -> None:
    image_obj = Image.open(image_path)
    rotated_image = image_obj.rotate(angle)
    rotated_image.save(image_path)
```

Listing 2. Python function to rotate an image by a specified angle (`rotate_image`)

https://maps.googleapis.com/maps/api/staticmap?center=55.780001,9.717275&zoom=18&size=400x400&maptype=satellite&key=wefhuwvjwekrlbvowilerbvkebvlearufhbebwe&signature=aqwhfunojlksdcnipwebfpwebfu=

Figure 11. Example of a signed URL and satellite image retrieved using the Google Maps Static API.

4.2 Loss Function Design

One of the most critical parts of designing a deep learning model, is the creation of the loss function that will guide the training. The loss function is a measure of how well a model is performing, and it is used to adjust the model's parameters during training. Therefore, the choice of loss function is crucial to the success of the model. In this section, I will discuss the design of the loss functions used to train the selected models. It will consist of a combination of different loss functions, each designed to capture different aspects of the problem at hand. Firstly, I will cover the utilization of a common classification loss function, the Cross-Entropy (CE) loss, which is used to measure the difference between the predicted and true distributions. CE will be the main driving force behind the models prediction of the correct labels for each path through the intersection.

Secondly, two different methods for enforcing topological constraints will be presented. First of these is the development of the novel "cold map"-based loss function, which is supposed to guide the model by penalizing points further away from the true path subject to some threshold. In combination with this, I will discuss the use of BCE loss. Finally, a topology-based loss function focusing on enforcing specific betti numbers will be presented.

4.2.1 Cross-Entropy Loss

4.2.2 Cold Map Loss ✓

The cold map loss is a novel loss function designed to enforce topological constraints on the predicted path. It is based on the idea of using a cold map, which is a grid of the same size as the input image, where the intensity of each cell is a value derived from the distance to the nearest path pixel. Before covering them and their creation, the BCE loss function is presented, as it is closely related the CE loss but for binary classification tasks, such as when looking at the structure of the predicted path.

4.2.2.1 Binary Cross-Entropy Loss ✓

The BCE loss is a commonly used loss function for binary segmentation tasks, which is relevant for the task at hand due to the pixel-subset nature of the problem, i.e. pixels can be either zero or non-zero. Furthermore, it is well-versed in handling heavily imbalanced data, which is the case when dealing with classification tasks where one class is much more prevalent than the other, like path and non-path pixels in an image. These properties make it ideal for this problem, as the background pixels are much more prevalent than the path pixels. The implementation is using the definition from PyTorch[2], which is defined as follows:

[2]Full implementation details can be found in the official documentation: https://pytorch.org/docs/stable/generated/torch.nn.BCELoss.html

$$\ell(x, y) = L = \{l_1, ..., l_N\}^T \tag{20}$$

with

$$l_n = -w_n[y_n \log x_n + (1 - y_n) \log(1 - x_n)] \tag{21}$$

where w_n is a weight parameter, y_n is the ground truth label, x_n is the predicted label, and ℓ is subject to

$$\ell(x, y) = \begin{cases} \text{mean}(L), & \text{if reduction} = \text{'mean'} \\ \text{sum}(L), & \text{if reduction} = \text{'sum'} \end{cases} \tag{22}$$

depending on the reduction parameter. Looking at the inner bracket (21), the left-hand side is activated when the ground truth label is 1. It evaluates how well the positive class's predicted probability x_n aligns with the ground truth. A smaller loss is achieved when the predicted probability is close to 1. The right-hand side of (21) is activated when the ground truth label is 0. It evaluates how well the negative class's predicted probability x_n aligns with the ground truth. A smaller loss is achieved when the predicted probability is close to 0. The BCE loss is then calculated as the sum or mean of the individual losses, depending on the `reduction` parameter. The weight parameter w_n can be used to scale the output of the loss function, which is key in making the cold map loss work.

Formally, BCE quantifies the dissimilarity between the predicted probabilities and the actual labels, giving a sense of how well the model is performing. For example, calculating the BCE with $y = 1$ and $x = 0.8$ gives

$$-1 \cdot (1 \cdot \log(0.8) + (1 - 1) \cdot \log(1 - 0.8)) = -\log(0.8) = 0.223$$

This value represents the dissimilarity between the predicted probability of 0.8 and the actual label of 1. A lower value indicates that the model's prediction is closer to the ground truth, suggesting a more accurate classification. Alternatively, the value can be near 1, indicating a large discrepancy between the predicted and actual labels. So, a value of 0.223 is a good result, as it indicates that the model is performing well, with some room for improvement. For contrast, if the predicted label is 0.4, but the true label is 1, the dissimilarity would be $-\log(0.4) \approx 0.916$. This higher value reflects a more significant error in prediction. From this, note that the function calculates the dissimilarity for both positive and negative classes. So, in the case of the predicted label being 0.4 and the true label being 0, the loss value would be much better at just $-\log(1 - 0.4) \approx 0.51$.

Other considerations for handling imbalanced data include methods like Dice [68] similarity coefficient and Focal loss [69], which can also be effective in certain contexts. The Dice similarity coefficient is a measure of overlap between two samples, and is particularly useful when dealing with imbalanced data. The Focal loss is designed to address the class imbalance problem by focusing on hard examples that are misclassified. These methods can be used in conjunction with the BCE loss to improve the model's performance, especially when dealing with heavily imbalanced datasets. But for the purposes of this project, the BCE loss is expected to be sufficient.

4.2.2.2 Cold Maps ✓

The cold map loss is the first of two loss functions that will improve topological soundness within the output from the models. This involves using the predicted path generated by the model, and comparing it to the cold map from the dataset. The creation of the cold maps are detailed in Section 4.3.1. Briefly, the cold maps are grids of the same size as the input image, where the intensity of each cell is a value derived from the distance to the nearest path pixel magnified beyond some threshold.

The main idea behind the cold map loss is to introduce spatial penalty that increases as the distance from the true path increases. Though it is similar to BCE, it differs in some key aspects. It is not pixel-wise, but rather a global loss that is calculated over the entire image. This means that slight deviations from the true path are penalized less than those with a larger discrepancy; BCE simply checks for classes. This property of the loss function is a desirable trait for path-planning tasks, as minor offsets from the true path are less critical than larger ones. This loss function is defined as follows:

$$\mathcal{L} = \sum_{i=1}^{H} \sum_{j=1}^{W} C_{ij} P_{ij} \tag{23}$$

where C_{ij} is the cold map value at pixel (i, j), and P_{ij} is the predicted path value at pixel (i, j). This version of the loss function is a simple dot product between the cold map and the predicted path. Thus, after flattening the cold map and the predicted path matrices, the loss is calculated as the dot product between the two vectors:

$$\mathcal{L}_{\text{cold}} = C \cdot P \tag{24}$$

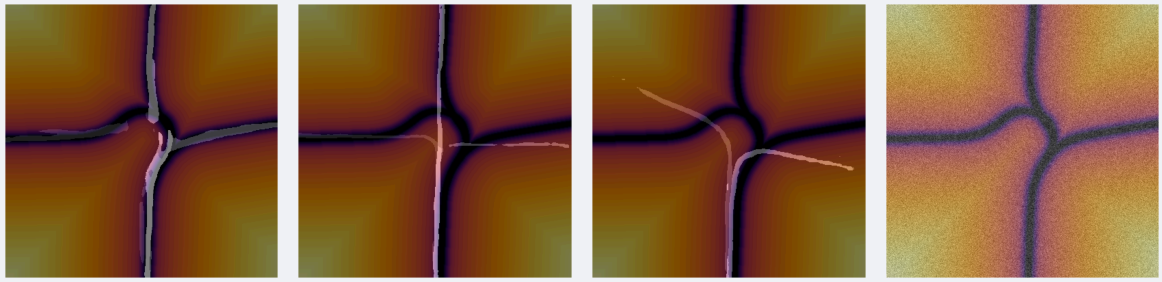
where C is the cold map vector and P is the predicted path vector, giving a scalar value, contributing to the total loss. However, in practice, this on its own does not drive an optimizer to optimize the models in any meaningful way. What an optimizer like Adam or SGD does, is to minimize the loss function by adjusting the model's parameters. This cold map loss does well to drive it towards removing activated pixels, but is does this in a very destructive way. By simply pushing every logit from the model towards $-\infty$ it achieves a perfect score: 0. This means that the model will not output any meaningful predictions, as it will simply push all logits to $-\infty$. This is because of the sigmoid function used inside the loss's implementation. Simply doing the aforementioned dot product requires the output of the model to be in the range of $[0, 1]$, which is not the case for the logits, so it is put through a sigmoid function. Thus, as the optimizers tries to lower the loss, it inevitably ends up with $\sigma(-\infty) = 0$, which is the perfect score.

To combat this and to make the cold map loss viable, two major changes are made. The equations shown hitherto do work in the sense that they penalize point further away, but they don't reward correct predictions either. This means that the loss penalizes false positives, but not false negatives. Thus, the first of the major changes is introduced. The opposite of the cold map is introduced, i.e. a heat map. This is simply the reversed cold map: $h_{\text{map}} = 1 - c_{\text{map}}$. This operation essentially invert the cold map, meaning that the penalty is higher the closer the ground truth path it is. This heat map penalizes false negatives, and in combination the two will drive models to output paths that are close to the true path.

The second major change is the fact that BCE is used with the cold and heat maps as the weights. Since the output from the models are of the same size as the cold map, these maps can be used to individually weight the loss for each pixel. However, simply adding these two together will give a combined map that is simply 1 all over, so a α and β terms are introduced to add more credence to one map than the other. The combined map is thus found by the following equation:

$$w = \alpha h_{\text{map}} + \beta c_{\text{map}} \tag{25}$$

where $\alpha = 1$ and $\beta = 0.5$ because it should reward true positives more than anything else. This, along with the output logits from the model, is passed to the BCE loss function, which will then calculate the loss as a weighted sum of the individual losses. Examples of this loss function in action is shown in Figure 12.



(a) Loss: 1.13 (b) Loss: 2.57 (c) Loss: 2.61 (d) Loss: 0.88

Figure 12. Paths drawn on top of a cold map with their associated loss.

Despite the fact that the loss value for Figure 12a is seemingly high, it is actually a good path. This might seem like an error in the loss function, but the loss value itself is not of the highest importance. The most important part is the fact that the optimizer is able to drive the model to a point where it can output a path that is close to the true path, i.e. lower the loss value as much as possible. As this cold map based loss function is a novel approach, it should be compared to related loss functions that aim to achieve the same goal. Therefore, the following section will present work already done in this area, concerning itself with a more algebraic approach to the topology problem.

4.2.3 Continuity Loss ✓

The second of the topology based loss functions is the continuity loss function. The continuity loss function will revolve around getting the output from the model to be a singular, continuous component, meaning it does not contain breaks or holes. Formally, this is done by aiming for specific Betti number values, ensuring that the predicted path has a single connected component and no loops. The details of this approach are discussed in the following sections.

Considerations of using existing topology existing methods. Dep *et al.* [70] introduced TopoNets and TopoLoss. This loss function revolves around penalizing jagged paths and encouraging smooth, brain-like topographic organization within neural networks by

reshaping weight matrices into two-dimensional cortical sheets and maximizing the cosine similarity between these sheets and their blurred versions. Cortical sheets are two-dimensional grids formed by reshaping neural network weight matrices to emulate the brain's spatial organization of neurons, enabling topographic processing. While initially interesting in the context of this project, simple testing showed that the values returned from this loss, did not give a proper presentation of the path's topology, outside of its smoothness. And while smoothness is a part of the topology, this will largely be handled by the CE loss.

This section will present the topology-based loss function designed for this project, with a focus on ensuring that the predicted path is continuous and does not contain any breaks or holes. This is a crucial aspect of the task at hand, as the goal is to create a path that a vehicle can follow. Breaks in a path would be unrealistic for a grounded vehicle to follow. As a starting point, it is important to understand the concept of Betti numbers:

Betti Numbers

Betti numbers [71] come from algebraic topology, and are used to distinguish topological spaces based on the connectivity of n-dimensional simplicial complexes. The nth Betti number, β_n, counts the number of n-dimensional holes in a topological space. The Betti numbers for the first three dimensions are:

- β_0: The number of connected components.
- β_1: The number of loops.
- β_2: The number of voids.

The logic follows that, in 1D, counting loops are not possible, as it is simply a line. This, if the number is greater than 1, means it is split into more than one component. In 2D, the number of loops is counted, i.e. a circled number of pixels. In 3D, this extends to voids.

With this, for the 2D images used in this project, the Betti numbers are β_0 and β_1. The continuity loss is designed to ensure that the predicted path has a single connected component and no loops. This is achieved by aiming for the Betti numbers to be $\beta_0 = 1$ and $\beta_1 = 0$. Higher dimensional Betti numbers are not relevant for this project, as the images are 2D. While Betti numbers are a powerful tool for topology analysis, they are not directly applicable to the loss function as they are discrete values. This means that they offer no gradient information, which is essential for training a neural network. Instead, persistent homology is deployed.

Persistent Homology (PH) is a mathematical tool used to study topological features of data. Homology itself is a branch of algebraic topology concerned with procedures to compute the topological features of objects. Persistent homology extends the basic idea of homology by considering not just a single snapshot of a topological space but a whole family of spaces built at different scales. Instead of calculating Betti numbers for one fixed space, a filtration is performed. This filtration is a sequence of spaces, where each space is a subset of the next, i.e. a nested sequence of spaces where each one is built by gradually growing the features by some threshold. As this threshold varies, topological features such as connected components and loops will appear (be born) and eventually merge or vanish (die).

This birth and death of features is recorded in what is known as a persistence diagram or barcode (See Figure 13). In these diagrams, each feature is represented by a bar (or a point in the diagram) whose length indicates how persistent, or significant, the feature is across different scales. Features with longer lifespans are generally considered to be more robust and representative of the underlying structure of the data, whereas those that quickly appear and disappear might be attributed to noise.

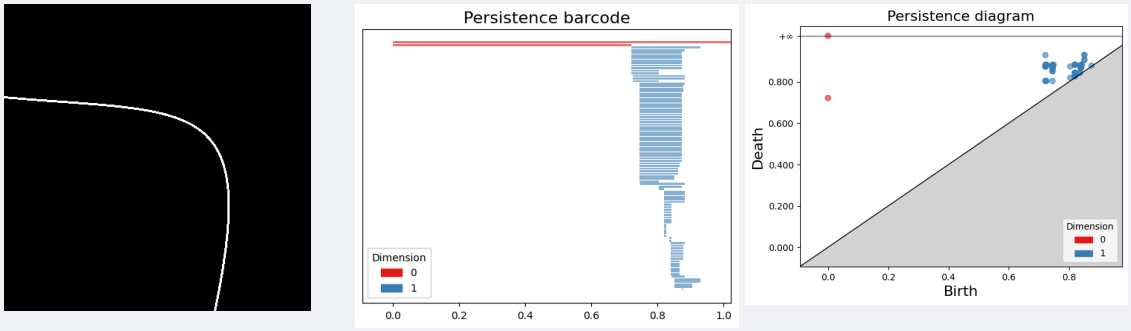


Figure 13. The top row shows a connected path along with its persistence barcode and persistence diagram, while the bottom row shows a disconnected path. The number of lines in the barcode, stems from the fact that the images are rather large in size and thus the number of built spaces are many.

The following will cover the method used to achieve the continuity loss function. It closely follows the work done by Clough *et al.* [72] with some minor changes that will be pointed out. Furthermore, the implementation is done using the Gudhi library, which is a Python library for computational topology. It provides a set of tools for computing persistent homology and other topological features of data. The library is designed to be efficient and easy to use, making it a good choice for this project. Lastly, PyTorch's ability to create custom autograd functions is used to implement the persistent homology loss function.

The output from the network is a tensor of size

$$\Omega = H \times W \tag{26}$$

where, in this project, $H = 400$ and $W = 400$. Then for each pixel $x \in \Omega$, the network outputs a logit vector:

$$L(x) \in \mathbb{R}^C \tag{27}$$

where C is the number of classes. A softmax function is applied to give class probabilities:

$$P_c(x) = \frac{e^{L_{c(x)}}}{\sum_{j=1}^{C} e^{L_{j(x)}}} \tag{28}$$

Here it is noted that the background class is $\mathrm{bg} = 0$. Following that, a foreground probability map is created by removing the background probabilities:

$$\mathrm{fg}(x) = 1 - P_{\mathrm{bg}}(x) \in [0, 1] \tag{29}$$

Now, persistent homology on images is usually computed through super-level sets. Clough *et al.* formulates PH on super-level sets, which is the opposite of sub-level sets. Sub-level

sets were chosen, as the alternative wouldn't work in practice, as Gudhi is kept in sub-level mode. Therefore, an inverted version of the foreground probability map is created:

$$f(x) = 1 - \mathrm{fg}(x) \in [0, 1] \tag{30}$$

where low values mark confident foreground pixels, and high values mark confident background pixels. The filtration, as generated inside Gudhi, is defined as

$$K_0 \subseteq K_{t_1} \subseteq \cdots \subseteq K_1 \tag{31}$$

where each sub-level set is

$$K_t = \{x \in \Omega \mid f(x) \leq t\} \tag{32}$$

This happens inside the `CubicalComplex` function, which is a Gudhi function that creates a cubical complex from the input. In short, the cubical complex is a data structure that represents the topological features of the input data. The Gudhi library is used to compute the persistent homology of the cubical complex, which is then used to calculate the Betti numbers.

From the cubical complex, persistence pairs are found, which are the birth and death values of the topological features. Gudhi returns a multiset for each homology dimension:

$$D_k(f) = \left\{ \left(b_i^{(k)}, d_i^{(k)} \right) \right\} \tag{33}$$

with the constraint $0 \leq b_i^{(k)} < d_i^{(k)} \leq \infty$, where $k = 0$ for connected components and $k = 1$ for loops. The component whose death time is $+\infty$ is discarded, as it is not influenced by the other pixels' values and all other pairs have finite deaths.

At this stage, some extra work is required, as Gudhi also returns *where* the feature pairs are born and die. In other words, flat Fortran-order indices of the pixels whose grey-values equal the birth and death values. These indices are remapped to PyTorch's row-major order with a small helper function:

$$\_\mathrm{F2C}(n) = r + cH \tag{34}$$

with $(r, c) = \mathrm{divmod}(n, H)$ where divmod returns the element-wise quotient and remainder of the two inputs.

Finally, the actual loss is found by defining the persistence, ignoring bars that are shorter than some threshold ε:

$$\mathrm{pers}_i^{(k)} = d_i^{(k)} - b_i^{(k)} > \varepsilon \tag{35}$$

where the threshold ε is typically a low value. Thus, the per-image loss, i.e. each image in a batch, is defined as:

$$\mathcal{L}(f) = w_0 \sum_{(b,d) \in D_0(f), d-b > \varepsilon} (d - b)^p + w_1 \sum_{(b,d) \in D_1(f), d-b > \varepsilon} (d - b)^p \tag{36}$$

where w_0 and w_1 are weights for the two dimensions, and p is a power parameter. The loss is then averaged over the batch size:

$$\text{Loss} = \frac{1}{B}\sum_{b=1}^{B}\mathcal{L}(f_b) \tag{37}$$

Because each summand in (36) is built from pixel values $f(x)$, gradients propagate exactly to those birth- and death-pixels:

$$\frac{\partial}{\partial f(x)}(d-b)^p = \begin{cases} -p(d-b)^{p-1}, & \text{if } x \text{ is the birth pixel,} \\ p(d-b)^{p-1}, & \text{if } x \text{ is the death pixel,} \\ 0, & \text{otherwise} \end{cases} \tag{38}$$

This comes from the fact that the partial derivatives are immediate, as the summand for any bar depends only on the two scalar values from f.

Examples of the PH-based loss function is shown in Figure 14. The far left image shows the output from a model, showing clear and separated components. This is the highly undesired trait this loss is designed to penalize. The loss value is 1.35, which is a very high value. The next image shows the same model after a backwards pass has been made by the loss function. Already, it is clear to see that the model has improved, as the components are now largely connected, with only some stragglers. Then after another two iterations, the model is nearing a loss of 0. The last image has a near 0 loss after just another iteration.

This highlights the effectiveness of this loss function. If this keeps going, the model will eventually reach a loss of 0. However, this rudimentary testing of the loss function also highlights the fact that it alone does not care for class labels. After just 20 iterations, it would label every pixel as the same class, as it is only concerned with the topology of the path. This is a trait of the loss function that is not desired, but when it works in conjunction with the other loss functions, it gets balanced to create more accurate results. This will be covered in depth in Section 4.4.5 where the training strategy is presented.

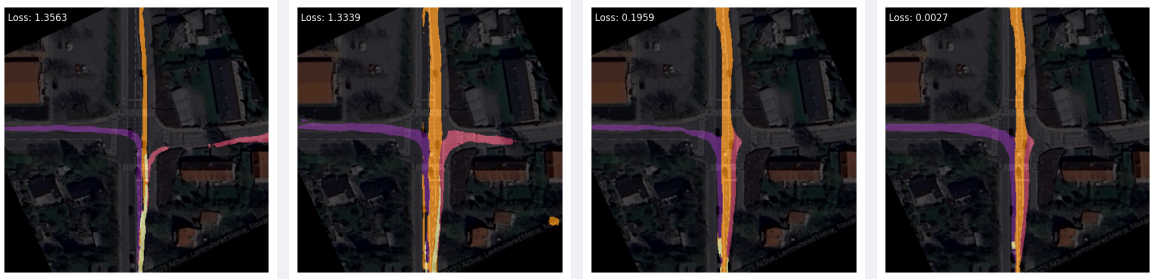


Figure 14. Results of refining trained model a few iterations on the same image.

Finally, before any loss function can be used, there needs to exist a dataset to train on. The dataset needs to contain content that it is desired for the model to learn. Furthermore, it is important for each entry to have their annotated ground truth, otherwise the loss function would not have anything to compare against. This is especially the case with models trained on supervised learning. Thus, the next section will cover the dataset created for this project.

4.3 Dataset Creation ✓

The creation of a proper dataset is crucial for making sure the models learn the task desired for them to perform. The dataset will have to work hand-in-hand with the model architecture and the loss function to ensure that the model learns the task effectively. Many aspects are to be considered when creating a dataset for a task as specific as this project sets out to create:

- It should be large enough to capture the complexity of the task. Size can be artificially increased through data augmentation.
- It should be diverse enough to capture the variety of scenarios that can occur at an intersection.
- It should allow for some leniency when it comes to generating paths, as the model should not be too stringent to a singular path.
- For the purposes of this project, its creation should seek to answer Research Question **RQ-3** by providing a dataset that allows for the training of a model that can generate paths that are not too stringent to a singular path.

By "not too stringent to a singular path" is meant that the model should be able to generate paths that are not too far from the desired path, while also allowing the model to generalize well.

4.3.1 Cold maps ✓

The deduced method for training the model, as detailed in Section 4.2, includes the use of a cold map. A cold map representation of the desired path was chosen for a small simplification in the loss function. It penalizes points that are further from the desired path, and does not do this for points that are on the path. Creating this cold map was done in several steps. First, a grid of the same size as the input image is created. The input image is the path drawn in white on a black background, as shown in centre Figure 17. This means that the only occupied pixels are those taken up by the possible paths. In this grid, the coordinates of the closest non-zero pixel is found by iterating over the entire input image containing the path. The complexity of this operation will be covered in the following sections. Next, the distance between the current pixel and the closest non-zero pixel is calculated. This distance is then compared to a threshold value to determine its value. If it is further away, the resulting penalty from the loss function should be higher. Different values for the threshold and the exponent of the distance calculation were tested to find the best combination. Lastly, the cold map is saved in a structured folder format for later use in training. Later, the created data is put through augmentation to inflate the size of the dataset and increase its diversity.

4.3.1.1 Finding the distance to the desired path ✓

The algorithm for finding the distance to the closest point on the desired path is shown in Listing 3.

