A MULTI-HEAD ATTENTION APPROACH WITH COMPLEMENTARY MULTIMODAL FUSION FOR VEHICLE DETECTION

by

Nujhat Tabassum

A Thesis

Submitted to the Faculty of Purdue University

In Partial Fulfillment of the Requirements for the degree of

Master of Science in Electrical and Computer Engineering

Department of Electrical and Computer Engineering

Indianapolis, Indiana

May 2024

THE PURDUE UNIVERSITY GRADUATE SCHOOL
STATEMENT OF COMMITTEE APPROVAL

Dr. Mohamed El-Sharkawy, Chair

Department of Electrical and Computer Engineering

Dr. Brian King

Department of Electrical and Computer Engineering

Dr. Maher Rizkalla

Department of Electrical and Computer Engineering

Approved by:

Dr. Brian King

Dedicated to

My husband- S M Abrar Jahin for his constant support and encouragement throughout my full journey

ACKNOWLEDGMENTS

I wish to extend my heartfelt thanks to my thesis advisor, Dr. Mohamed El-Sharkawy. His expert guidance, continual encouragement, and unwavering support have been indispensable. His methodical approach greatly aided my comprehension of the concepts and methodologies essential for conducting this research.

It is with great respect and appreciation that I acknowledge the contributions of Dr. Brian King, Chair of the ECE Department, and Dr. Maher Rizkalla, for their invaluable roles on my graduate committee and their significant input to my research. Additionally, I extend my profound gratitude to the esteemed faculty of the Department of Electrical and Computer Engineering at IUPUI and my special gratitude to Sherrie Tucker.

I am indebted to the Purdue School of Engineering & Technology for financially supporting the development of my thesis. Similarly, my gratitude extends to the IoT Collaboratory at IUPUI for welcoming me into their community. The IoT lab was a crucial resource, providing the hardware and software support necessary for this research.

Lastly, I must express my profound thanks to my husband for his unwavering support, motivation, and assistance in brainstorming and troubleshooting throughout this journey.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABBREVIATIONS

MVDNet	Multimodal Vehicle Detection Network
CNN	Convolutional Neural Networks
R-CNN	Region-Based Convolutional Neural Network
DEF	Deep Fusion
ANN	Artificial Neural Network
ViT	Vision Transformer
FMCW	Frequency Modulated Continuous Wave
ADAS	Advanced Driver-Assistance System
DL	Deep Learning
AI	Artificial Intelligence
VGG	Visual Graphics Group
ReLU	Rectified Linear Unit
RPN	Region Proposal Network
RFN	Region Fusion Network
NMS	Non-Maximum Suppression
RoI	Region of Interest
IoU	Intersection over Union
GPU	Graphics Processing Unit
LiDAR	Light Detection and Ranging
SGD	Stochastic Gradient Descent
ORR	Oxford Radar RobotCar
AP	Average Precision
AR	Average Recall
TP	True Positive
FP	False Positive
FN	False Negative
GPS	Global Positioning System
CUDA	Compute Unified Device Architecture

PNG	Portable Network Graphics
COCO	Common Objects in Context
NLP	Natural Language Processing

ABSTRACT

The advancement of autonomous vehicle technologies has taken a significant leap with the development of an improved version of the Multimodal Vehicle Detection Network (MVDNet), distinguished by the integration of a multi-head attention layer. This key enhancement significantly refines the network's capability to process and integrate multimodal sensor data, an aspect that becomes crucial in the face of challenging weather conditions. The effectiveness of this upgraded Multi-Head MVDNet is rigorously verified through an extensive dataset acquired from the Oxford Radar Robotcar, demonstrating its enhanced performance capabilities. Notably, in complex environmental conditions, the Multi-Head MVDNet shows a marked superiority in terms of Average Precision (AP) compared to existing models, underscoring its advanced detection capabilities.

The transition from the traditional MVDNet to the enhanced Multi-Head Vehicle Detection Network signifies a notable breakthrough in the arena of vehicle detection technologies, with a special emphasis on operation under severe meteorological conditions, such as the obscuring presence of dense fog or the complexities introduced by heavy snowfall. This significant enhancement capitalizes on the foundational principles of the original MVDNet, which skillfully amalgamates the individual strengths of lidar and radar sensors. This is achieved through an intricate and refined process of feature tensor fusion, creating a more robust and comprehensive sensory data interpretation framework. A major innovation introduced in this updated model is the implementation of a multi-head attention layer. This layer serves as a sophisticated replacement for the previously employed self-attention mechanism. Segmenting the attention mechanism into several distinct partitions enhances the network's efficiency and accuracy in processing and interpreting vast arrays of sensor data.

An exhaustive series of experimental analyses was undertaken to determine the optimal configuration of this multi-head attention mechanism. These experiments explored various combinations and settings, ultimately identifying a configuration consisting of seven distinct attention heads as the most effective. This setup was found to optimize the balance between computational efficiency and detection accuracy. When tested using the rich radar and lidar datasets from the ORR project, this advanced Multi-Head MVDNet configuration

consistently demonstrated its superiority. It not only surpassed the performance of the original MVDNet but also showed marked improvements over models that relied solely on lidar data or the DEF models, especially in terms of vehicular detection accuracy. This enhancement in the MVDNet model, with its focus on multi-head attention, not only represents a significant leap in the field of autonomous vehicle detection but also lays a foundation for future research. It opens new pathways for exploring various attention mechanisms and their potential applicability in scenarios requiring real-time vehicle detection. Furthermore, it accentuates the importance of sophisticated sensor fusion techniques as vital tools in overcoming the challenges posed by adverse environmental conditions, thus paving the way for more resilient and reliable autonomous vehicular technologies.

1. INTRODUCTION

1.1 Overview

The technological landscape of autonomous vehicles has experienced rapid advancements, particularly in the quest for achieving Level Five automation - the pinnacle of autonomous driving where no human intervention is required under any driving conditions [1]. A critical component of this ambition hinges on the development of robust all-weather object detection systems. These systems are indispensable for the accurate and reliable identification and localization of objects in various weather conditions, including challenging scenarios like fog, rain, or snow.

Autonomous vehicles today are typically equipped with an array of sophisticated sensors [2][3][4], including radar, LiDAR (Light Detection and Ranging), and cameras. These sensors, through their integrated functionalities, collectively enhance the vehicle's ability to detect and interpret its surroundings. The fusion of these diverse sensory inputs is particularly vital in overcoming the inherent limitations of each sensor type. However, most existing object detection methodologies, which primarily integrate LiDAR and camera data [5][6][7], are heavily reliant on visibility. This dependence becomes a significant challenge in adverse weather conditions, where the effectiveness of visual sensors is considerably reduced [8].

Radar technology, known for its efficacy in navigating through foggy conditions, emerges as a crucial element in the sensor suite of autonomous vehicles [2][3]. The advantage of radar in such scenarios is attributed to its use of millimeter-wave signals, which can penetrate or circumvent obstacles like fog particles more effectively than other wavelengths [9]. Despite this, the exploration of radar data in autonomous driving applications has been somewhat limited due to the scattered nature of radar data. A notable advancement in this field is the Oxford Radar Robotcar (ORR) dataset [10], which introduced an innovative approach by incorporating a rotating horn antenna radar system. This system is designed to provide a comprehensive 360° view of the vehicle's surroundings with an impressive azimuth resolution of 0.9°, marking a significant step forward in the use of radar data for autonomous vehicles.

Building on these developments, this thesis introduces a multimodal deep fusion model that leverages the multi-head attention mechanism to enhance vehicle detection in foggy

Figure 1.1. Comparison of lidar-only and image only data with their fusion[2].

conditions. The multi-head attention mechanism, a concept derived from advances in visual transformers (ViT) [11], is adept at processing complex datasets by dividing the attention process into multiple segments or 'heads' [12]. Each head is allowed for a parallel and more nuanced analysis. This approach enables the extraction of a richer spectrum of information from lidar and radar data, surpassing the capabilities of single attention mechanisms. The integration of the multi-head attention mechanism into MVDNet is a pioneering step in the field of autonomous vehicle technology. It represents a significant enhancement in the model's ability to interpret and analyze sensor data, particularly in challenging weather conditions, thereby pushing the boundaries of what is possible in the pursuit of fully autonomous driving.

1.2 Motivation

The motivation behind this research lies in the pressing need to advance the efficacy of autonomous vehicle detection systems, especially under challenging foggy weather conditions. A notable limitation in current sensor fusion methodologies, primarily those combining lidar and camera data, is their heavy reliance on visibility, which becomes critically compromised in adverse weather. This research is motivated by the opportunity to bridge this gap through the innovative application of multi-head attention mechanisms to the late fusion of lidar and radar data [13].

The concept of multi-head attention mechanisms, which has garnered significant success in the realm of image data processing [14], offers a promising avenue for processing and interpreting the complex data sets typical of autonomous vehicle sensors. The application of this mechanism to lidar and radar data fusion in the MVDNet model is particularly compelling. It represents a shift from traditional approaches, moving towards a more sophisticated, nuanced analysis of sensor data.

Multi-head attention mechanisms enable simultaneous focus on different attributes or aspects of the data, such as the spatial positioning of elements or the varying intensities of signals. This parallel processing capability is critical for ensuring that the MVDNet model can comprehensively analyze the sensor data, identifying and focusing on the most relevant features for object detection. The adaptation of this mechanism from visual transformers to the fusion of lidar and radar data is not just a technical enhancement but a strategic move to leverage the distinct characteristics of each sensor modality effectively.