```python
occupied = []
for i in range(binary.shape[0]):
  for j in range(binary.shape[1]):
    if binary[i, j] != 0:
      occupied.append((i, j))

h, w = binary.shape
nearest_coords = np.zeros((h, w, 2), dtype=int)

for i in range(binary.shape[0]):
  for j in range(binary.shape[1]):
    if binary[i, j] == 0:
      min_dist = float('inf')
      nearest_coord = (i, j)
      for x, y in occupied:
        d = hypot(i - x, j - y)
        if d < min_dist:
          min_dist = d
          nearest_coord = (x, y)
      nearest_coords[i, j] = nearest_coord
    else:
      nearest_coords[i, j] = (i, j)
```

Listing 3. Non-parallelized code for finding the nearest point on the path.

The algorithm in Listing 3 starts by creating an array of coordinates based on the `binary` map created with the `threshold` function from the OpenCV library. This `binary` map contains every non-black pixel in the input image, which in this case is the paths drawn on a black background. With these occupied pixels stored in an array, the algorithm then iterates over every grid point of the `nearest_coords` grid, created to be the same size as the input image. For every point in the grid, the algorithm checks if the point is on the path. If it is, the algorithm assigns the current point's coordinates to the `nearest_coords` grid. If the point is not on the path, the algorithm iterates over every occupied pixel and calculates the distance between the current point and the occupied pixel. If the distance is less than the current minimum distance, the minimum distance is updated and the coordinates of the closest point are saved. This is repeated for every occupied pixel, and the coordinates of the closest point are saved in the `nearest_coords` grid. This process is repeated for every point in the grid until every point has been assigned the coordinates of the closest point on the path. This grid will later be used under the name `coords`.

The shown algorithm is not parallelized and has a complexity of $\mathcal{O}(n^2)$, where n is the size of the input image. This is due to the nested `for`-loops used in the algorithm. While not a great complexity, it is a vast improvement over its earlier iteration which was $\mathcal{O}(n^4)$ [3]. The actual implementation of this algorithm is parallelized, but the non-parallelized form is shown here. The first iteration of the algorithm took 73 minutes to complete on a 400×400 image, while the parallelized version took 8 minutes on an 8-core CPU. This non-parallelized version takes roughly 30 seconds to complete on the same image, with the parallelized version taking just a few seconds on a full 400×400 image with very narrow paths. Further improvements are likely possible to be made both to the complexity of the implementation and parallelization could be distributed to a GPU or the cloud for even faster computation, but this remains future work.

[3] The original implementation can be seen in `dataset/lib.py:process_rows` in the GitLab repository.

4.3.1.2 Creating the cold map ✓

To start the creation of the cold map, a distance grid is created using Pythagoras' theorem between the coordinates of the point of the grid and the coordinates saved within, retrieved from the aforementioned `coords` variable. A masking grid is then created by comparing the distance grid to a threshold value. This results in each grid point being calculated using:

$$d_{ij} = \sqrt{\left(i - c_{ij0}\right)^2 + \left(j - c_{ij1}\right)^2} \tag{39}$$

$$dt_{ij} = \begin{cases} d_{ij} & \text{if } d_{ij} < t \\ t + \left(d_{ij} - t\right)^e & \text{otherwise} \end{cases} \tag{40}$$

where $c =$ `coords`, $c_{ij0} =$ `coords[i, j][0]`, t is the threshold value, and e is the exponent value. All three of these can be seen as function parameters in the function declaration in Listing 4. The distance grid is then normalized to a range of 0 to 255 to minimize space usage such that it fits within a byte, i.e. an unsigned 8-bit integer. This is also how PNG files store RGB values for each pixel. This is done by subtracting the minimum value and dividing by the range of the values. Alternatively, the `normalize` parameter can be set to another value, as usage within a loss function would prefer a value between 0 and 1 (as detailed in Section 4.2). The resulting grid is then saved as a cold map. The resulting cold map can be seen in the rightmost image in Figure 17.

```python
def coords_to_coldmap(coords, threshold: float, exponent: float, normalize: int = 1):
  rows, cols = coords.shape[0], coords.shape[1]

  distances = np.zeros((rows, cols), dtype=np.float32)
  for i in range(rows):
    for j in range(cols):
      distances[i, j] = hypot(i - coords[i, j][0], j - coords[i, j][1])

  distances_c = distances.copy()
  mask = distances > threshold
  distances_c[mask] = threshold + (distances[mask] - threshold) ** exponent

  distances_c_normalized = normalize * (distances_c - distances_c.min()) / (distances_c.max() - distances_c.min())

  return_type = np.uint8 if normalize == 255 else np.float32

  return distances_c_normalized.astype(return_type)
```

Listing 4. Non-parallelized code for finding the nearest point on the path.

To figure out the optimal values for the threshold and exponent, a grid search was performed. The grid search was done by iterating over a range of values for both the threshold and the exponent. The resulting cold maps were then evaluated by a human to determine which combination of values resulted in the most visually appealing cold map. For a 400×400 image, the optimal values were found to be $t = 10$ and $e = 0.5$. This combination meant that the resulting penalty for pixels further away, grew very quickly after the cold map is normalized. The grid of results can be seen in figure Figure 15. In testing, the value of $e = 1$ was excluded as it had no effect on the gradient produced in the cold map, meaning all values of t produced the same map.

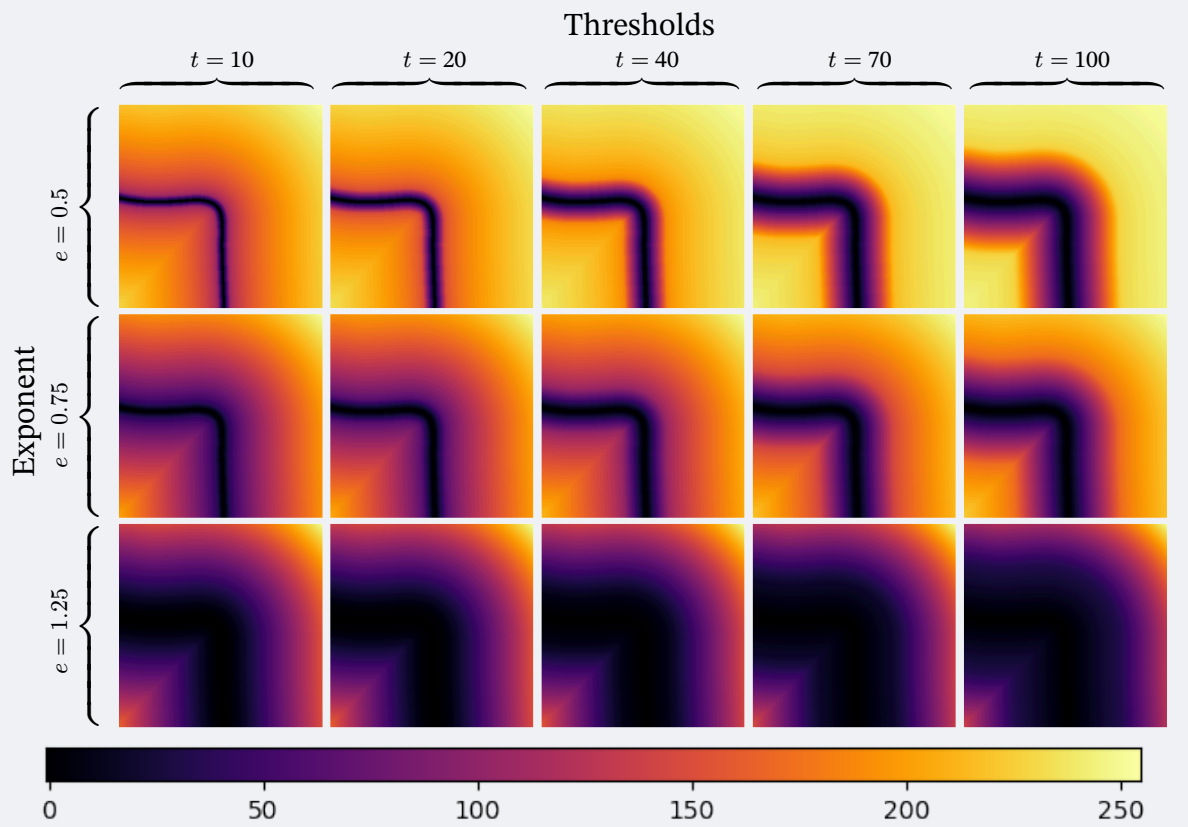


Figure 15. Results of testing threshold values $t \in \{10, 20, 40, 70, 100\}$ and exponent values $e \in \{0.5, 0.75, 1.25\}$. The colour map is shown beneath the results[4].

While pretty, these cold maps can be difficult to understand. Therefore, Figure 16 shows the 3D plots of the generated cold maps with exponent values $e \in \{0.50, 0.75, 1.25\}$. The 2D plots, particularly for an exponent like $e = 0.5$, might initially suggest an extremely sharp penalization of distant points. However, the 3D plots provide crucial clarification: while points far from the true path are indeed effectively penalized with a steep gradient that correctly orients towards the path, the key benefit of $e = 0.5$ emerges for points closer to the true trajectory. For these nearby points, the gradient becomes notably gentle, providing the desired leniency. This prevents minor, acceptable deviations from being harshly penalized and avoids overly aggressive corrections when a point is already close to the target. This behaviour — a strong corrective slope for distant points coupled with a forgiving gradient for proximate points — is precisely the desired characteristic for the loss function. Therefore, after careful analysis of the 3D plots, the exponent value $e = 0.5$ was selected, along with a threshold of $t = 10$, as the default for the function in Listing 4.

[4]The colour map is retrieved from the matplotlib docs: https://matplotlib.org/stable/users/explain/colors/colormaps.html

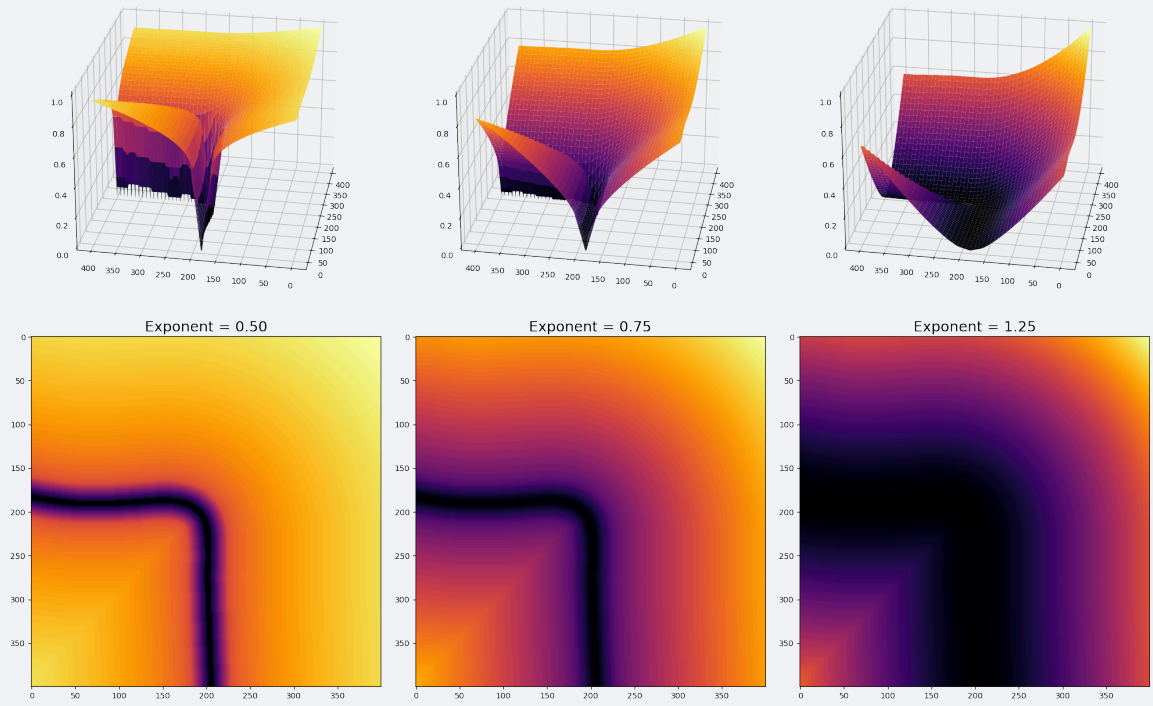


Figure 16. 3D plots of cold maps along with their 2D counterparts.

Finally, a comparison between the retrieved satellite image of an intersection, the optimal path through it, and the cold map generated by the process described above are shown in Figure 17. This highlights the importance of the cold map in the training process as opposed to the single line path. The cold map allows for a more lenient path to be generated, as the model is not penalized for deviating slightly from the path.

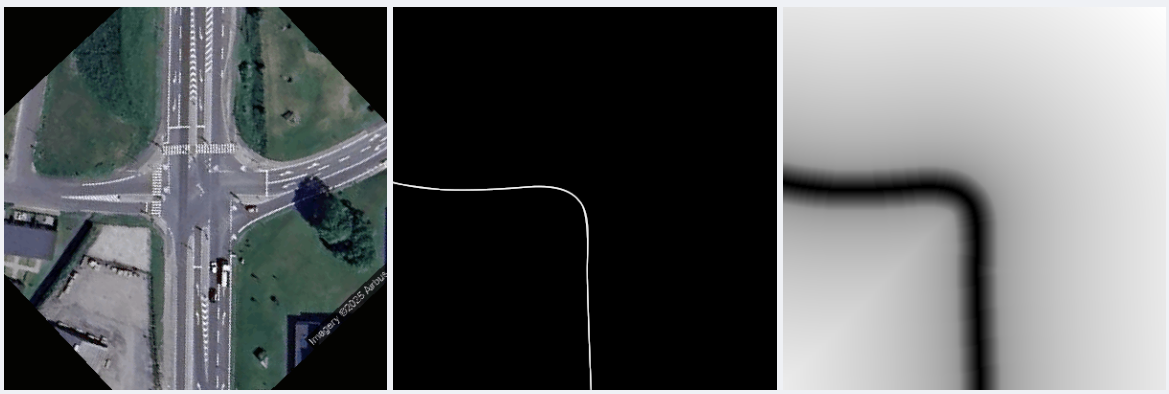


Figure 17. Example of satellite image, next to the desired path through with. To the far right, the generated cold map is shown with threshold $t = 10$ and exponent $e = 0.5$. Notice how it is only the points very close to the path that are very cold, while the rest of the map is warmer the further away it is.

At this point, the dataset is still very small, consisting of only 112 intersections. This is hardly enough to train models of any complexity, as overfitting is very likely to occur. Therefore, the dataset underwent severe inflation in the form of data augmentation. The dataset was augmented in several ways, including colouration, distortion, cropping, and zooming, covered in detail in the following section.

4.3.2 Data Augmentation ✓

Creating large datasets is a very time consuming tasks, scaling directly with the complexity and workflows structured around its creation. For the dataset created during this project, the workflow was as follows: Find a suitable intersection for the dataset. Copy the coordinates for the center of the intersection. Use the found coordinates in the satellite script described in Section 4.1.2 to download satellite images. Through trial and error, rotate the downloaded satellite image to align entry with bottom of the image.

Once a bunch of satellite images were downloaded, a small python script was used to automatically distribute each intersection image to their own folder and within each folder, create the structure shown in Listing 5. GNU Image Manipulation Program (GIMP) was chosen as the software to draw the paths through the intersections. First, another small script distributed a `.xcf` file to each intersection folder. `.xcf` is the file format for GIMP projects. This base `.xcf` was defined to be 400x400 pixels and contained a black background and three empty layers named “left”, “right”, and “ahead”. This massively simplified the process of creating the paths by not having to create a new project every time a new intersection was to be processed.

For each of the paths drawn in GIMP, they were saved individually as a `.png` file. Yet another small script then used these images of the path to create the corresponding JavaScript Object Notation (JSON) files containing the entry and exit coordinates of the path as well as generate the cold map. The cold map generated is stored as a `.npy` files, as the values of the cold map are simply between 0 and 1, meaning it does not make sense to store as a Portable Network Graphics (PNG) as there is not high enough values to be discernible to the human eye. The small scripts mentioned can be seen in Appendix X. The JSON file containing the entry and exit coordinates were at one stage to be used in a loss function that enforced entry and exit points, but has been left as future works, hence they remain in the dataset.

As described, this is a very time consuming process, and despite many hours of work being put into it, the training dataset only consisted of 112 intersections, some of which have very similar satellite images. To enlarge this dataset dramatically, the dataset underwent augmentation. Augmentation can be done in many ways with different methods. A variety of augmentations were chosen for this dataset, including: colouration, distortion, cropping, and zooming. The reason for choosing these will be discussed in their respective sections.

COLOURATION is the augmentation regarding the colours making up the image. In this project, this is achieved by adjusting the saturation and hue of the HSV colour space for an image. HSV stands for Hue, Saturation, and Value. Changing the hue of an image is changing the colour tone, meaning that a red image can be turned into a blue image. Changing the saturation is changing the intensity of the colours in an image, resulting in a more vibrant or dull image. Changing the value is changing the brightness of the image, meaning that a dark image can be turned much brighter and vice versa. HSV is generally more intuitive than the RGB colour space, as it is more closely related to how humans perceive colour.

Concretely, the colouration augmentation was done by randomly changing the hue and saturation of the image. This was done to help the models focus on structural features rather than specific colour cues, i.e. become better at generalizing. Colour augmentations also help the model become more robust to changes in lighting conditions. This is especially prominent when using satellite images from all kinds of areas. Some satellite images appear to have a very low image saturation, while others are more vibrant and sharp. Therefore, teaching the model to understand these different conditions is crucial. So, these colouration augmentations help make the models more adept at generalizing to different conditions.

The hue augmentation function randomly changes the hue of the image by a value between a lower and upper bound. These values are the defaults from the official documentation of the function. The saturation augmentation function randomly changes the saturation of the image by a value between 6, 8, and 10. These values were chosen as they were found to be the most visually appealing and interesting when testing the augmentations. Finally, a greyscale augmentation is also implemented, which is simply a call of the saturation augmentation function with the value 0. This is done to further enhance the models understanding of the structural features of the image.

`Hue augmentation`

The hue augmentation was done by randomly changing the hue of the image. The hue was changed by a value between -0.5 and 0.5. The image was then converted to a tensor and the hue was adjusted using the `adjust_hue` function from the `torchvision.transforms.functional` module. The resulting image was then converted back to a PNG image for easier handling.

`Saturation augmentation`

The saturation augmentation was done by randomly changing the saturation of the image. The saturation was changed by a value between 6, 8, and 10. The image was then converted to a tensor and the saturation was adjusted using the `adjust_saturation` function from the `torchvision.transforms.functional` module. The resulting image was then converted back to a PNG image for easier handling.

Examples of the colouration augmentations can be seen in Figure 18. Each column shows an intersection and its augmented variations. The top row is the original image, the second row is the greyscale augmented image, the third row is the hue adjusted image, and the fourth row is the saturation adjusted image. The greyscale images highlights the structural features of the image. By adjusting the hue, the dominant tones of the image are altered, resulting in dominant parts like vegetation appears as a variety of colours, such as blue or even purple. Adjusting saturation then makes the colours more vibrant or muted, creating anything from intensely vivid scenes to nearly colourless landscapes as also highlighted by the greyscale image.

Seeing these colour augmentation examples, it is clear to see that a large amount of diversity has been introduced to the dataset. Rather than training on a dataset where the colours might be very similar, hence not capturing the real world, the models are trained on a dataset that encapsulates real world colour variations. Furthermore, this motivates the

models to focus on structural features rather than specific colour cues, which is crucial for generalization.

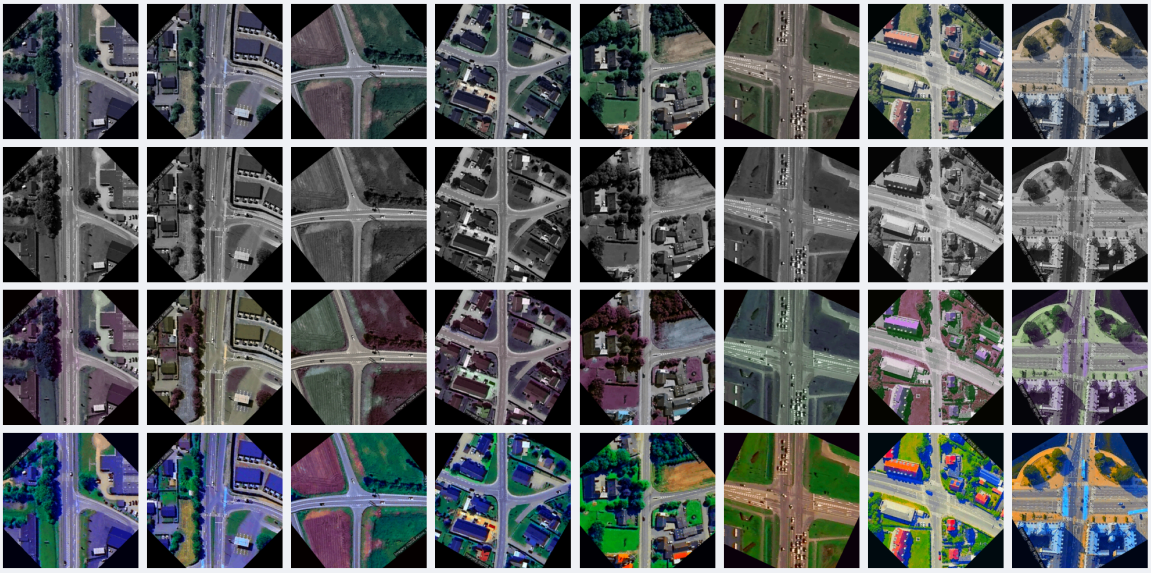


Figure 18. Example of the colouration augmentations. The top row is the original image, second row is the greyscale augmented image, third row is the hue adjusted image, and fourth row is the saturation adjusted image.

DISTORTION is the augmentation regarding the quality of the image. In this project, this is achieved by applying two methods of intentionally deteriorating the image quality. The first method is noise augmentation and the second is blurring the image. These augmentations were chosen, as they represent common downfalls when working with satellite images. Depending on the area where images are taken, the images are often more blurry in smaller towns and rural areas, while they are noticeably sharper and more pristine in larger cities like capitol cities. Thus, by incorporating distortion augmentations into the dataset, it becomes much more diverse and a better representation of the quality of images that the models are expected to work with.

The noise augmentation was chosen as it is a common issue with images in general, not just satellite images. Noise is produced in images through various outside factors, such as atmospheric conditions, sensor limitations, and environmental interferences. By adding this noise augmentation to the images, it helps the models learn to ignore these artifacts and focus on the relevant features of the image. The noise augmentation was done by generating Gaussian noise through the `randn` function from the PyTorch library. Furthermore, this kind of augmentation has been noted to act as a form of regularization on it own [73], as it prevents the model from overfitting to these clean, ideal, and pristine images that are common for very populated areas.

The blur augmentation was chosen as blur is a common issue with satellite images, particularly when using images from smaller cities, rural areas, older images, or far out of the cities on the roads where population density is significantly lower. Applying a Gaussian blur simulates these conditions by intentionally deteriorating the image quality, making it more representative of real-world scenarios. This helps the model learn to extract structural

features from the image, despite the images being obscured. This augmentation is particularly useful because it forces the model to focus on the structural features of the image, rather than the fine details. This is crucial for the model to generalize well to unseen data, as it is not expected to see the same images during inference as it did during training.

The noise augmentation function generates random noise using a normal distribution with a mean of 0 and a standard deviation between 0.1 and 0.5. The noise is then added to the image and the resulting image is clamped to a value between 0 and 1 before being returned as an image. The blur augmentation function applies a Gaussian blur to the image using a kernel size of 5, 7, or 9 and a sigma value of 1.5, 2, or 2.5. The kernel size and sigma values were chosen through testing to find the most visually appealing results, i.e. distortions great enough to distort the details of the image, but not so much that the image becomes unrecognizable.

Noise augmentation

The noise augmentation was done by generating random noise and adding it to the image. The noise was generated using a normal distribution with a mean of 0 and a standard deviation between 0.1 and 0.5. The noise was then added to the image and the resulting image was clamped to a value between 0 and 1. The resulting image was then converted back to a PNG image for easier handling.

Blur augmentation

The blur augmentation was done by applying a Gaussian blur to the image. Through testing, the kernel sizes of 5, 7, and 9 were chosen, as well as the sigma values of 1.5, 2, and 2.5. After randomly selecting the combination of kernel size and sigma, the image is converted to a tensor and the blur is applied. The resulting image is then converted back to a PNG image for easier handling.

Examples of the distortion augmentations can be seen in Figure 19. The image to the far left-hand side is the original image, the top row is the noise augmented images, and the bottom row is the blur augmented images. The noise augmented images show the image with added noise, making certain features difficult to see and the image more obscured. The blur augmented images show the image with a Gaussian blur applied, making the image more obscured and less sharp. Seeing these examples, it is clear to see that, again, a large amount of diversity has been introduced to the dataset. The introduction of noise and blur further help the model generalize better by focusing on structural features rather than specific details. This addition to the dataset broadens the ability of the models and teaches them to perform better on suboptimal images.

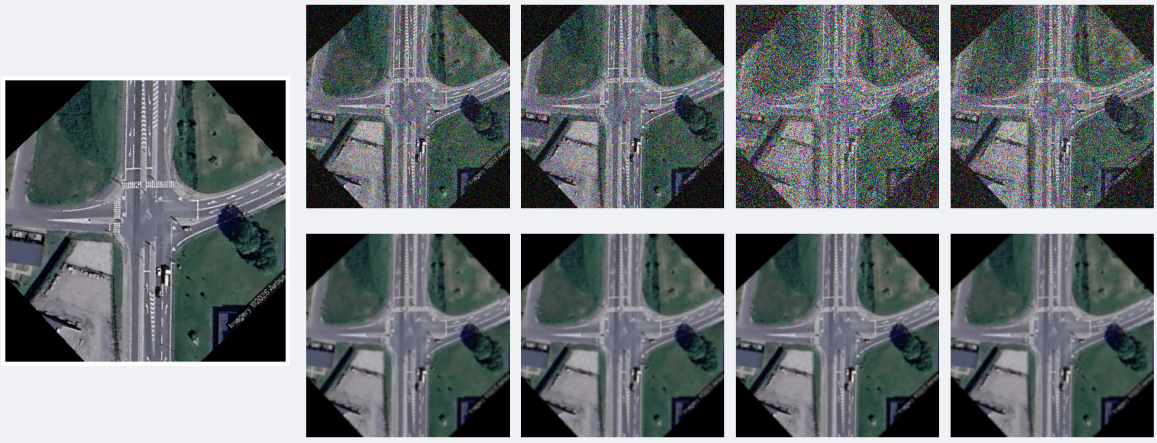


Figure 19. Example of the distortion augmentations. The far left image is the original image, the top row is the noise augmented images, and the bottom row is the blur augmented images.

CROP AND ZOOM are the last augmentations adopted to expand the dataset. They are very common spatial augmentations used on image datasets. They are typically selected to make the models trained more robust against non-centred images and to make the models more adept at generalizing to different scales. If centring is not an issue, then cropping is still a common method for classification tasks, as it helps the model focus on the relevant features of the image. Zooming is also a common augmentation, as it helps the model learn different spatial scales of whatever it is being taught to handle. Also, as discussed earlier in Section 4.1, the zoom level of images in the dataset is set to a value of 18, but these images still appear to consist of vastly different sized intersection, thanks to the fact that some intersection are simply larger than others. Thus, using a zoom augmentation, helps the models generalize better to different scales of intersections. Furthermore, this should also help the model gain a better understanding of road width and not overfit to any particular width present in the dataset.

So, both of these augmentations help the trained models become scale invariant, which is a desired trait of models set to perform the task at hand, as the scale and size of intersection images gotten through satellite images can vary greatly. The crop augmentation was done by randomly cropping the satellite image and its corresponding paths. The crop size is determined by a factor of the original image size, with a default value of 0.8. The cropped image and paths are then resized back to the original dimensions to maintain consistency. This augmentation helps the model focus on different parts of the image and improves its robustness to non-centred features, as is highly relevant for this project. The zoom augmentation was done by resizing the image to a larger size and then cropping it back to the original size. Through testing, the zoom factors of 1.4 to 1.9 were chosen. After randomly selecting a zoom factor, the image is resized and cropped to simulate zooming in.

`Crop augmentation`
The crop augmentation was done by randomly cropping the satellite image and its corresponding paths. The crop size is

`Zoom augmentation`
The zoom augmentation was done by resizing the image to a larger size and then cropping it back to the original size.

determined by a factor of the original image size, with a default value of 0.8. The cropped image and paths are then resized back to the original dimensions to maintain consistency.

Through testing, the zoom factors of 1.4 to 1.9 were chosen. After randomly selecting a zoom factor, the image is resized and cropped to simulate zooming in.

Examples of the crop and zoom augmentations can be seen in Figure 20. The far left image is the original image, the top row is the crop augmented images, and the bottom row is the zoom augmented images. The crop augmented images show the image cropped to a different part of the image, with the associated paths being cropped by the same factors. This was a necessary step, separating these augmentations from the previous, as the paths through the image were not the exact same as when the image underwent colouration or distortion augmentations. The same is true for the zoom augmented image. Here, the paths also needed to undergo the same augmentation as the satellite image. Once again, these augmentation have clearly impacted the diversity of the dataset, as the models are now trained on images that are not centred and images that are zoomed in, which is important for generalization.

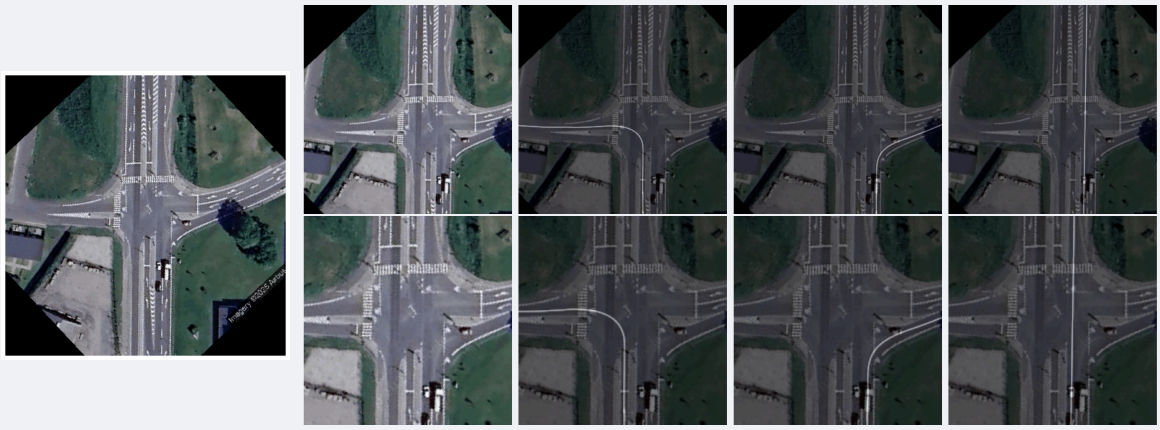


Figure 20. Example of the crop and zoom augmentations. The far left image is the original image, the top row is the crop augmented images, and the bottom row is the zoom augmented images. The three rightmost columns shows how these augmentations are also applied to the paths, saving a lot of time by not having to redraw them.

All in all, these augmentations were selected to expand the dataset by introducing real-world inspired variations of the satellite images. These augmentation mimic the unpredictable conditions posed by using satellite imagery. By employing techniques such as hue and saturation adjustments, greyscale conversions, noise injection, blur filtering, and spatial transformations like cropping and zooming, the augmented dataset now contains diverse lighting, quality, and scale entries. This strategy of employing such a variety of augmentations not only ballooned the size of the dataset immensely, but also made the dataset more representative of the real-world conditions the models are expected to work with, meaning that it should be more robust and better at generalizing to unseen data.

Other augmentation techniques were considered for this project. Rotation, for example, was considered to further increase the diversity of the dataset by introducing variations

in orientation, which could help the model learn to recognize paths from different angles. However, early on in this project, it was decided that the entry for each path should be at the bottom of the satellite image. Thus, rotation the images could have undesired consequences for the models' ability to generalize with this constraint. Flipping was also considered, but was ultimately left out, as it was deemed to have too little of a potential impact since many intersections already go in all directions. Translation was also considered, but was left out, as it would require the paths to be redrawn as they would no longer reach the edges of the image, which was a factor this entire set out to combat. Finally, a very common augmentation used in segmentation, is the act of using cutmix and its constituent parts, namely cutout and mixup. These were, however, also left out, as it is assumed that the satellite images do not contain holes or other artefacts that would be introduced by these augmentations. Furthermore, cutmix might disrupt the spatial features understood by the model. These may be introduced in future work, as they may ultimately increase the robustness of the models in environments that have blocking features.

4.3.3 Dataset Structure ✓

To maintain an ease-of-use principle for this project, the dataset was structured in a way that allows for easy loading of the data. This includes building the dataset in a logical way, and creating a python class that can load the dataset gracefully. This is especially important as the paths in a dataset can vary in number, so custom loading is necessary. Thus, the dataset is structured like shown in the listing below