The motivation for this research also stems from the potential to significantly improve the safety and reliability of autonomous vehicles. By enhancing the vehicle detection capabilities of the MVDNet model, particularly in foggy conditions, this research aims to contribute to the development of autonomous driving technologies that are robust and dependable, even in the most challenging environments. This advancement is not just a technical achievement but a crucial step towards realizing the full potential of autonomous vehicles, ensuring their safe integration into everyday life.

1.3 Research Objectives

The primary objectives of this research are meticulously outlined to address key areas in the advancement of autonomous vehicle technologies.

Implementation of Multi-Head Attention in MVDNet: Central to this research is the objective to augment the MVDNet model's efficacy in detecting vehicles under foggy weather conditions. The introduction of a novel multi-head attention mechanism into the model is anticipated to significantly enhance its ability to process and analyze lidar and radar data. This enhancement is expected to result in a more accurate and reliable detection of vehicles, considering the diverse and complex nature of sensor data in such challenging environments.

Optimized Fusion of lidar and Radar Data: A key goal is to explore and maximize the potential of combining lidar and radar data using the multi-head attention approach. This objective seeks to extend the capabilities of sensor data utilization in autonomous vehicles, going beyond traditional methods of data fusion. By doing so, the research aims to uncover new insights and methodologies that could lead to more sophisticated and effective sensor data integration.

Advanced Data Analysis Techniques: This research intends to implement a comprehensive data analysis mechanism within the existing MVDNet model. By dynamically adjusting focus areas based on attention scores for various data segments, the model is expected to conduct a more thorough and detailed object detection process. This approach will leverage the multi-head attention mechanism's ability to parallel-process different features of the data, thereby enhancing the overall performance of the model.

Enhanced Reliability and Safety in Adverse Conditions: Ultimately, this research aims to contribute significantly to the safety and reliability of autonomous vehicles, especially in challenging weather conditions. By improving vehicle detection capabilities, the research aspires to address one of the critical challenges in the field of autonomous driving. The goal is to ensure that autonomous vehicles can operate safely and efficiently, irrespective of environmental conditions, thereby advancing the field towards the realization of fully autonomous driving.

Through these objectives, the research endeavors to make a substantial contribution to the field of autonomous vehicle technologies, focusing on enhancing the capabilities and reliability of autonomous vehicles in navigating complex and unpredictable environments.

1.4 Challenges

In the course of developing enhancements to the MVDNet model for autonomous vehicle technology, several significant challenges were encountered.

One of the initial and crucial challenges was setting up a compatible environment for the existing MVDNet model. This process involved identifying and installing the appropriate software and their specific version numbers. The significance of version compatibility cannot be overstated; different versions of software, libraries, and tools can have varying functionalities and compatibility issues. Ensuring that all components of the environment were harmonized in terms of versioning was essential to maintain the stability and functionality of the model. This meticulous setup was necessary to prevent potential conflicts and dependency issues that could arise from version mismatches.

To improve the performance of the MVDNet model, extensive modifications were required. This task presented a substantial challenge, necessitating a comprehensive understanding of the existing codebase and its operational mechanisms. This deep dive into the code not only involved understanding the structure and flow of the existing algorithms but also required the identification of potential areas for optimization and enhancement.

A pivotal aspect of this research was the integration of a multi-head attention layer into the fusion network unit of the MVDNet, replacing the existing self-attention layer. This implementation was challenging due to the complexity of the multi-head attention mechanism and the need to seamlessly integrate it into the existing architecture without disrupting the core functionalities. The multi-head attention layer needed to be carefully designed and optimized to work in tandem with the other components of the model, ensuring that the enhanced model could effectively process and analyze the lidar and radar data.

A critical technical challenge was determining the optimal number of heads for the multi-head attention mechanism. The chosen number of heads had to allow for the equal distribu-

tion of the input data (feature tensors of lidar and radar signal) among the heads, which is crucial for the efficient and accurate processing of the data. An incorrect number of heads, not aligning with the input data dimensions, could lead to issues such as data redundancy or incomplete data analysis. This decision required careful consideration and testing to ensure that the multi-head attention mechanism operated optimally within the context of the MVDNet model.

Each of these challenges represented a significant hurdle in the enhancement of the network. Overcoming them required a combination of technical expertise, careful planning, and rigorous testing to ensure that the modifications and enhancements made to the model would yield the desired improvements in performance and reliability.

1.5 Contributions

- Proposed the integration of a novel multi-head attention layer within the MVDNet model for improved vehicle detection in adverse weather conditions.
- Proposed an empirical optimization of the number of attention heads in the network, ensuring efficient data processing and alignment with input dimensions of the sensors.
- Research paper titled Vehicle Detection in Adverse Weather: A Multi-head Attention Approach with Multimodal Fusion pending submission.

2. LITERATURE REVIEW

2.1 Sensors for Autonomous Vehicles

The automotive industry has seen a remarkable evolution with the integration of advanced sensors like cameras, radars, and LiDAR, each playing a vital role in the development of autonomous and assisted driving technologies. Camera sensors provide critical visual data for functions such as traffic monitoring and lane detection, although their effectiveness can be hindered by poor lighting or adverse weather conditions. Radar sensors, on the other hand, excel in these challenging environments by using radio waves to detect objects' distance and speed, making them indispensable for collision avoidance and adaptive cruise control. LiDAR sensors complement these by using laser technology to create detailed three-dimensional maps of a vehicle's surroundings, crucial for precise navigation in autonomous driving, despite challenges such as high costs and sensitivity to certain weather conditions[15]. The synergy of these sensors enhances vehicle safety and functionality, marking a significant stride in automotive technology toward more sophisticated and reliable transportation solutions [16].

2.2 LiDAR

LiDAR technology operates by emitting laser pulses and measuring the time taken for these pulses to return after reflecting off a target. This process, although conceptually straightforward, becomes intricate due to the requirement for precise time resolution and the need to maintain a minimal signal-to-noise ratio. LiDAR systems, diverse in their technological makeup, cater to various sectors including automotive, military, robotics, surveillance, and topographical mapping through airborne laser scanners. Each application demands unique specifications; for instance, the automotive industry prioritizes cost-effective, rapid, and high-resolution scanners. Leading manufacturers in this domain include Velodyne, Ibeo, and Valeo, predominantly utilizing mechanical beam steering, optical laser diodes for pulse emission, and avalanche photodiodes for detection [8]. However, recent trends suggest a

shift towards solid-state technology, promising enhanced durability and maintenance for these scanners.

In the context of the Oxford Radar Robotcar dataset, the Velodyne HDL-32E 3D LiDAR has been employed. This sensor provides a comprehensive scanning capability with a 360° horizontal and 40° vertical field of view [10]. It's adept at generating as many as 700,000 data points per second, facilitating a thorough and accurate three-dimensional depiction of its surroundings. While its effective range reaches up to 100 meters, this can fluctuate depending on factors like the reflectivity of objects and varying environmental conditions. The inclusion of 32 channels in the sensor design significantly enhances its ability to capture high-resolution data.

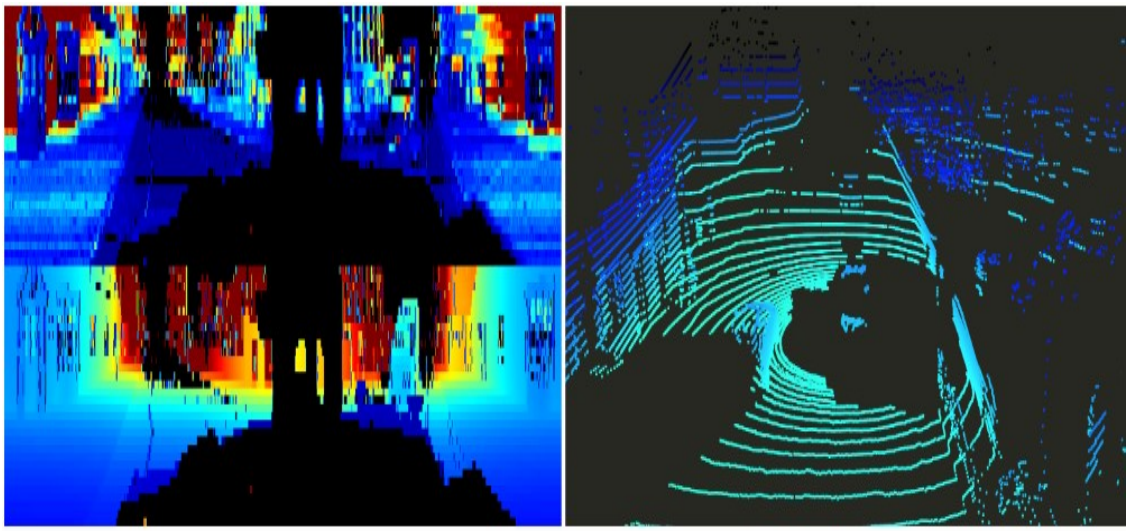

Figure 2.1. Velodyne lidar scan in Pointcloud form (right side) and Velodyne lidar scan in Polar form (left side) [10].

LiDAR technologies boast a higher resolution than their radar counterparts and demonstrate greater resilience to weather-related disruptions compared to camera systems. This attribute renders them exceptionally proficient in generating precise environmental representations for vehicles [17] [18]. Nonetheless, LiDAR systems are accompanied by their own set of challenges. They necessitate advanced computational power and intricate data processing capabilities, and their operational efficacy can be influenced by various environmental elements [19]. This combination of sophisticated functionality and complex technical require-

ments highlights the critical importance of LiDAR in the progression of vehicle technology and the accurate sensing of environmental conditions [20].