```
dataset/
├── train/
│   ├── intersection_001/
│   │   ├── satellite.png
│   │   ├── class_labels.npy
│   │   ├── class_label_cold_map.npy
│   │   ├── paths/
│   │   │   ├── path_1/
│   │   │   │   ├── path_line.png
│   │   │   │   ├── path_line_ee.json
│   │   │   │   ├── cold_map.npy
│   │   │   ├── path_2/
│   │   │   │   ├── path_line.png
│   │   │   │   ├── path_line_ee.json
│   │   │   │   ├── cold_map.npy
│   │   │   ├── path_3/
│   │   │   │   ├── path_line.png
│   │   │   │   ├── path_line_ee.json
│   │   │   │   ├── cold_map.npy
├── test/
│   ├── intersection_001/
│   │ ...
```

Listing 5. Folder structure of the dataset. Each `intersection_XXX` folder contains a satellite image of an intersection, class labels and their corresponding cold map, and a `paths` folder containing the paths through the intersection. Each path folder contains the path line, the path's entry and exit points, and the cold map for the path in a `.npy` format.

Firstly, the dataset is split into two separate parts, `train` and `test`. This is done to ensure that the model is not overfitting to the training data. This is achieved by training the model

on the `train` dataset and testing/validating it on the `test` dataset. To ensure that the models generalize well to the task as hand, the `test` dataset should contain intersections that are completely absent from the `train` dataset. This is done to ensure it does not fall into the simple trap of memorizing the training data and created really good results that can be considered false positives as it supposedly has never seen the data before.

This `train`/`test` split in the dataset is created in the folder structure instead of using the simpler functionalities offered by PyTorch. PyTorch offers a `random_split` function from its utility sub-library. This function takes in some dataset declared as a PyTorch `Dataset` object, as shown below in Section 4.3.3.1, and splits it based on a given ratio. This is a simple way to split the dataset, but, as is the case of the created dataset, some images are very similar, meaning that the split does not achieve the desired effect and the model overfits to the training data. Thus, a completely different set of intersections is used for the `test` dataset.

Each `intersection_XXX` folder contains a satellite image saved as a PNG. The ground truth class labels are also saved along with the corresponding cold map. Accompanying these, is the `paths` folder, which contains a folder for each path through the intersection. Each path folder contains the path line image, currently saved as a PNG as well, a JSON file containing the entry and exit points of the path in relation to the image, not the global coordinates, and the corresponding cold map saved as a `.npy` file.

4.3.3.1 Dataset class ✓

To be able to easily load a satellite image and its corresponding paths, entry/exit points, and cold maps, a `IntersectionDataset` class was created, built on top of the PyTorch `Dataset` class. To implement this class, three functions must be created, namely `__init__`, `__len__`, and `__getitem__`.

`__init__` is the function called when the class is instantiated. It initializes the dataset with the root directory of the dataset, a transform function, and a path transform function. The root directory is the directory where the dataset is stored, the transform function is a function that can be applied to the satellite image, and the path transform function is a function that can be applied to the path line. The root directory passed to the instantiation should that of either the training or test dataset within the dataset root folder. The transforms are simply `ToTensor` functions provided by PyTorch. The `__init__` function also creates a list of all intersection folder found at the root directory. The code for the `__init__` function can be seen in Listing 6 below.

```python
def __init__(self, root_dir, transform=None, path_transform=None):
  self.root_dir = root_dir
  self.transform = transform
  self.path_transform = path_transform

  self.intersections = [
    os.path.join(root_dir, f)
    for f in os.listdir(root_dir)
    if os.path.isdir(os.path.join(root_dir, f))
  ]
```

Listing 6. Code snippet of the `__init__` function for the dataset.

`__len__` is another function required by the PyTorch `Dataset` class. It returns the length of the dataset. Thanks to the initialization of the dataset in the `__init__` function, the length of the dataset is simply the number of intersection in the root directory. The code for the `__len__` function can be seen in Listing 7 below.

```python
def __len__(self):
  return len(self.intersections)
```

Listing 7. Code snippet of the `__len__` function for the dataset.

`__getitem__` is one of the most crucial functions of the dataset class. First, the function retrieves the directory of the intersection at the given index. Then, it loads the satellite image from the intersection directory and applies the transform function to it if one is provided. It then loads the `class_labels.npy` and `class_label_cold_map.npy` files from the same intersection directory. If a `path_transform` is provided, it is applied to both `class_labels` and `class_labels_cmap`. These transforms are typically `ToTensor` as provided by PyTorch. All of this data is then stored in a dictionary with keys `satellite`, `class_labels`, and `class_labels_cmap` and returned as the sample. The code for the `__getitem__` function can be seen in The code for the `__getitem__` function can be seen in Listing 8 below.

```python
def __getitem__(self, idx):
  intersection_dir = self.intersections[idx]

  satellite_path = os.path.join(intersection_dir, 'satellite.png')
  satellite_img = Image.open(satellite_path).convert('RGB')

  if self.transform:
    satellite_img = self.transform(satellite_img)

  class_labels_path = os.path.join(intersection_dir, 'class_labels.npy')
  class_labels = np.load(class_labels_path)

  class_labels_cmap = os.path.join(intersection_dir, 'class_label_cold_map.npy')
  class_labels_cmap = np.load(class_labels_cmap)

  if self.path_transform:
    class_labels = self.path_transform(class_labels)
    class_labels_cmap = self.path_transform(class_labels_cmap)

  sample = {
    'satellite': satellite_img,
    'class_labels': class_labels,
    'class_labels_cmap': class_labels_cmap
  }
  return sample
```

Listing 8. Code snippet of the `__getitem__` function for the dataset.

The dataset is then simply instantiated as such:

```python
dataset = IntersectionDataset(root_dir=dataset_dir,
                              transform=ToTensor(),
                              path_transform=ToTensor())
```

Listing 9. Instantiation of the dataset.

where `dataset_dir` is the path to either the training or test dataset folders. Creating the dataloader is as simple as:

```
dataloader = DataLoader(dataset,
                        batch_size=b,
                        shuffle=True,
                        num_workers=num_workers)
```

Listing 10. Creating a dataloader for the dataset.

Arguments passed to the `DataLoader` initializer are the dataset from Listing 9, the batch size, whether the dataset should be shuffled, the number of workers to use for loading the data, and the ability to give it a custom collate function. `num_workers` is found using the `multiprocessing` library as it easily finds the number of available computation cores. Finally, the dataset is then ready to be used in a training loop. The dataloader will yield batches of data, where each batch is a dictionary containing the satellite image, class labels, and class label cold map. This makes it easy to iterate over the dataset and train the models on the data.

With the loss functions and dataset in place, the models which are to be subjected to these will be presented in the next section. They will set out to answer Research Question **RQ-1**, as two of them are convolution based and the other two are transformer based.

4.4 The Models ✓

With all the previous sections covered, the next step is to define and discuss the models that will be used to predict the path through intersection. The goal is to create and test various models that do this task well, evaluating their performance and comparing them to each other. The models used are commonly used as the back-bone of other, larger models designed for specific tasks. This approach was chosen as it may provide a better understanding of what kind of backbone for a model might yield the greatest results in the context of path-planning.

Chosen for this project are two groups of models, those that are based on convolution and those that are based on transformers. This is done to answer Research Question **RQ-1**. The specific models chosen, as presented in detail in the following sections, are the convolution-based U-Net and DeepLabV3+ models, and the transformer-based Vision Transformer and Swin Transformer models. These models were chosen largely due to their popularity within the field of computer vision, and their ability to perform well on a variety of tasks. Comparing these distinct groups of models, will also highlight the differences between them, and how they perform on the task at hand. Additionally, these models were chosen in their bare-bone state, as hardware limitations hindered the usage of vast and complex networks.

Furthermore, it should be noted that other DL methodologies were considered. First, is reinforcement learning, which was not chosen as it would add unnecessary complexity to the project. Second, is the use of generative models, which were also not chosen as they introduce a massive change in training algorithm. Both of these points highlight the fact

that their implementation would require a significant paradigm shift in this project, which was deemed to go beyond the scope of this thesis.

The following sections will present the chosen model, starting with the convolution-based models, followed by the transformer-based models. First, the U-Net model proposed in 2015 [37] will be presented, focusing on its reliance of classical convolution and usage of alternative convolutional methods and skip connections. DeepLab was originally proposed in 2016 [74] with the implemented V3+ version introduced in 2018 [75]. With this model, the focus will be on the use of atrous convolution and the ASPP module. The first of the transformer-based models is the Vision Transformer (ViT), presented in Section 2.2.1, which introduced the transformer architecture to the field of computer vision. Finally, the Swin Transformer is a hierarchical transformer model that introduced shifted windows to the transformer architecture, allowing for a more efficient computation of self-attention.

4.4.1 U-Net ✓

The U-Net architecture was released in a landmark paper 2015, having since garnered over 100,000 citations. U-Net gets its very literal name from its architecture, which resembles a wide letter U. The U-Net architecture is classed as a Fully Convolutional Network (FCN), which is a type of network that is particularly effective for image segmentation tasks.

The architecture consists of two main parts: the encoder and the decoder. Referring to Figure 21, the encoder part of the network is the left-hand side portion down to the bottleneck, with the decoder being the following layers. First, the encoder increases the number of channels while decreasing the spatial dimensions of the input image. This is done in layers. The first layer is a double convolutional layer●, which consists of two convolutional layers, each followed by a batch normalization and ReLU activation. This is followed by a max pooling layer●, which reduces the spatial dimensions of the image. The second layer is another double convolutional layer●, which again increases the number of channels while decreasing the spatial dimensions. This process continues until the bottleneck is reached.

The decoder then reverses this process. After the bottleneck, the decoder begins with an upsampling layer●, which increases the spatial dimensions of the image again. This is followed by a concatenation with the corresponding encoder layer, which allows the model to retain spatial information lost during the downsampling process. Skip connections are commonly used in various architectures, as they help to retain spatial information. Following the concatenation, another double convolutional layer● is applied, which reduces the number of channels. This process continues until the final output layer is reached. In the final layer, a double convolution is initially applied after the concatenation, followed by a final convolutional layer●. This final convolutional layer reduces the number of channels to the number of classes in the segmentation task. The final output is a segmentation map, which indicates the predicted class for each pixel in the input image.

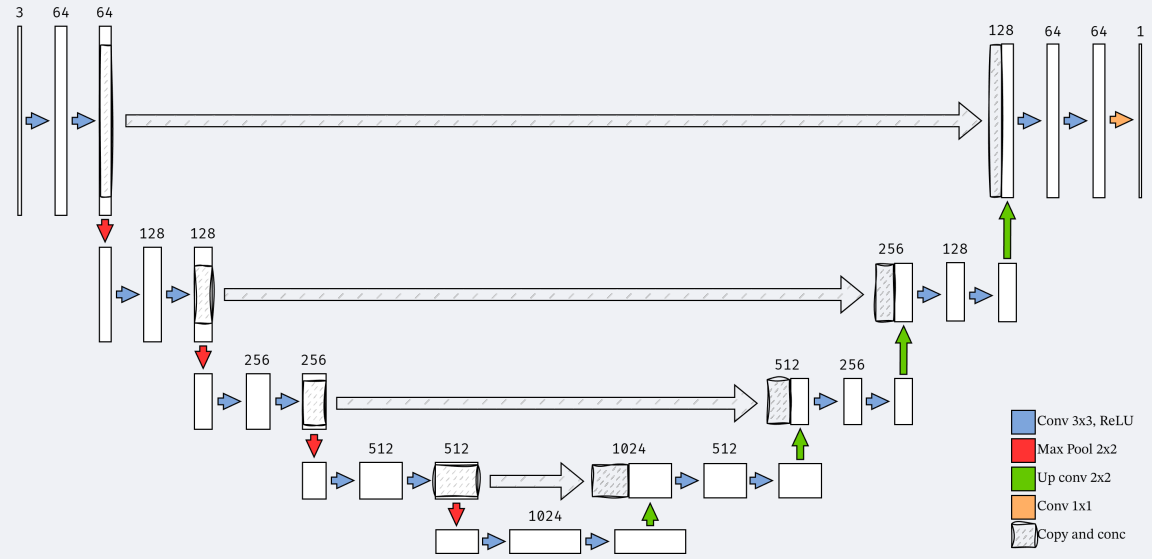


Figure 21. U-Net Architecture.

This architecture significantly advanced the field of image segmentation. One of its key innovations is its extensive use of skip connections, which directly fuse the contextual information from the encoder with the spatial detail from the decoder. This mechanism helps the model recover fine-grained spatial features that are often lost during the down-sampling process. Additionally, U-Net employs double convolutional layers rather than single convolutions. This structure allows the model to extract more complex and abstract features from the input data. The architecture is also fully convolutional, meaning it avoids fully connected layers altogether. As a result, U-Net can be applied to images of varying sizes without needing architectural modifications, making it highly versatile across different datasets and tasks.

U-Net was originally introduced to address the challenge of limited labelled data in the medical imaging field, where precise segmentation is critical and annotated samples are scarce. However, its effectiveness has extended far beyond this initial scope. It has inspired a wide range of variants, including U-Net++, Attention U-Net, ResUNet, and Mobile U-Net, each adapted for specific applications or resource constraints.

Despite its strengths, U-Net is not without limitations. It can be memory-intensive and computationally demanding, particularly due to the skip connections. These connections require storing high-resolution feature maps from earlier layers in memory, which increases the model's footprint and can limit scalability. Nevertheless, U-Net's widespread success, ability to generalize well on small datasets, and strong performance in segmentation tasks make it an ideal candidate for this project.

For this project, the U-Net architecture was implemented using the implementation by milesial[5] with minor modifications. This implementation features a modular design with clearly defined components for the encoder and decoder paths. Each downsampling step comprises a max-pooling layer followed by a double convolutional block, while each

[5]Their github repository is found here: https://github.com/milesial/Pytorch-UNet with the implemented version found in this project's GitLab repository.

upsampling step includes either bilinear interpolation or transposed convolution, concatenation with the corresponding encoder feature map, and another double convolutional block. The model supports flexibility in input image sizes due to its fully convolutional nature. Training routines are provided as well, but new ones were ultimately implemented for easier switching between this project's models.

In this project, U-Net is the simplest of the models explored for the task of predicting paths through an intersection based solely on satellite imagery. Rather than serving as a segmentation backbone or part of a larger system, the U-Net model is trained end-to-end to directly output the predicted traversal path. While U-Net is traditionally used for semantic segmentation tasks, its encoder-decoder structure and spatial information retention via skip connections make it a suitable candidate for this kind of spatially grounded prediction. It is acknowledged, however, that U-Net alone is unlikely to yield highly precise results in such a complex task. Nonetheless, its simplicity and proven performance on pixel-wise prediction problems make it a valuable baseline for evaluating how well standard architectures can handle intersection-based path planning.

4.4.2 DeepLabV3+ ✓

The DeepLab family of models is a series of models designed for semantic segmentation tasks by Google. The original DeepLab model was introduced in 2016 and has since evolved into several versions, with DeepLabV3+ being the latest iteration.

DeepLab, contextually DeepLabV1, was the first model of the family to be released in 2016 [74]. It introduced several innovations, chief amongst which was the idea of using atrous convolution, also known as dilated convolution. Shown in Figure 22, atrous convolution is a method where holes are introduced in the convolutional kernels. This allows each pixel in the resulting feature map to capture a wider context, without having to using massive kernels. Formally, this technique allows for the extraction of multi-scale features without losing resolution, making it particularly effective for semantic segmentation tasks. Furthermore, the model, and its successor, used a fully connected conditional random field (CRF). This is a post-processing step used to refine the pixel-level labellings produced by the network. It operates by defining an energy function that includes costs for assigning labels to individual pixels, called unary potentials, and costs based on the label assignments of all possible pairs of pixels, called pairwise potentials. The successor, DeepLabV2, also used this post-processing step, but not in later versions.

DeepLabV2, utilized the concept of Atrous Spatial Pyramid Pooling (ASPP), which employs multiple parallel atrous convolutions with different rates to capture features at various scales. The "pyramid" part of its name comes from the fact that it uses multiple atrous convolutions with different rates, effectively creating a pyramid of features at different scales. The introduction of this technique can be compared to the popularization of skip connection's usage in U-Net, garnering over 25,000 citations.

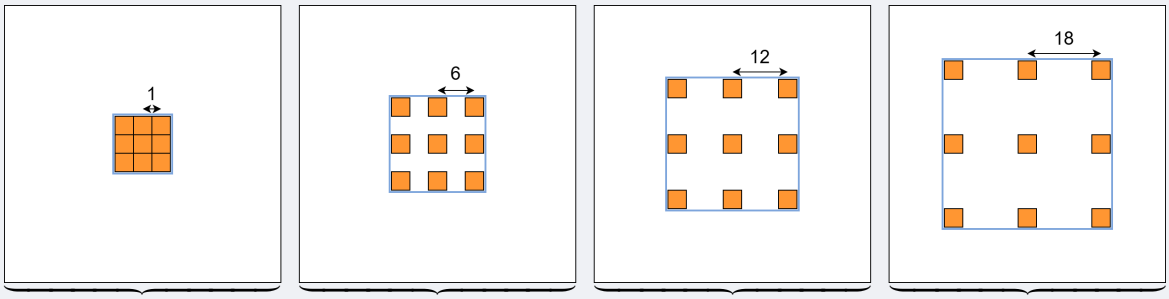


Figure 22. Atrous Convolution.

DeepLabV3 was released in 2017 [76], introducing an enhanced ASPP module, image pooling, and batch normalization. The ASPP module was improved by adding image pooling, which allows the model to capture global context information. This is done by applying a global average pooling operation to the feature map. Finally, DeepLabV3+ was released in 2018 [77], which turned the DeepLab architecture into an encoder-decoder structure. This was done by adding a decoder module to the DeepLabV3 architecture. So, the encoder part● of the models is the same as the DeepLabV3 model, which consists of a backbone network, in this project MobileNetV3 [78], and the ASPP module. The top half of Figure 23 shows the encoder part of the model. Five features maps are concatenated in the encoder. First, a simple 1x1 convolution● is applied. Second, atrous convolution is used to create 3 different feature maps●, each with a different dilation rate. Finally, an image pooling● feature map is concatenated to capture global context.

In the decoder part●, the model uses a low-level feature map● from the encoder. This feature map is concatenated with the output of the ASPP module after being passed through another 1x1 convolution●. A final segmentation head is then applied to the concatenated feature map before being upscaled to the original image size. The segmentation head consists of a 3x3 convolutional layer followed by a bilinear upsampling layer. This allows the model to produce a segmentation map that is the same size as the input image.

This family of models achieves high results on various datasets. From the outset, DeepLabV1 achieved a mean intersection over union (mIoU) of 71.6% on the PASCAL VOC 2012 dataset using VGG-16 as the backbone of the network. DeepLabV2 improved this to 79.7% mIoU using ResNet-101 as the backbone on the same dataset. DeepLabV3 further improved this to 82.7% mIoU using ResNet-101 as the backbone on the PASCAL VOC 2012 dataset. Finally, DeepLabV3+ achieved 89.0% mIoU using Xception as the backbone on the same dataset, and 82.1% mIoU on the Cityscapes dataset. This shows that the DeepLab family of models is capable of achieving high results on various datasets.

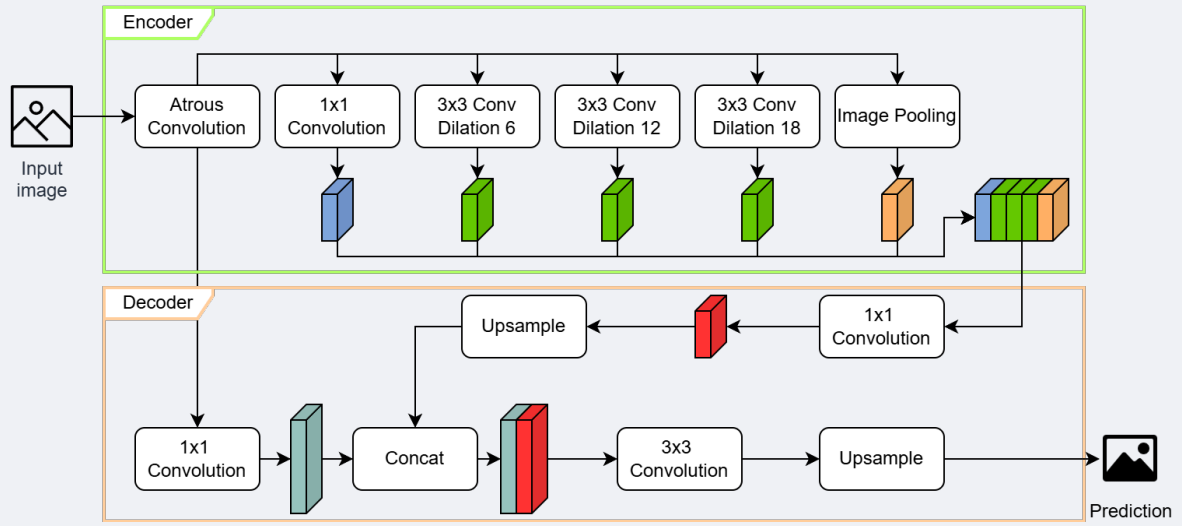


Figure 23. DeepLabV3+ Architecture.

The implementation of this model follows the network shown in Figure 23. The backbone, as mentioned, is the MobileNetV3 network. This network was chosen as it is a lightweight network, capable of running on the machine used in this project and for a more simplistic implementation. Other backbones are available and have active communities. The original paper was proposed with the ResNet-101 and Xception backbones, considered to be classic CNNs. Larger networks like EfficientNet-L2 and ConvNeXt are also viable options for more complex tasks. The base version of ConvNeXt was tested, namely ConvNeXt-base, but simply crashed when initialized. There's even work being done to integrate the DeepLab architecture with transformer-based backbones, such as ViT and Swin Transformer.