2.3 Radar

The utilization of radar sensors is on the rise, particularly in systems designed to aid autonomous driving, such as Adaptive Driver-Assistance Systems (ADAS). These systems frequently leverage radar for functions like automatic braking systems and adaptive cruise control. Radar sensors are increasingly recognized for their ability to enhance camera-based systems, contributing to capabilities such as collision avoidance and the identification of pedestrians and cyclists. Among the various types of radar technologies employed in the automotive industry, the Frequency Modulated Continuous Wave (FMCW) radar is notably prevalent. This popularity stems from several advantages, notably its cost-effective hardware and the reduced computational resources required for signal processing [21].

The fundamental principle of radar technology hinges on the transmission and reception of radio waves, operating similarly to the way an echo works. Analogous to how a human voice echoes in a cavern, radar emits frequency waves that reflect off nearby objects. This echoed signal, once returned to the radar system, is analyzed to determine both the distance and the orientation of the objects in question. This echo-based mechanism enables the radar to accurately assess the positioning and movement of nearby entities, forming an integral component of modern vehicular safety and navigation systems [22].

The Navtech CTS350-X used in the ORR dataset represents a Frequency Modulated Continuous Wave (FMCW) scanning radar, distinguished by its absence of Doppler capabilities. It is adept at generating 3768 power readings across 400 azimuths with a fine resolution of 4.38 cm, operating at a frequency of 4 Hz. This setup allows for a maximum detection range of 163 meters and an azimuth resolution of 0.9°. While there are variants of the Navtech CTS350-X that can detect objects at distances exceeding 650 meters and at higher rotational speeds, the chosen configuration is particularly optimized for urban environments, where extended straight-line distances beyond 163 meters are uncommon. Operating within the frequency range of 76 GHz to 77 GHz, the radar ensures reliable performance even in ad-

verse conditions like dust, rain, and snow. It features a main beam spread of 1.8° horizontally and vertically between the -3 dB points and includes a secondary cosecant squared beam pattern extending up to 40° below the horizontal, enhancing its ability to detect objects located below the primary beam [10].

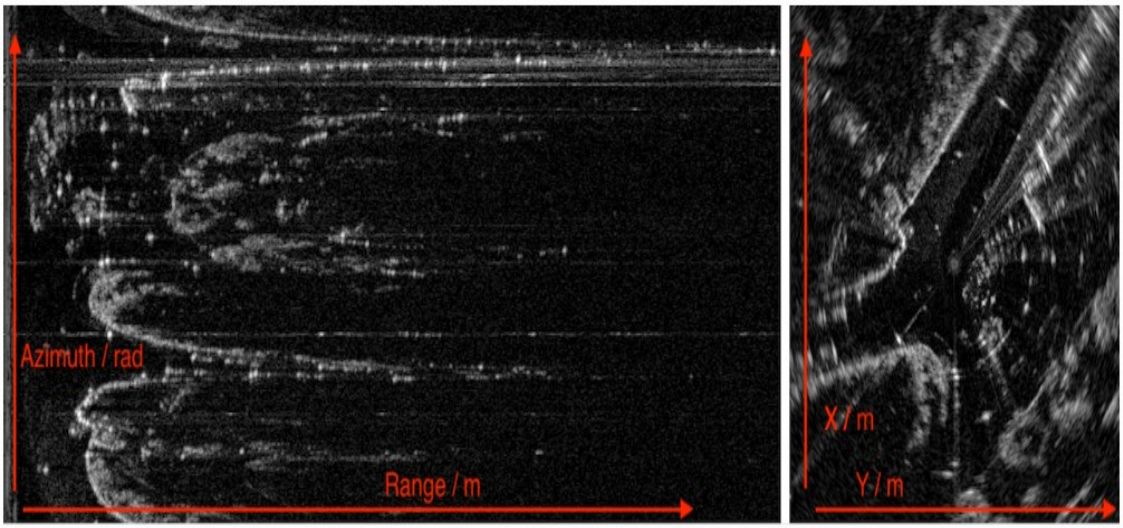

Figure 2.2. Radar scan in Polar form (left side) and after conversion in Cartesian form (right side) [10].

A notable strength of this radar system lies in its consistency and reliability across various weather conditions. However, it's not without limitations. One such challenge is the occurrence of noise in the point cloud data, which can occasionally result in the detection of ghost objects or false positives [23]. Additionally, the nature of radar data limits its utility in object classification, as discerning specific shapes and details from radar signals alone can be challenging. This balance of advantages, such as robustness in diverse conditions, against certain technical constraints, highlights the complex yet critical role of the Navtech CTS350-X radar in urban navigation and safety applications.

2.4 Multimodal Fusion

In the rapidly evolving field of automotive technology, multimodal sensor fusion stands as a key innovation. This technology integrates data from diverse sensors such as cameras, LiDAR, radar, and ultrasonics to enhance the vehicle's perception and decision-making ca-

pabilities. Multimodal sensor fusion involves combining and processing data from various sensors to create a unified and accurate representation of the vehicle's surroundings. For instance, cameras provide detailed visual information, LiDAR offers precise distance measurements, and radar is effective in adverse weather. By fusing these data streams, vehicles gain a comprehensive understanding of their environment [24].

The primary advantage of sensor fusion lies in its ability to provide a more reliable and robust system. It mitigates the limitations of individual sensors and ensures continuous functionality, even if one sensor type becomes compromised. This redundancy is vital for critical applications like autonomous driving, where safety and accuracy are paramount. In Advanced Driver-Assistance Systems (ADAS), sensor fusion plays a crucial role. It enhances features such as adaptive cruise control, collision avoidance, and lane-keeping assist, making driving safer and more efficient. The fused data allows for more accurate predictions and responses to dynamic road conditions [25].

Despite its benefits, sensor fusion poses challenges like the need for high computational resources to process the diverse data inputs in real-time. Ensuring synchronization and calibration among different sensor types is also crucial to maintain accuracy. Additionally, the complexity of integrating these systems can impact the vehicle's design and cost.

2.5 Deep Learning and Neural Networks

In the arena of Artificial Intelligence (AI), neural networks and deep learning have emerged as groundbreaking technologies, reshaping how machines understand and process information. Neural networks, inspired by the human brain's structure, are intricate systems of algorithms that identify patterns and process data in a manner akin to human cognition. Comprising input, hidden, and output layers, these networks simulate the function of biological neurons, creating a structure that processes incoming data, applies various transformations through hidden layers, and outputs the processed information [26].

Deep Learning, an advanced iteration of neural networks, involves more complex networks known as deep neural networks due to their numerous layers. These networks excel at handling and analyzing large volumes of unstructured data. Deep learning architectures like

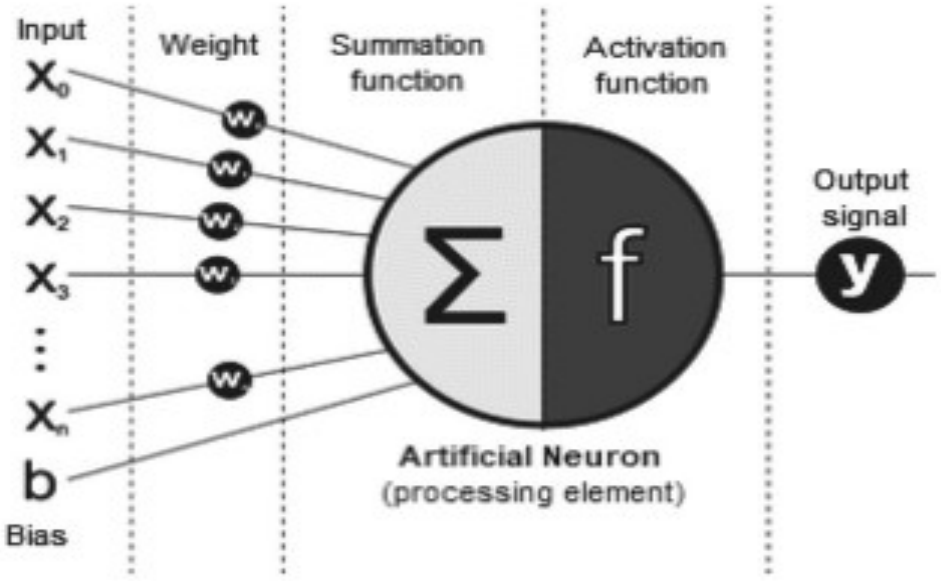

Figure 2.3. An artificial neuron with basic processing elements [26].

Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) have been instrumental in areas such as image and speech recognition, as well as language processing. The efficacy of neural networks is honed through a training process where the network adjusts its internal parameters (weights) to minimize the error in its outputs. Techniques like backpropagation, coupled with optimization algorithms such as gradient descent, are critical in refining these networks' performance [27].

2.5.1 Convolutional Neural Network

In the realm of deep learning, Convolutional Neural Networks (CNNs) emerge as a pivotal architecture, primarily tailored for analyzing data that inherently exhibits a grid-like topology. This architectural design finds its most prevalent application in image processing, attributed to the grid-structured nature of images. A defining characteristic of CNNs, distinguishing them from conventional neural networks, is their composition of three distinct types of layers: the convolution layer, the pooling layer, and the fully connected layers. The first two layers are primarily dedicated to the extraction of features, whereas the fully connected

layer, which may consist of multiple sub-layers, is instrumental in correlating these extracted features with the final output values. It is this innovative approach toward feature extraction that sets CNNs apart, rendering them exceptionally proficient in detecting patterns within images for object recognition.

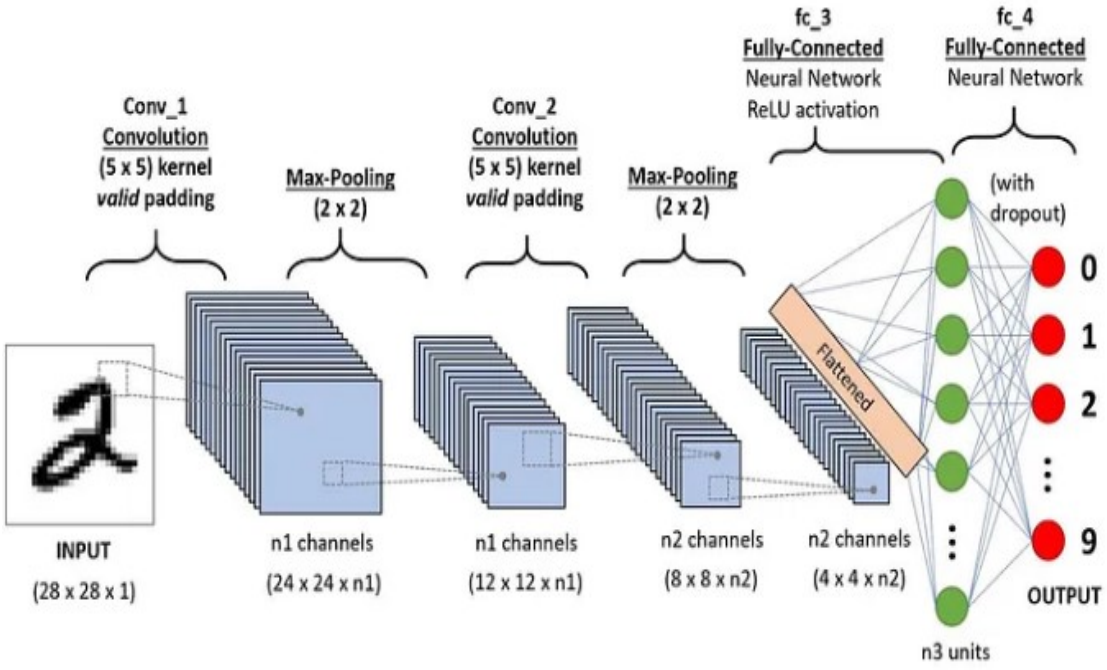

Figure 2.4. Basic architecture of convolutional neural networks (CNNs)[28].

The inception of any neural network is marked by the input layer. For CNNs, this layer is predominantly represented by an image. However, CNNs are versatile enough to accommodate various non-image data forms, including signal, audio, or time-series, provided they are pre-processed to align with the architecture's expected input specifications.

Following the input layer is the convolutional layer, which performs linear operations crucial for the extraction of features. A key element of this layer is the use of a kernel, a diminutive 2D matrix, typically odd-dimensional, which is tasked with discerning the relationships between a central pixel and its adjacent pixels in a given area. This kernel traverses across the expanse of the larger input image, altering the central pixel of the input image in the process.

The output dimension of the convolutional layer is contingent upon the stride number, which determines the kernel's movement, and the padding method, which entails the addition of zeros to facilitate convolution at each pixel location. The endpoint of this layer is characterized by the input undergoing convolution with an array of kernels, thereby generating a feature map.

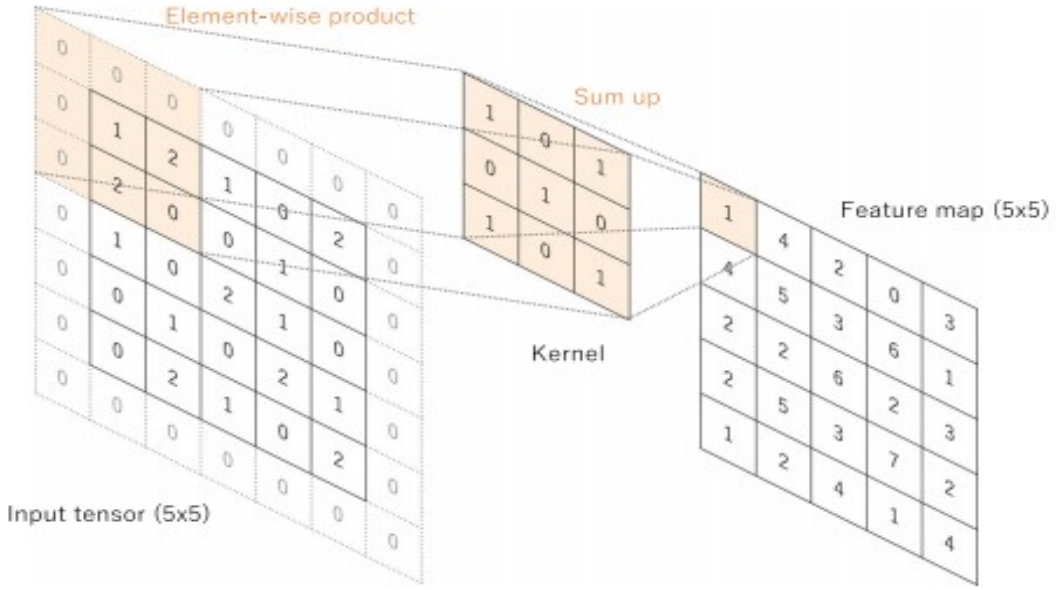

Figure 2.5. Convolution operation with zero padding, 3×3 kernel and stride 1 [29].

In the context of Convolutional Neural Networks (CNNs), pooling is a strategic process that enhances the generalization of features derived from convolutional filters. This technique plays a pivotal role in enabling the network to identify features regardless of their spatial positioning within the image. By integrating pooling, CNNs are equipped to maintain a robust recognition capacity even when the features undergo spatial variations. This aspect of pooling contributes significantly to the versatility and efficacy of CNNs in image analysis and recognition tasks.

This feature map is then subjected to a non-linear activation process, most commonly through a ReLU function. This step is crucial to ensure that the feature map is devoid of any negative values. The cycle of convolution, pooling, and activation is repeated in succes-

sion until the result is a compact matrix, within which the activation becomes increasingly responsive to complex features.

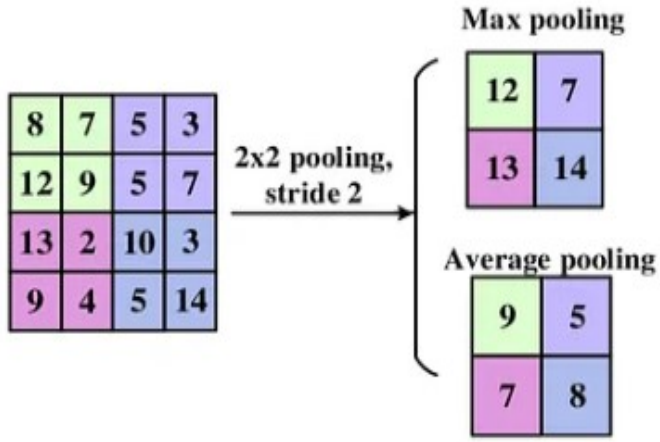

Figure 2.6. Max pooling and Average pooling operation with pool size 2×2.

The network's architecture culminates in the fully connected layer. This layer receives the compact matrix, a product of the previous layers, and transforms it into a one-dimensional array. In this layer, each input connects to every output through a specific weight, a feature inherent to its design. The number of neurons in the output layer is reflective of the classification categories required.

In the terminal layer, the softmax function is typically employed as the activation function, especially in scenarios involving multi-class classification. A few activation functions are also shown in Figure- 2.7. This function plays a crucial role in converting the output values from the fully connected layer into a probabilistic distribution, ranging from 0 to 1, indicative of the respective classification probabilities. Additionally, the use of other common activation functions is also delineated for further reference.

2.5.2 VGG-16 (Visual Graphics Group) Network

VGG-16, a seminal model in the convolutional neural network (CNN) landscape, was developed by the Visual Graphics Group at Oxford University. It garnered significant attention

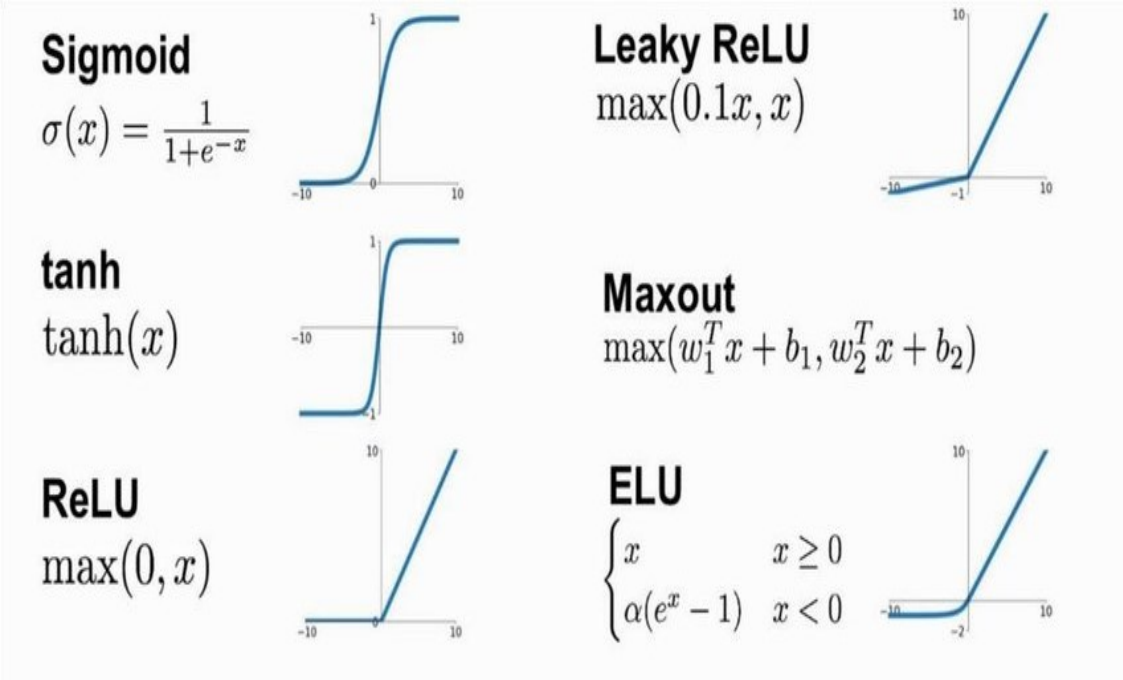

Figure 2.7. Example of the commonly used activation functions [30].

for its exceptional performance in the ImageNet Large Scale Visual Recognition Challenge. This 16-layer deep network is specifically tailored for image recognition, offering a benchmark in CNN architectures.