In summary, the two convolution-based models chosen for this project are the U-Net and DeepLabV3+ models. The U-Net model is a fully convolutional network that uses skip connections to retain spatial information, while the DeepLabV3+ model uses atrous convolution and ASPP to capture multi-scale features. Both models have been shown to achieve high results on various datasets, making them suitable for the task at hand. The next sections will present the transformer-based models chosen for this project.

4.4.3 Vision Transformer ✓

For this project, two of the most widely used transformer-based models were chosen, the Vision Transformer (ViT) and the Swin Transformer. The ViT was presented in detail in Section 2.2.1, highlighting it being the first connection of the NLP method called Attention and computer vision.

The ViT model is a pure transformer model, meaning it does not use any convolutional layers. Instead, it relies on self-attention mechanisms to process the input data. The model consists of several key components, including a linear projection layer●, an embedding layer●, an encoder●, and a segmentation head●. The linear projection layer is responsible for transforming the input data into a format suitable for the transformer architecture. The embedding layer then maps the input data into a higher-dimensional space, allowing

the model to capture more complex relationships between the input features. The encoder consists of multiple transformer blocks that apply self-attention and feed-forward networks to process the input data. Finally, the segmentation head generates the output predictions based on the processed features.

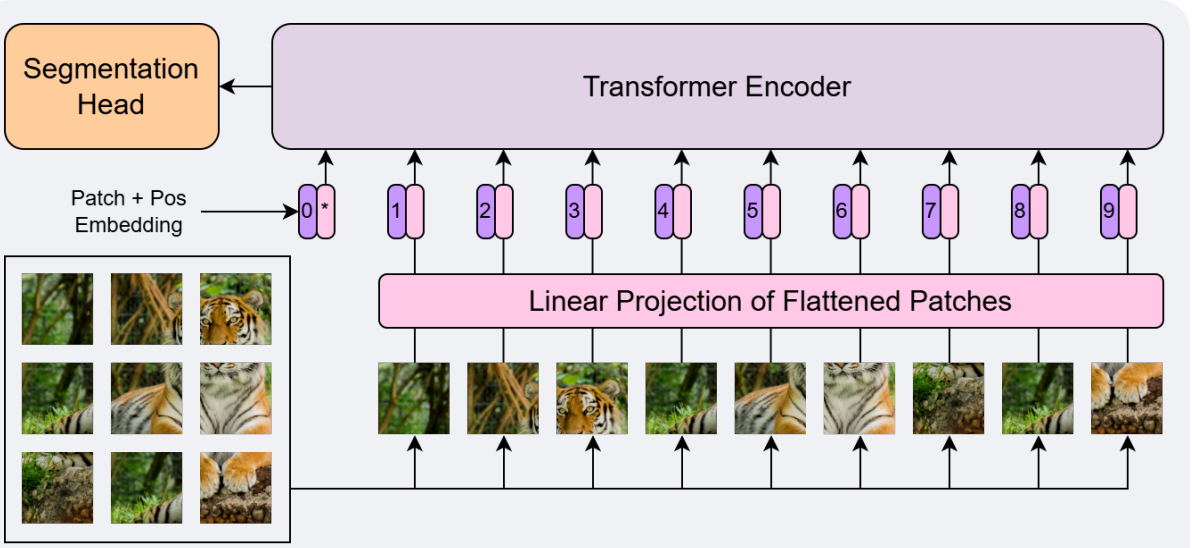


Figure 24. Vision Transformer.

Much of this functionality is offered by the PyTorch framework, where the ViT model is implemented as a class. The largest difference from the base ViT presented in Section 2.2.1 is the change from the classification head to the segmentation head. The core of the network therefore remains unchanged: a single-stride convolutional projection decomposes the 400×400 input image into non-overlapping 16×16 patches ($G = 25$ along each axis), each of which is flattened and linearly embedded. A learnable class token is prepended so that the model structure remains identical to the original ViT, but it is discarded after the encoder since the downstream task is dense prediction rather than classification.

The transformer encoder is an unmodified stack of twelve blocks with 768 hidden units, 12 self-attention heads, and a 3072-dimensional MLP. A dropout of 0.2 follows the encoder output. The sequence of patch embeddings is then reshaped back to its 2-D layout, yielding a feature map of shape [B, 768, 25, 25]. A lightweight segmentation head — implemented as a single 1×1 convolution—projects this map to the required five classes; bilinear upsampling restores the original spatial resolution.

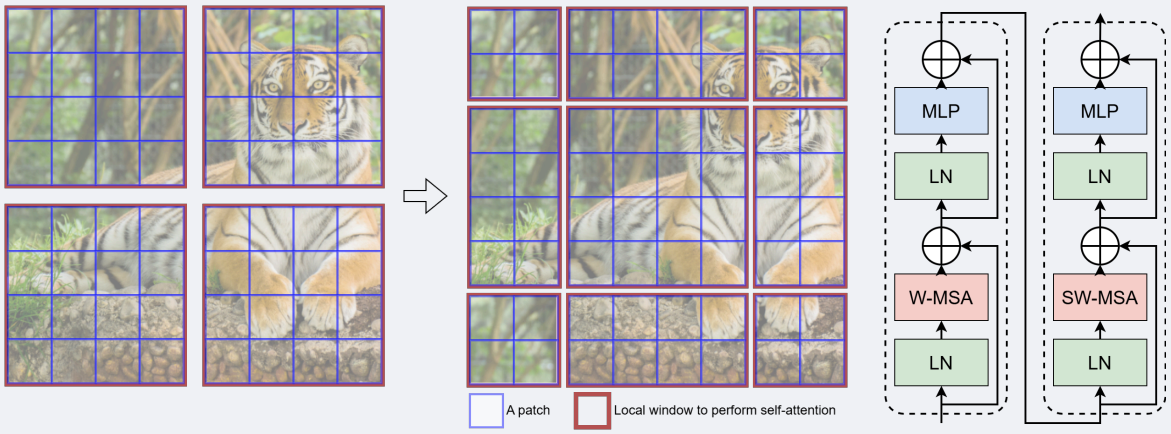
Training from random initialisation removes any dependence on large-scale external datasets and keeps the experimental comparison fair with respect to U-Net and DeepLabV3+. The trade-off is a significantly longer convergence time and higher memory usage: the quadratic cost of self-attention makes the ViT approximately four times more memory-intensive than U-Net at 400×400 resolution. Even so, the architecture's ability to capture global context is attractive for the intersection-traversal task, where the correct path at one corner may rely on cues many tens of metres away in the satellite frame.

In summary, the ViT configuration employed here can be viewed as the minimal modification of the canonical ViT architecture for dense prediction: keep the patch projection and encoder intact, drop the class token after encoding, and append a shallow 1×1 con-

volutional head plus up-sampling. Despite the absence of pre-training, this arrangement preserves ViT's long-range modelling capability while producing pixel-aligned outputs suitable for the task at hand. While the ViT offers a compelling baseline for transformer-only segmentation, its flat token structure and quadratic attention cost can become prohibitive at higher resolutions. Recent research therefore turns to hierarchical designs that marry the long-range reasoning of self-attention with the computational efficiency of local windows.

4.4.4 Swin Transformer ✓

The Swin Transformer extends the pure-attention idea of ViT with a hierarchical, window-based design that scales gracefully to high-resolution images. As reviewed in Section 2.2.1, Swin partitions the feature map into fixed-size, non-overlapping windows● and computes self-attention only within each window. Figure 25a visualises this arrangement for a 2×2 window composed of 4×4 patches●. The quadratic cost of attention is now bound by the window area rather than the entire image. Figure 25b shows two successive Swin transformer blocks. These ensure the exchange of information across window borders, where the second block● in a pair shifts the window grid by half the window size, so that the pixels sitting on window edges in the first block● lie at the centre of a window in the next. This "shifted window" scheme allows cross-window interactions with only a negligible increase in computation, while still preserving the locality that makes the model efficient.



(a) Shifted Window.

(b) Two Successive Swin Transformer Blocks.

Figure 25. Swin Transformer components.

A single Swin transformer block, shown in Figure 25b, follows the canonical transformer ordering of LayerNorm●, multi-head self-attention●, and a feed-forward network●, but with two key modifications. First, the attention module is either a standard Window-MSA (W-MSA) or its shifted counterpart (SW-MSA), depicted by the ●-coloured block. Second, the feed-forward network is implemented as a two-layer MLP● with a GELU activation between the layers. Gaussian Error Linear Units (GELU) is an activation function akin to RELU, but has a Gaussian distribution function, resulting in what is considered to be

a smooth version of RELU. Residual connections wrap both the attention and MLP sub-modules, enabling stable training of deep hierarchies.

The original Swin architecture is built from four such stages, each consisting of an alternating sequence of W-MSA and SW-MSA blocks. Between stages, a patch-merging layer concatenates neighbouring 2×2 tokens and projects them to twice the channel dimension, halving both spatial resolution and token count while increasing representational capacity. The resulting $\frac{H}{4}, \frac{H}{8}, \frac{H}{16}, \frac{H}{32}$ feature pyramid closely mirrors the spatial hierarchy of convolutional backbones and therefore integrates naturally with encoder-decoder segmentation heads.

The implementation employed in this work instantiates the `swin_base_patch4_window7_224` variant through `timm`, but with three modifications to suit the task. First, the image size is set to 400×400 so that the initial 4×4 patch projection still yields a token grid divisible by the 7×7 window. Second, the model is created with features_only=True, exposing the output of each stage; however, only the final stage is used for the present, lightweight decoder. Third, a small segmentation head replaces the classification head: a 3×3 convolution reduces the 1024-channel backbone output to 256, a ReLU introduces non-linearity, and a 1×1 projection produces the five class logits, which are finally up-sampled bilinearly to the original resolution.

Finally, with all the models presented, the next step is to discuss the training strategy used for all models. The training strategy is a crucial part of the model, as it defines how the model will be trained and what kinds of loss functions will be used and how they will be used together.

4.4.5 Training Strategy ✓

This section will give a comprehensive overview of the training strategy used for all models. All models will be trained using a combination of loss functions, an optimizer with weight decay, and a learning rate scheduler. Finally, a different number of epochs will be used for each combination of loss functions, as some are simply slower than others, resulting in very long training times.

4.4.5.1 Loss Functions ✓

There are three loss functions that will be used in this project. Cross-entropy (CE) loss is the main driving force behind training the models to correctly label each pixel in the image. CE will be used in combination with both of the topology loss functions separately. The topology loss functions will not trained with the intention of reaching the same level of results, thus it is only the novel cold map based loss that will be used to train the models on its own. This is mainly to show the effectiveness of the loss function in shaping the model's predictions.

To train with a combination of loss functions, the CE loss will be combined with the topology based ones. It is a common tactic to combine loss functions in order to achieve better results. Loss functions are typically combined by simply adding them together, with a

weight for each loss function. This is done to balance the contribution of each loss function to the overall loss. In order to achieve stable training, this weight is typically set such that each loss function contributes equally to the overall loss, i.e. the weights add up to 1. This is achieved by combining the loss functions like this:

$$\mathcal{L}_{\text{total}} = \alpha \cdot \mathcal{L}_1 + (1 - \alpha) \cdot \mathcal{L}_2 \tag{41}$$

thus by setting $0 \leq \alpha \leq 1$ the contribution of each loss function can be controlled. For example, if $\alpha = 0.5$, then both loss functions contribute equally to the overall loss. And of course if you set $\alpha = 1$, then only the first loss function contributes to the overall loss and vice versa if $\alpha = 0$. For this project, the value of α is chosen dynamically during training. The value for alpha follows the following function:

$$\alpha(\text{epoch}) = \begin{cases} \alpha_{\text{hi}}, & \text{if epoch} < T_{\text{warm}} \\ \alpha_{\text{hi}} - (\alpha_{\text{hi}} - \alpha_{\text{lo}}) \cdot r, & \text{if } T_{\text{warm}} \leq \text{epoch} < N_{\text{epochs}} \end{cases} \tag{42}$$

with $r = \frac{\text{epoch} - T_{\text{warm}}}{\max(1, N_{\text{epochs}} - T_{\text{warm}})}$, where α_{hi} and α_{lo} are the high and low values for alpha, respectively, T_{warm} is the warm-up period, i.e. the stage where α_{hi} is kept as the α value, and N_{epochs} is the total number of epochs. Once past T_{warm}, the value of α will linearly decrease to α_{lo}, which it will hit at the end of training. Below is a table showing the values for α_{hi}, α_{lo}, and T_{warm} for each combination of loss functions, as well as the number of epochs they are trained on:

	Cross-entropy	CE + Cold Map	CE + Continuity	Cold Map
α_{hi}	—	0.9	0.99	—
α_{lo}	—	0.5	0.5	—
T_{warm}	—	10	30	—
N_{epochs}	300	100	100	50
S_{epochs}	{10, 20, 50, 100, 300}	{10, 20, 50, 100}	{10, 20, 50, 100}	{10, 20, 50}

Table 1. Values for α_{hi}, α_{lo}, T_{warm}, and N_{epochs} for the different loss function combinations. S_{epochs} is the set of epochs at which the models are checkpointed.

What this means, is that the CE loss will always be the main driving force during training, while topology based losses will gradually become more influential, but never take over more than half of the total loss. The values shown are for all models. The table also shows the number of epochs used for both CE and the cold map loss as standalone. CE on its own is meant to serve as a baseline for the other combined losses. Finally, the epochs at which the models are checkpointed are shown by S_{epochs}.

4.4.5.2 Optimizer and Scheduler ✓

The optimizer used for all models is the AdamW optimizer [79], which is a variant of the Adam optimizer that includes weight decay. Unlike the standard Adam optimizer, which incorporates L2 regularization by adding a penalty term to the loss function, AdamW decouples weight decay from the gradient-based optimization step. This distinction is important because, in Adam, the interaction between L2 regularization and adaptive

learning rates can lead to suboptimal convergence behaviour. In contrast, AdamW applies weight decay directly to the weights during the parameter update step, independently of the gradient computation.

L2 regularization in Adam is implemented top modify the gradient like this:

$$g_t = \nabla f(\theta_t) + w_t \theta_t \tag{43}$$

where w_t is the regularization coefficient, or rate of decay, and θ_t is the weight at time t. This blends the decay term with the gradient, making it sensitive to the optimizer's internal adaptive mechanisms. What AdamW does instead is to adjust the weight decay term to appear in the gradient update step:

$$\theta_{t+1,i} = \theta_t, i - \eta \left(\frac{1}{\sqrt{\hat{v}_t + \varepsilon}} \cdot \hat{m}_t + w_{t,i} \theta_{t,i} \right), \forall t \tag{44}$$

where $\hat{m}_t$ and $\hat{v}_t$ are the bias-corrected first and second moment estimates of the gradients, respectively, and $w_{t,i}$ is the weight decay coefficient. This means that the weight decay is applied directly to the weights during the update step. By separating the decay term from the loss gradient, AdamW ensures a more consistent regularization effect and improves generalization. Thus, AdamW is chosen as the optimizer for all models, as it is an enhanced version of Adam that provides better performance in many scenarios.

In combination with the AdamW optimizer, a learning rate scheduler is used to adjust the learning rate during training. The scheduler used is the cosine annealing scheduler, which gradually reduces the learning rate from an initial value to a minimum value over a specified number of epochs before resetting it to the initial value. The cosine annealing scheduler is defined as:

$$\eta_t = \eta_{\min} + \frac{1}{2}(\eta_{\max} - \eta_{\min}) \left(1 + \cos \left(\frac{T_{\text{cur}}}{T_i} \pi \right) \right) \tag{45}$$

where $\eta_{\min}$ and $\eta_{\max}$ are the ranges of the learning rate, with $\eta_{\max}$ being the initial learning rate, T_{cur} account for how many epochs have been performed since the last restart, and T_i is the total number of epochs. This form of cosine annealing is known as warm restarts, and is used to improve the performance of the model by allowing it to escape local minima and explore the loss landscape more effectively. This is the main trait desired of the scheduler for this project, as the models should really explore the loss landscape to find the best possible solution.

Other considerations for the scheduler include exponential decay and plateau decay. Exponential decay is a simple and effective way to reduce the learning rate over time, but it can be too aggressive and lead to premature convergence. Plateau decay, on the other hand, is more adaptive as it lowers the learning rate if the validation loss does not improve for a certain number of epochs.