The model's architecture is noted for its straightforward design, comprising a sequential arrangement of 13 convolutional layers, interspersed with max-pooling layers, and concluding with 3 fully connected layers. The convolutional layers utilize compact 3x3 receptive fields, advancing with a one-pixel stride, which facilitates the extraction of intricate details from images while maintaining a manageable number of parameters. Following each convolutional layer, the ReLU activation function is employed, introducing necessary non-linearity that allows the network to capture complex image features [31].

A distinctive feature of VGG-16 is its integration of max-pooling layers following specific convolutional layers. These layers execute spatial down-sampling, effectively compressing the feature maps, thereby reducing computational demands and enhancing the network's ability to detect dominant and positionally invariant features.

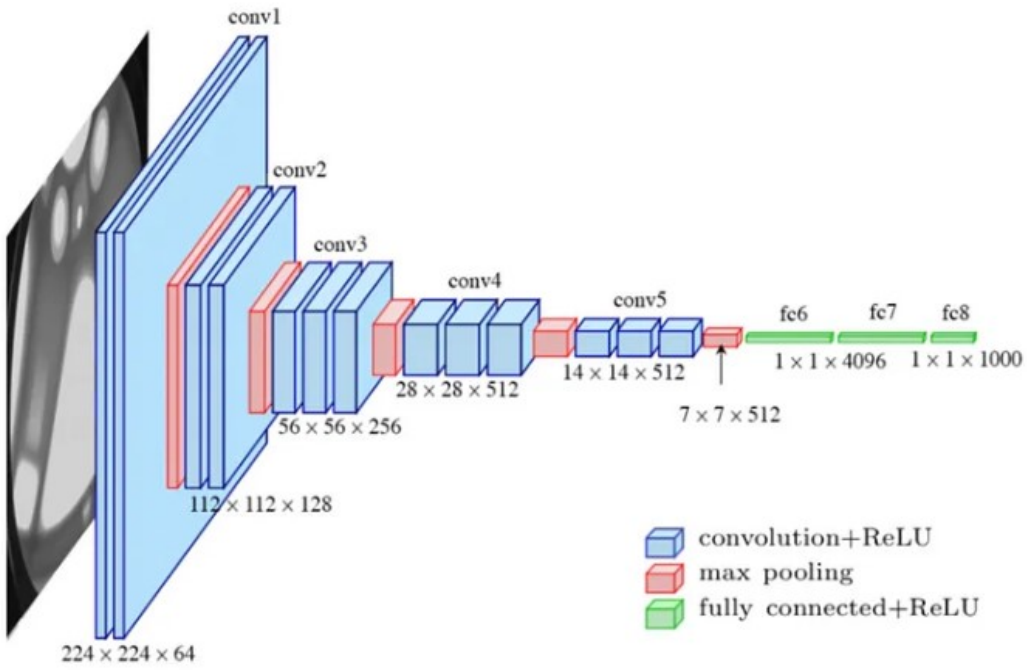

Figure 2.8. Standard architecture of a VGG-16 network.

At the network's culmination, the fully connected layers serve to synthesize the features learned for classification purposes. The initial two fully connected layers consist of 4096 nodes each, and the final layer's size is determined by the number of target classes in the classification task, typically utilizing a softmax activation function for probabilistic class representation.

To mitigate overfitting, VGG-16 incorporates dropout layers within its fully connected segments. This regularization strategy randomly deactivates certain neurons during training, promoting model generalization.

The training methodology for VGG-16 is also noteworthy. The model undergoes training on extensive datasets like ImageNet, which boasts millions of images across a diverse range of categories. Employing mini-batch gradient descent coupled with momentum and weight decay, the model's learning rate is meticulously adjusted during training to optimize its performance.

In essence, VGG-16 stands as a pivotal deep learning model in computer vision, renowned for its depth and uniformity in architecture. It excels in extracting detailed features from images, influencing subsequent model designs in the field, and maintaining its status as a fundamental reference in image classification.

2.5.3 Faster R-CNN

Faster Region-based Convolutional Neural Network (Faster R-CNN) is a state-of-the-art deep learning framework, primarily engineered for object detection tasks. This model marks a significant progression in the R-CNN series, offering marked improvements in speed and accuracy over its predecessors, R-CNN and Fast R-CNN.

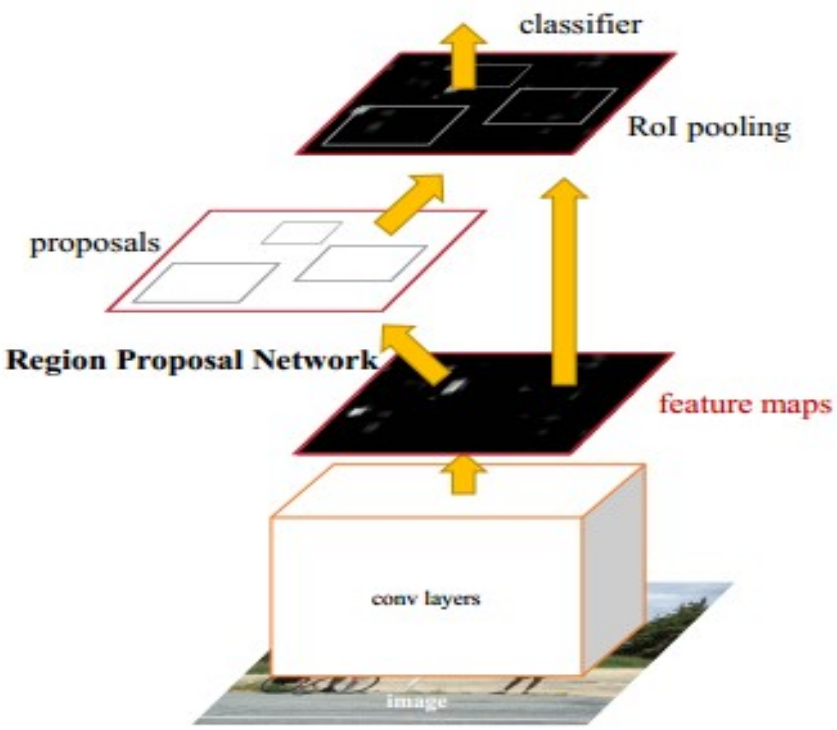

Figure 2.9. Faster R-CNN network for object detection [32].

At its core, the architecture of Faster R-CNN is bifurcated into two integral segments: the Region Proposal Network (RPN) and the Fast R-CNN detector. This dual-component struc-

ture is pivotal, enabling concurrent region proposal and object detection, thereby enhancing the model's processing speed significantly.

The RPN stands as a seminal innovation within Faster R-CNN. Operating as a fully convolutional network, the RPN generates predictions on object boundaries and objectness scores at various image locations. It employs a sliding window mechanism across the image, producing a set of object proposals, each accompanied by an objectness score. These scores signify the likelihood of object presence within those regions. The RPN is meticulously trained to produce high-quality region proposals, which are subsequently relayed to the Fast R-CNN detector for the object detection phase.

The Fast R-CNN detector, upon receiving these proposals, extracts features using a Region of Interest (RoI) pooling layer. This layer is designed to transform each proposal into a uniform-sized feature vector, suitable for classification and bounding box regression. A distinctive feature of Faster R-CNN is the extraction of these features from the final shared convolutional layer's feature map, thus reducing computational demands and enhancing overall model efficiency.

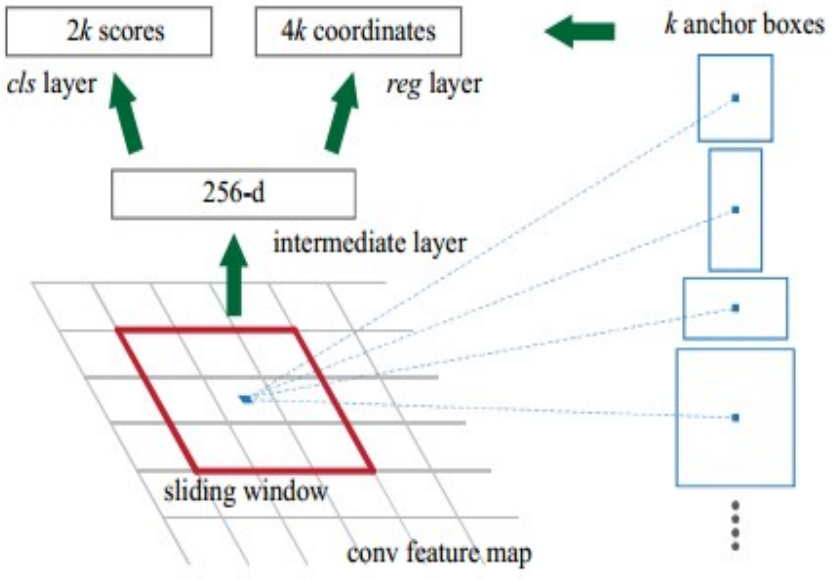

Figure 2.10. Region Proposal Network (RPN) operation [32].

Faster R-CNN is also notable for its innovative fusion of region proposal and object detection tasks. By sharing convolutional features between the RPN and the Fast R-CNN detector, the model achieves remarkable efficiency. This means the intensive process of generating feature maps is executed just once, and these maps are then utilized for both region proposal generation and object detection.

For training, Faster R-CNN employs a composite multi-task loss function that concurrently trains the RPN and the Fast R-CNN detector. This integrated training regimen is designed to optimize the network for both tasks, enhancing its overall detection performance.

In the realm of object detection, Faster R-CNN has set new benchmarks in terms of speed and precision. Its applications span various fields, including security surveillance, autonomous vehicles, and advanced image analysis. The model's proficiency in accurately detecting and localizing multiple objects in images has solidified its position as a preferred choice among computer vision researchers and practitioners.

3. BASELINE ARCHITECTURE

MVDNet, a sophisticated multimodal deep fusion model, has been crafted to tackle the complexities of vehicle detection in foggy conditions, a vital component for the progression of autonomous driving systems. The architecture of MVDNet is meticulously organized into two critical phases, each contributing significantly to the enhancement of vehicle detection's accuracy and dependability. The initial phase is focused on the independent generation of preliminary suggestions from LiDAR and radar data. This separation is crucial for maximizing the distinct attributes of each sensor type. Subsequently, MVDNet progresses to a more complex stage involving a well-thought-out fusion strategy. This strategy aims to amalgamate the features discerned from both LiDAR and radar, applying a temporal analysis via 3D convolutional methods. The model's late fusion technique, pivotal in its design, prioritizes processing in essential areas, termed regions of interest (RoI). This focus not only heightens the efficiency in generating proposals but also augments the precision in merging sensor data.

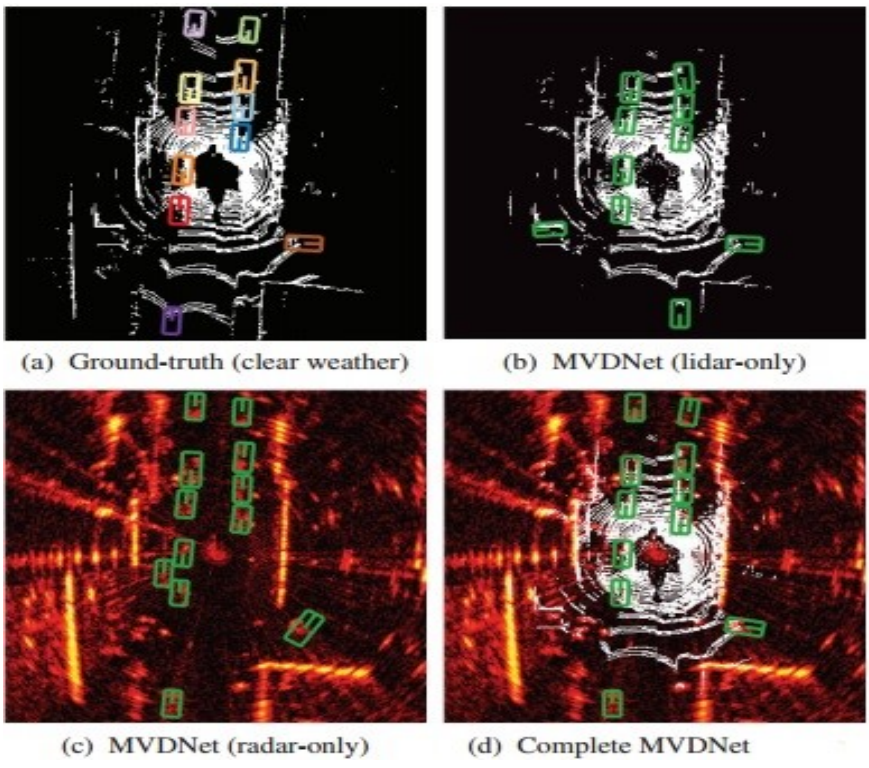

Figure 3.1. Illustration of performance comparison of the baseline MVDNet [13].

Despite offering a more extensive dataset, ORR radar data remains considerably less precise and more susceptible to noise when compared to its visual sensor, lidar, as depicted in Figure- 3.1 (a) and (c). Consequently, treating radar data in the same manner as lidar point clouds can lead to inaccuracies such as false alarms and substantial regression errors. To effectively identify vehicles in foggy weather conditions, it is essential to harness the strengths of both lidar (which provides detailed imagery within its range of visibility) and radar (known for its resilience to fog) while simultaneously mitigating their respective limitations. This unique contribution forms the core innovation of MVDNet.

The performance of MVDNet is comprehensively illustrated in the provided Figure- 3.1. Figure- 3.1(a) presents a 360° birds eye view of the 3D lidar point cloud, along with precisely labeled ground-truth information for different vehicles. In this context, the vehicle outfitted with both lidar and radar sensors is centrally positioned. In foggy weather scenarios, the lidar-only version of MVDNet (as shown in Figure- 3.1(b)) tends to overlook vehicles at greater distances due to the obscuring effects of fog and may mistakenly categorize non-vehicular elements as vehicles. On the other hand, the radar-only variant (Figure- 3.1(c)) is prone to generating false positives and imprecise bounding boxes, primarily due to the inherent noise in radar data. The complete potential of MVDNet is realized through the deep fusion of lidar and radar data, as depicted in Figure- 3.1(d). This integration empowers MVDNet to accurately detect vehicles, effectively overcoming the individual limitations of each sensing modality. By leveraging the detailed imaging capabilities of lidar and the fog-penetrating capabilities of radar, MVDNet offers a robust and reliable solution for vehicle detection in challenging weather conditions.

The MVDNet model is architecturally segmented into two primary phases. The first phase involves the Region Proposal Network (RPN), which handles the assimilation of inputs from LiDAR and radar sensors. This is achieved by deriving feature maps from the sensor inputs and then formulating proposals founded on these maps. In the subsequent phase, known as the Region Fusion Network (RFN), the focus shifts to the amalgamation and integration of region-centric features gathered from the data of each sensor.

3.1 Region Proposal Network (RPN)

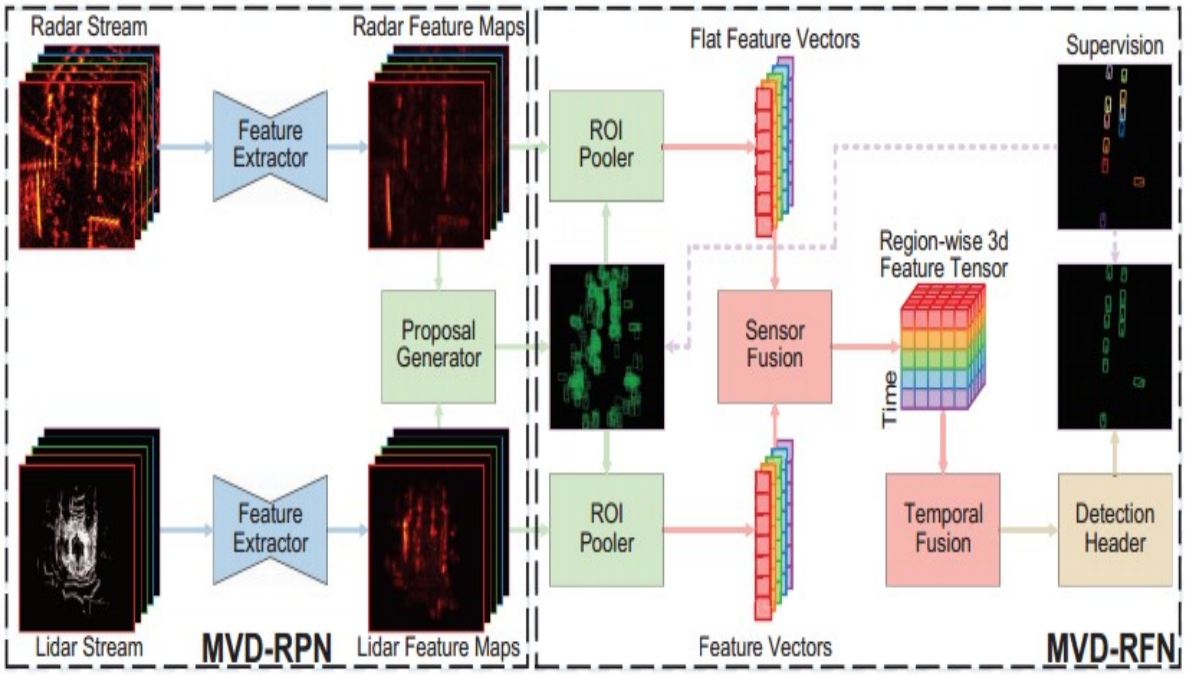

Figure 3.2. Architecture of the baseline MVDNet model [13].

3.1.1 Feature Extraction Unit

In the MVDNet architecture, as detailed in Figure- 3.2, the system incorporates two distinct feature extractor modules for LiDAR and radar inputs, both sharing a uniform structural design. Notably, the LiDAR section is equipped with a significantly larger number of feature channels, a choice influenced by the denser channel array found in LiDAR data compared to radar signals. The feature extraction process is composed of four convolutional layers, each with a 3x3 kernel, aimed at initially extracting features at a resolution parallel to the inputs. Following this, the architecture involves a downsampling stage using max-pooling. Importantly, the model integrates a transposed convolution layer for the purpose of upscaling the feature maps. The resulting upscaled outputs are then seamlessly integrated with the earlier high-resolution feature maps through a skip connection. This technique

ensures a comprehensive and unified set of feature maps for both LiDAR and radar inputs, with the skip connection playing a critical role in combining multi-resolution features.