With the training strategy now fully defined — including the choice of loss functions, optimizer, and scheduler — the stage is set to evaluate how each model performs under

these conditions. The following chapter presents the results of these experiments, comparing both the quantitative metrics and qualitative outputs of each configuration.

5 Results

This section details the experiments conducted

6 Discussion

In this section...

6.1 Integration with existing systems

6.2 Shortcomings

6.3 Other considerations

6.4 Ablation

MAYBE, if time allows

Hyperparameters, scheduler (cosann vs exp), optimizer (adam vs adamw), weight initialization

7 Conclusion

References

[1] C. M. University, “The Carnegie Mellon University Autonomous Land Vehicle Project.” [Online]. Available: https://www.cs.cmu.edu/afs/cs/project/alv/www/index.html

[2] C. Thorpe, M. Hebert, T. Kanade, and S. Shafer, “Vision and navigation for the Carnegie-Mellon Navlab,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 10, no. 3, pp. 362–373, 1988, doi: 10.1109/34.3900.

[3] J. Billington, “The Prometheus project: The story behind one of AV's greatest developments.” [Online]. Available: https://www.autonomousvehicleinternational.com/features/the-prometheus-project.html

[4] A. S. Francisco, [Online]. Available: https://abc7news.com/post/wrong-waymo-driverless-car-goes-oncoming-traffic-tempe-arizona/15238556/

[5] “Technical milestone in road safety: experts praise Volkswagen's Car2X technology.” [Online]. Available: https://www.volkswagen-newsroom.com/en/press-releases/technical-milestone-in-road-safety-experts-praise-volkswagens-car2x-technology-5914

[6] K. Dresner and P. Stone, “A multiagent approach to autonomous intersection management,” *Journal of artificial intelligence research*, vol. 31, pp. 591–656, 2008.

[7] A. P. Chouhan and G. Banda, “Autonomous Intersection Management: A Heuristic Approach,” *IEEE Access*, vol. 6, no. , pp. 53287–53295, 2018, doi: 10.1109/ACCESS.2018.2871337.

[8] Z. Zhong, M. Nejad, and E. E. Lee, “Autonomous and Semiautonomous Intersection Management: A Survey,” *IEEE Intelligent Transportation Systems Magazine*, vol. 13, no. 2, pp. 53–70, 2021, doi: 10.1109/MITS.2020.3014074.

[9] M. Cederle, M. Fabris, and G. A. Susto, “A Distributed Approach to Autonomous Intersection Management via Multi-Agent Reinforcement Learning.” [Online]. Available: https://arxiv.org/abs/2405.08655

[10] On-Road Automated Driving (ORAD) committee, “Taxonomy and definitions for terms related to on-road motor vehicle automated driving systems,” SAE International, 400 Commonwealth Drive, Warrendale, PA, United States, 2017.

[11] On-Road Automated Driving (ORAD) committee, “Taxonomy and definitions for terms related to driving automation systems for on-road motor vehicles,” SAE International, 400 Commonwealth Drive, Warrendale, PA, United States, 2021.

[12] S. E. Staff, “The 6 Levels of Vehicle Autonomy Explained.” [Online]. Available: https://www.synopsys.com/blogs/chip-design/autonomous-driving-levels.html

[13] K. Buchholz, “Infographic: Cars Increasingly Ready for Autonomous Driving.” [Online]. Available: https://www.statista.com/chart/25754/newly-registered-cars-by-autonomous-driving-level/

[14] K. Buchholz, “Autonomous driving is taking over.” [Online]. Available: https://www.weforum.org/stories/2023/02/charted-autonomous-driving-accelerating-mobility/

[15] F. A. Mall, “The History of Cruise Control | Folsom Auto Mall.” [Online]. Available: https://www.folsomautomall.com/blog/2022/november/21/the-history-of-cruise-control.htm

[16] F. Danmark, “Ford BlueCruise | Håndfri kørsel i Blue Zones | Ford DK.” [Online]. Available: https://www.ford.dk/om-os/foererassistancesystemer/ford-bluecruise

[17] C. Hoffmann, [Online]. Available: https://www.shop4tesla.com/en/blogs/news/tesla-fsd-supervised-europa-2025

[18] P. Davies, [Online]. Available: https://autoseu.com/you-can-now-drive-partially-hands-free-in-france-this-is-whats-changing/

[19] C. Murray, [Online]. Available: https://www.designnews.com/autonomous-vehicles/here-s-why-level-5-autonomous-cars-may-still-be-a-decade-away

[20] J. Marie, “The Power of Perception: Cameras in Self-Driving Cars.” [Online]. Available: https://supplybridge.com/the-power-of-perception-cameras-in-self-driving-cars/

[21] J. Cohen, “Sensor Fusion - Fusing LiDARs & RADARs in Self-Driving Cars.” [Online]. Available: https://www.thinkautonomous.ai/blog/sensor-fusion/#:~:text=A%20Kalman%20filter%20can%20be,prediction

[22] D. J. Yeong, K. Panduru, and J. Walsh, “Exploring the Unseen: A Survey of Multi-Sensor Fusion and the Role of Explainable AI (XAI) in Autonomous Vehicles,” *Sensors*, vol. 25, no. 3, 2025, doi: 10.3390/s25030856.

[23] J. Cohen, “9 Types of Sensor Fusion Algorithms.” [Online]. Available: https://www.thinkautonomous.ai/blog/9-types-of-sensor-fusion-algorithms/

[24] W. Franklin, “Kalman Filter Explained Simply - The Kalman Filter.” [Online]. Available: https://thekalmanfilter.com/kalman-filter-explained-simply/

[25] W. S. McCulloch and W. Pitts, “A logical calculus of the ideas immanent in nervous activity,” *Bulletin of Mathematical Biophysics*, vol. 5, no. 4, pp. 115–133, Dec. 1943.

[26] Z. Parvez, “The Pioneers of AI: Marvin Minsky and the SNARC.” [Online]. Available: https://zahid-parvez.medium.com/history-of-ai-the-first-neural-network-computer-marvin-minsky-231c8bd58409

[27] [Online]. Available: https://www.ibm.com/history/early-games

[28] F. Rosenblatt, “The perceptron: a probabilistic model for information storage and organization in the brain,” *Psychol. Rev.*, vol. 65, no. 6, pp. 386–408, Nov. 1958.

[29] R. Karjian, “History and evolution of machine learning: A timeline.” [Online]. Available: https://www.techtarget.com/whatis/feature/History-and-evolution-of-machine-learning-A-timeline

[30] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, “Learning representations by back-propagating errors,” *Nature*, vol. 323, no. 6088, pp. 533–536, Oct. 1986.

[31] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," in *Advances in Neural Information Processing Systems*, F. Pereira, C. Burges, L. Bottou, and K. Weinberger, Eds., Curran Associates, Inc., 2012, p. . [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf

[32] [Online]. Available: https://www.wikiwand.com/en/articles/Gradient_descent

[33] B. News, "Artificial intelligence: Google's AlphaGo beats Go master Lee Se-dol." [Online]. Available: https://www.bbc.com/news/technology-35785875

[34] OpenAI, "OpenAI Five defeats Dota 2 world champions." [Online]. Available: https://openai.com/index/openai-five-defeats-dota-2-world-champions/

[35] G. Boesch, "Computer Vision Tasks (Comprehensive 2025 Guide) - viso.ai." [Online]. Available: https://viso.ai/deep-learning/computer-vision-tasks/

[36] [Online]. Available: https://openai.com/index/introducing-4o-image-generation/

[37] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation." [Online]. Available: https://arxiv.org/abs/1505.04597

[38] J. Deign, "How ChatGPT changed… well, almost everything." [Online]. Available: https://newsroom.cisco.com/c/r/newsroom/en/us/a/y2024/m12/how-chatgpt-changed-well-almost-everything.html

[39] A. Vaswani *et al.*, "Attention Is All You Need." [Online]. Available: https://arxiv.org/abs/1706.03762

[40] A. Dosovitskiy *et al.*, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale." [Online]. Available: https://arxiv.org/abs/2010.11929

[41] Z. Liu *et al.*, "Swin Transformer: Hierarchical Vision Transformer using Shifted Windows." [Online]. Available: https://arxiv.org/abs/2103.14030

[42] P. Sun *et al.*, "Waymo Open Dataset: An Autonomous Driving Dataset for Perception and Prediction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2020.

[43] "NuScenes." [Online]. Available: https://www.nuscenes.org/

[44] B. Wilson *et al.*, "Argoverse 2: Next Generation Datasets for Self-Driving Perception and Forecasting." [Online]. Available: https://arxiv.org/abs/2301.00493

[45] W. Zhan *et al.*, "INTERACTION Dataset: An INTERnational, Adversarial and Cooperative moTION Dataset in Interactive Driving Scenarios with Semantic Maps," *arXiv:1910.03088 [cs, eess]*, Sep. 2019.

[46] spacenet.ai, "spacenet.ai - Accelerating Geospatial Machine Learning." [Online]. Available: https://spacenet.ai/

[47] [Online]. Available: http://deepglobe.org/

[48] S. Yoo *et al.*, "End-to-End Lane Marker Detection via Row-wise Classification." [Online]. Available: https://arxiv.org/abs/2005.08630

[49] D. Jagula, “Satellite Imagery for Everyone.” [Online]. Available: https://spectrum.ieee.org/commercial-satellite-imagery#:~:text=The%20best%20commercially%20available%20spatial,Moderate

[50] E. W. Dijkstra, “A note on two problems in connexion with graphs,” *Numer. Math. (Heidelb.)*, vol. 1, no. 1, pp. 269–271, Dec. 1959.

[51] P. E. Hart, N. J. Nilsson, and B. Raphael, “A Formal Basis for the Heuristic Determination of Minimum Cost Paths,” *IEEE Transactions on Systems Science and Cybernetics*, vol. 4, no. 2, pp. 100–107, 1968, doi: 10.1109/TSSC.1968.300136.

[52] A. Stentz, “Optimal and efficient path planning for partially-known environments,” in *Proceedings of the 1994 IEEE International Conference on Robotics and Automation*, 1994, pp. 3310–3317. doi: 10.1109/ROBOT.1994.351061.

[53] A. (Tony) Stentz, “The Focussed D* Algorithm for Real-Time Replanning,” in *Proceedings of 14th International Joint Conference on Artificial Intelligence (IJCAI '95)*, Aug. 1995, pp. 1652–1659.

[54] S. Koenig and M. Likhachev, “Fast replanning for navigation in unknown terrain,” *IEEE Transactions on Robotics*, vol. 21, no. 3, pp. 354–363, 2005, doi: 10.1109/TRO.2004.838026.

[55] O. Khatib, “Real-Time Obstacle Avoidance for Manipulators and Mobile Robots,” in *Autonomous Robot Vehicles*, I. J. Cox and G. T. Wilfong, Eds., New York, NY: Springer New York, 1990, pp. 396–404. doi: 10.1007/978-1-4613-8997-2_29.

[56] Y. Koren and J. Borenstein, “Potential field methods and their inherent limitations for mobile robot navigation,” in *Proceedings. 1991 IEEE International Conference on Robotics and Automation*, 1991, pp. 1398–1404. doi: 10.1109/ROBOT.1991.131810.

[57] V. Gazi, “Swarm aggregations using artificial potentials and sliding-mode control,” *IEEE Transactions on Robotics*, vol. 21, no. 6, pp. 1208–1214, 2005, doi: 10.1109/TRO.2005.853487.

[58] N. Leonard and E. Fiorelli, “Virtual leaders, artificial potentials and coordinated control of groups,” in *Proceedings of the 40th IEEE Conference on Decision and Control (Cat. No.01CH37228)*, 2001, pp. 2968–2973. doi: 10.1109/CDC.2001.980728.

[59] S. M. LaValle, “Rapidly-exploring random trees : a new tool for path planning,” *The annual research report*, 1998, [Online]. Available: https://api.semanticscholar.org/CorpusID:14744621

[60] S. Karaman and E. Frazzoli, “Sampling-based Algorithms for Optimal Motion Planning.” [Online]. Available: https://arxiv.org/abs/1105.1186

[61] R. Cui, Y. Li, and W. Yan, “Mutual Information-Based Multi-AUV Path Planning for Scalar Field Sampling Using Multidimensional RRT*,” *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 46, no. 7, pp. 993–1004, 2016, doi: 10.1109/TSMC.2015.2500027.

[62] M. Xanthidis *et al.*, “Navigation in the Presence of Obstacles for an Agile Autonomous Underwater Vehicle,” in *2020 IEEE International Conference on Robotics and Automation (ICRA)*, 2020, pp. 892–899. doi: 10.1109/ICRA40945.2020.9197558.

[63] C. S. TAN, “A Collision Avoidance System for Autonomous Underwater Vehicles.” [Online]. Available: https://pearl.plymouth.ac.uk/secam-theses/302/

[64] C. Lamini, S. Benhlima, and A. Elbekri, “Genetic Algorithm Based Approach for Autonomous Mobile Robot Path Planning,” *Procedia Computer Science*, vol. 127, pp. 180–189, 2018, doi: https://doi.org/10.1016/j.procs.2018.01.113.

[65] L. Zadeh, “Fuzzy sets,” *Information and Control*, vol. 8, no. 3, pp. 338–353, 1965, doi: https://doi.org/10.1016/S0019-9958(65)90241-X.

[66] L. Zadeh, “Fuzzy algorithms,” *Information and Control*, vol. 12, no. 2, pp. 94–102, 1968, doi: https://doi.org/10.1016/S0019-9958(68)90211-8.

[67] “Use a Digital Signature.” [Online]. Available: https://developers.google.com/maps/documentation/maps-static/digital-signature

[68] C. H. Sudre, W. Li, T. Vercauteren, S. Ourselin, and M. Jorge Cardoso, “Generalised Dice Overlap as a Deep Learning Loss Function for Highly Unbalanced Segmentations,” in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, Springer International Publishing, 2017, pp. 240–248. doi: 10.1007/978-3-319-67558-9_28.

[69] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, “Focal Loss for Dense Object Detection.” [Online]. Available: https://arxiv.org/abs/1708.02002

[70] M. Deb, M. Deb, and A. R. Murty, “TopoNets: High performing vision and language models with brain-like topography,” in *The Thirteenth International Conference on Learning Representations*, 2025. [Online]. Available: https://openreview.net/forum?id=THqWPzL00e

[71] [Online]. Available: https://mathworld.wolfram.com/BettiNumber.html

[72] J. R. Clough, N. Byrne, I. Oksuz, V. A. Zimmer, J. A. Schnabel, and A. P. King, “A Topological Loss Function for Deep-Learning Based Image Segmentation Using Persistent Homology,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 12, pp. 8766–8778, Dec. 2022, doi: 10.1109/tpami.2020.3013679.

[73] H. Noh, T. You, J. Mun, and B. Han, “Regularizing Deep Neural Networks by Noise: Its Interpretation and Optimization.” 2017.

[74] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, “DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs.” [Online]. Available: https://arxiv.org/abs/1606.00915

[75] S.-H. Tsang, “Review: DeepLabv3+ — Atrous Separable Convolution (Semantic Segmentation).” [Online]. Available: https://sh-tsang.medium.com/review-deeplabv3-atrous-separable-convolution-semantic-segmentation-a625f6e83b90

[76] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, “Rethinking Atrous Convolution for Semantic Image Segmentation.” [Online]. Available: https://arxiv.org/abs/1706.05587

[77] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, “Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation.” [Online]. Available: https://arxiv.org/abs/1802.02611

[78] A. Howard *et al.*, “Searching for MobileNetV3.” [Online]. Available: https://arxiv.org/abs/1905.02244

[79] I. Loshchilov and F. Hutter, “Decoupled Weight Decay Regularization.” [Online]. Available: https://arxiv.org/abs/1711.05101

Appendix

A: Branch loss function tests

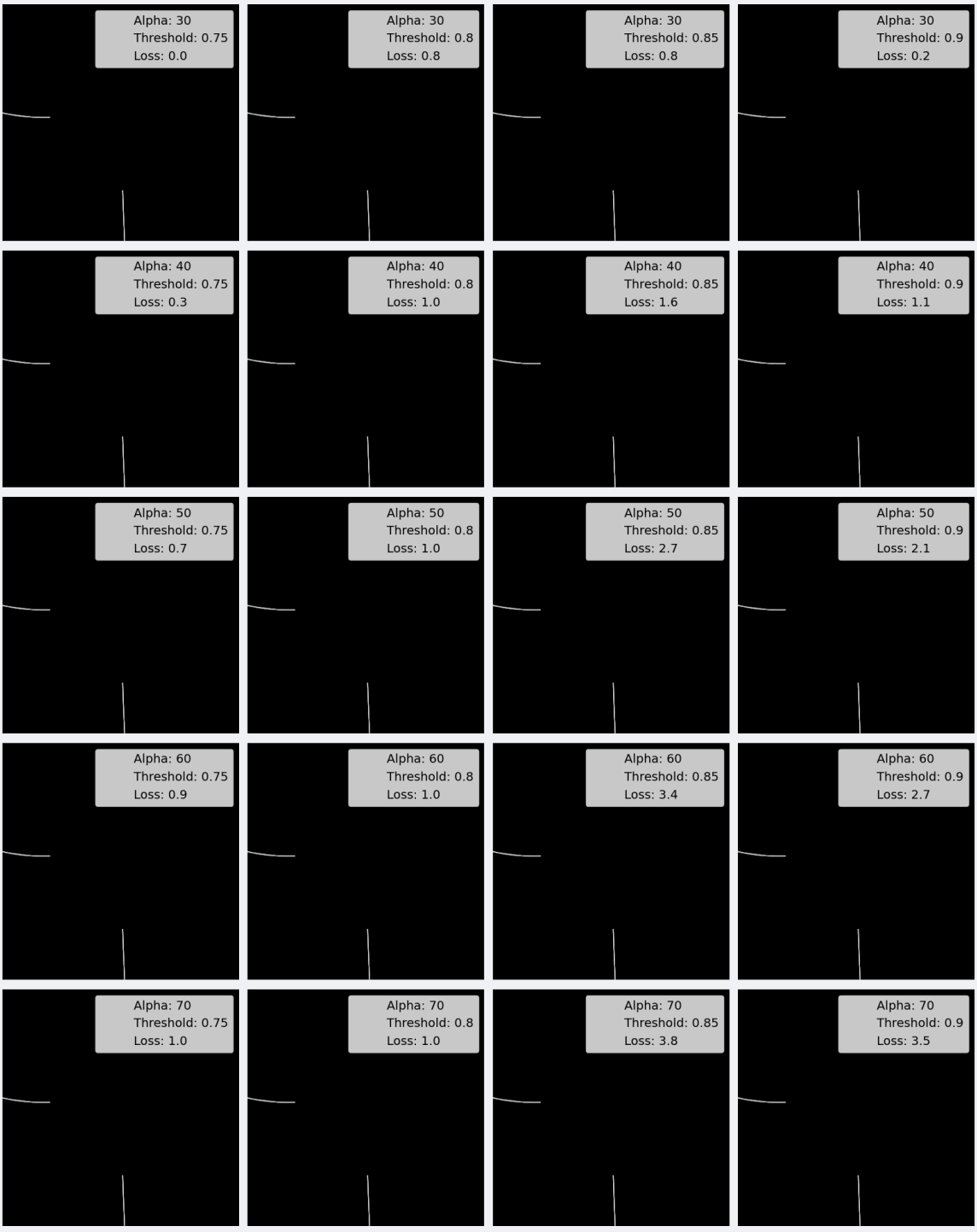


Figure 26. Showcase of various combinations of alpha and threshold values for the branch loss part of the topology loss function.