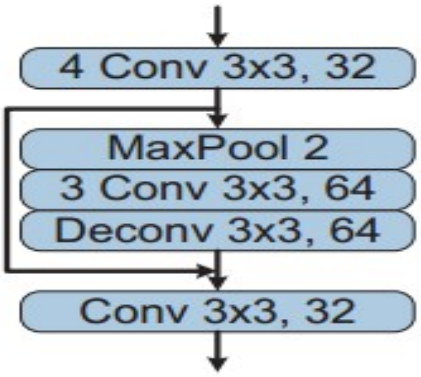

Figure 3.3. Operation in feature extraction unit [13].

3.1.2 Proposal Generation Unit

In the proposal generation stage of the model, the merged feature maps from the sensors are the focal point. This method diverges from conventional approaches that generate proposals using feature maps from singular frames. Such a deviation is crucial, given the unpredictable movement patterns of vehicles, which might be captured in varying positions across multiple sensor frames. The model's strategy involves the concatenation of the feature maps from all frames of each sensor, followed by their integration using a convolution layer. This process results in a combined feature map for each sensor, which is then utilized independently to calculate objectiveness scores and to determine the regression of proposal locations. In the final step, the proposals derived from both sensors are amalgamated using non-maximum suppression (NMS), ensuring a more accurate and comprehensive proposal generation process.

3.2 Region Fusion Network (RFN)

In the MVDNet model, the proposals derived from the region proposal network (RPN) are utilized in the Region of Interest (RoI) poolers to generate features specific to each region. This process involves executing a pooling operation on the designated region within

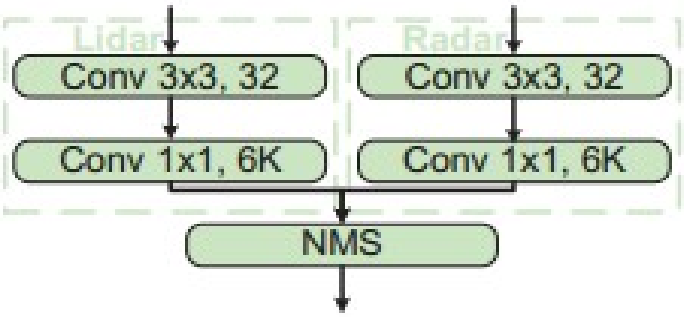

Figure 3.4. Operation in proposal generation unit [13].

the feature map for every sensor and frame corresponding to each proposal. Consequently, this generates feature tensors with dimensions of L×W×C, where 'C' represents the count of channels in the feature maps, and 'L×W' denotes the dimensions of the 2D pooling area. Following this, MVDNet undertakes the fusion of these feature tensors for each proposal through a two-stage process, comprising sensor fusion and temporal fusion. Sensor fusion amalgamates the data from different sensors, while temporal fusion integrates information across various time frames, ensuring a comprehensive and multi-dimensional analysis of the proposals.

3.2.1 Sensor Fusion

In the MVDNet framework, sensor fusion is a crucial process that combines the feature tensors of synchronized pairs of LiDAR and radar frames. This fusion acknowledges the varying significance of LiDAR and radar data in different scenarios, necessitating a weighted approach to their contributions. For instance, in conditions where fog completely obscures a vehicle, LiDAR data yields no points, and thus, the corresponding feature tensors from LiDAR should be assigned lesser weight in the fusion process. Conversely, situations where radar data presents strong signal peaks from non-vehicular background objects, which might mimic the signal peaks of vehicles, require a reduction in the weight of the radar features. This adjustment is informed by the corresponding LiDAR data, which can provide clarity

in such ambiguous scenarios. MVDNet addresses this challenge by adaptively fusing LiDAR and radar features, employing an extended attention block mechanism. This method enables dynamic weighting of sensor data, ensuring that the fusion process is responsive to the specific context and characteristics of the sensor inputs.

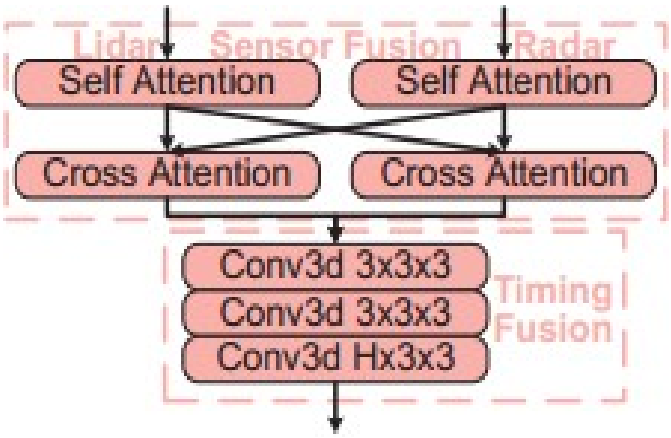

Figure 3.5. Attention mechanism in sensor fusion unit [13].

3.2.2 Temporal Fusion

In the MVDNet architecture, temporal fusion plays a pivotal role by integrating the feature tensors from various frames after the attention layer. Moving away from the use of time-consuming and memory-intensive recurrent structures, MVDNet adopts a different approach. It concatenates the attended feature tensors from distinct frames along an additional dimension, resulting in the formation of 4D feature tensors. Subsequently, 3D convolution layers are applied to these tensors. This technique facilitates the interchange of information along the temporal axis, effectively capturing the dynamics over time.

The final layer in this sequence of 3D convolutions serves a crucial function: it compresses the temporal dimension, yielding a singular, fused tensor of features. This tensor, embodying the essence of temporal dynamics, is then flattened. The flattened tensor undergoes further processing through fully-connected layers. It is within these layers that MVDNet performs critical functions such as inferring objectiveness scores and regressing the locations for the

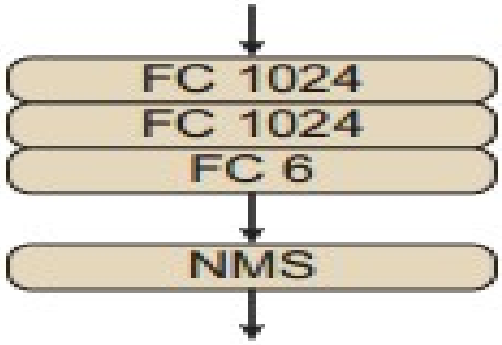

Figure 3.6. Operation in temporal fusion unit of MVDNet [13].

final detections. This process exemplifies how MVDNet efficiently manages and interprets spatiotemporal data, leading to more accurate and temporally coherent detection outcomes.

4. SOFTWARE AND DATASET

4.1 Software and Hardware Requirements

- **NVIDIA GeForce GTX 1080M GPU:** GPUs (Graphics Processing Units) are essential in deep learning for their adeptness at executing parallel computations efficiently. They possess formidable computational capabilities and are compatible with various deep learning frameworks, facilitating accelerated processing and model training.

- **Linux Ubuntu 16.04:** Ubuntu 16.04 Linux is popular due to its extended support duration, robustness, and secure framework, suitable for both development and server use. It offers broad compatibility with various applications and tools, backed by a supportive community, and provides the benefits of open-source customization, all free of charge.

- **CUDA 10.2:** CUDA represents a release of NVIDIA's CUDA platform, tailored for parallel processing on NVIDIA graphics cards. It includes various libraries and tools aimed at boosting computing efficiency, making it especially useful in areas like machine learning, scientific simulations, and graphic computations.

- **Python 3.7:** Python 3.7 is a release of the Python programming language, known for its enhanced capabilities and performance improvements. It's widely used in model development across various disciplines, favored for its ease of use, comprehensive libraries, and robust community backing.

- **Torch 1.8.1 and Torchvision 0.9.1:** Torch and torchvision refer to particular releases of the PyTorch framework and its associated vision library. These versions provide essential functionalities for constructing and training neural networks, making them valuable for artificial intelligence and machine learning projects.

- **Pycocotools:** Pycocotools is a Python toolkit designed for handling the COCO dataset, providing features for loading, analyzing, and annotating images, widely used in computer vision applications.

- **Detectron-2:** Detectron2, a sophisticated software library created by Facebook's AI Research team, specializes in detection of objects and segmentation within images and videos [33]. As an evolution of its predecessor, Detectron, this library is built on the PyTorch framework and offers several notable features-

  Functional Scope: It is adept at performing a variety of computer vision tasks, including object detection, instance segmentation, person keypoint detection, and panoptic segmentation, enabling detailed and precise analysis of visual data.

  Advanced Architectures: The library encompasses several leading-edge model architectures like Mask R-CNN, Faster R-CNN, and RetinaNet, renowned for their efficacy in object detection.

Figure 4.1. Detection of persons from an image using Detectron-2 library [33].

Design Flexibility: The library's modular structure allows for significant experimentation with algorithmic and architectural variations, enhancing its adaptability for diverse vision-related challenges.

Performance Enhancements: Detectron2 benefits from optimizations in the PyTorch framework, resulting in better speed and accuracy compared to its predecessor.

In the architectural design of the proposed Multi-Head MVDNet, Detectron-2 functions as the fundamental framework or 'backbone'. This integration signifies that Multi-Head MVDNet leverages the robust and versatile capabilities of Detectron-2. It provides a solid and reliable base upon which the specialized functionalities of Multi-Head MVDNet are built. This synergy enhances the overall performance and capability of the Multi-Head MVDNet.

4.2 Dataset

The field of autonomous vehicle technology is swiftly progressing and heavily depends on a range of datasets for its development, testing, and validation processes. Below are some of the most renowned and extensively utilized datasets in this area.

- **Waymo Open Dataset:** Originating from Waymo, a front-runner in autonomous driving technologies, this dataset is renowned for its extensive scope and high-fidelity sensory data. It incorporates an array of sensor outputs, including LiDAR and high-resolution cameras, offering a holistic environmental perspective. Waymo dataset predominantly employed for tasks like perception, 3D object identification, and movement forecasting in varied driving settings such as urban and highway scenarios. Its comprehensive and detailed nature makes it an indispensable resource for developing precise and robust perception systems in the autonomous vehicle sector [34].

- **KITTI Dataset:** The KITTI dataset, a collaborative creation by the Karlsruhe Institute of Technology and Toyota Technological Institute, stands as a trailblazer in the domain of autonomous vehicle research. It encompasses a spectrum of sensory data, including stereo imagery, LiDAR point clouds, and GPS/IMU data. This dataset

extensively utilized for benchmarking computer vision functionalities like stereo vision, optical flow, 3D object recognition, and tracking in real-world driving conditions. KITTI has been pivotal in setting performance standards for algorithmic approaches in authentic driving scenarios [35].

- **nuScenes Dataset:** The nuScenes dataset, provided by Aptiv, is a large-scale collection of data specifically curated for advancing autonomous driving technologies. This dataset is characterized by its complete sensor suite, which includes a 32-beam LiDAR, six cameras, and five radars, offering a comprehensive environmental view. It is ideal for a variety of applications such as 3D object detection, motion prediction, and semantic segmentation. With its extensive annotations and diverse situational coverage, nuScenes serves as a vital tool for training and validating autonomous driving systems [3].

- **Oxford Radar RobotCar Dataset:** The Oxford Radar RobotCar Dataset, part of the RobotCar initiative by the University of Oxford, focuses on promoting sustained autonomous vehicle functionality. The ORR dataset is distinguished by its emphasis on radar data with high-resolution radar, the dataset also includes LiDAR, GPS, and monocular images. Especially it is beneficial for research in radar-based perception and localization, particularly under challenging weather conditions and in dynamic urban settings [10].

In the proposed model, the Oxford Radar RobotCar Dataset is utilized to advance the detection of vehicles in challenging weather conditions like fog, leveraging its unique combination of radar and LiDAR technologies. The radar component is particularly valuable for its capability to function effectively in foggy environments, a limitation often encountered with optical sensors. LiDAR contributes by delivering crucial high-resolution spatial data. The selection of the Oxford dataset is further justified by its proven efficacy in a variety of weather conditions, providing a realistic and demanding testbed for the proposed detection models. Its comprehensive and accurate ground truth data is vital for the rigorous evaluation and refinement of these models. Additionally, the datasets emphasis on sustained

functionality in urban settings aligns well with the proposed models aim to develop a detection system that remains effective in diverse urban environments and adverse weather, making it an exemplary choice for exploring fog-related challenges in autonomous vehicle technology.

4.2.1 Sensor Configuration in RobotCar

The data for this study was gathered utilizing the Oxford RobotCar platform, specifically a Nissan LEAF modified for autonomous capabilities, as depicted with its sensor arrangement in Figure**. In this updated release, the RobotCar was outfitted with additional sensors not present in the original configuration. These include:

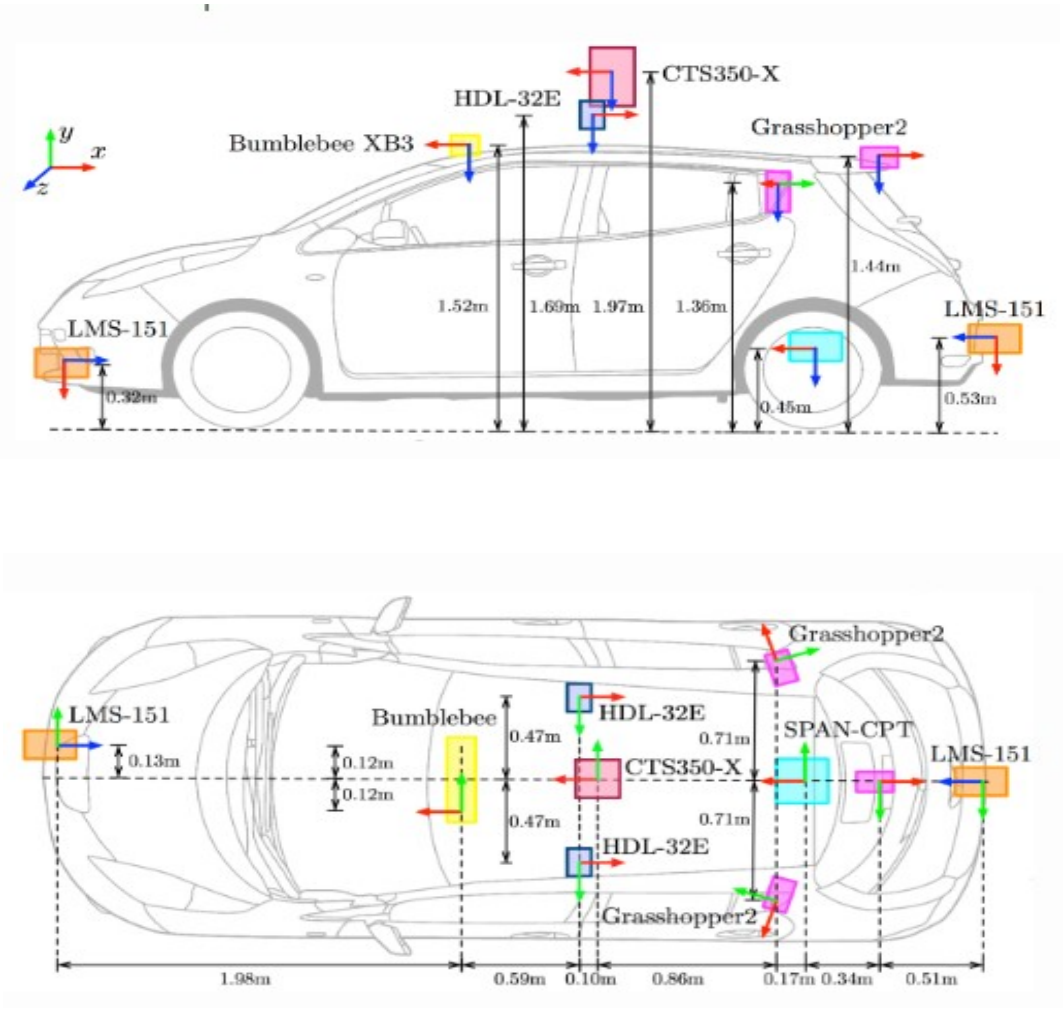

Figure 4.2. Sensor setup within Oxford Radar RobotCar [10].

One - Navtech CTS350-X Millimetre-Wave FMCW Radar:

- Frequency: 4 Hz
- Range: 163 m
- Beamwidth: 1.8 degrees
- Range Resolution: 4.38 cm and 400 measurements/rotation.

Two - Velodyne HDL-32E 3D LIDARs:

- Frequency: 20 Hz
- Range: 100 m
- Horizontal Field of View (HFoV): 360 degrees
- Vertical Field of View (VFoV): 41.3 degrees
- Range Resolution: 2 cm and 32 Channels.

These were in addition to the original sensor array:

One - Point Grey Bumblebee XB3 Trinocular Stereo Camera:

- Frequency: 16 Hz
- Resolution: 1280×960×3
- Horizontal Field of View (HFoV): 180 degrees
- Global shutter and Lens: 3.8 mm

Three - Point Grey Grasshopper2 Monocular Cameras:

- Frequency: 11.1 Hz
- Resolution: 1024×1024
- Horizontal Field of View (HFoV): 180 degrees

- Lens: 2.67 mm fisheye

Two - SICK LMS-151 2D LiDARs:

- Frequency: 50 Hz
- Range: 50 m
- Field of View (FoV): 270 degrees
- Resolution: 0.5 degrees

One - NovAtel SPAN-CPT ALIGN Inertial and GPS Navigation System:

- Frequency: 50 Hz
- 6 axes
- Dual antenna system and GPS/GLONASS compatible.

4.2.2 Data Formatting

Figure- 4.3 presents the standard directory layout for an individual dataset. Each zip file download contains the entire sensor data for a single dataset traversal or processed outputs like stereo VO, without segmenting the sensor data into smaller files.

Radar scans: They are stored as losslessly compressed PNG files in a polar format. In these files, each row corresponds to the sensor's azimuth reading, and each column indicates the raw power return at a specific range. The file structure follows the pattern of <dataset>/radar/<timestamp>.png, <timestamp> represents the UNIX timestamp marking the start of the capture, measured in microseconds. The utilized configuration comprises 400 azimuths per sweep (rows) and 3768 range bins (columns).

Embedded within the first 11 columns of each PNG file, the dataset includes metadata for each azimuth:

- The UNIX Timestamp is recorded as an int64 in columns 1-8.

```
oxford-radar-robotcar-dataset
├── yyyy-mm-dd-HH-MM-SS-radar-oxford-10k or yyyy-mm-dd-HH-MM-SS-radar-oxford-10k-partial
│   ├── gt
│   │   └── radar_odometry.csv  # ground truth SE2 radar odometry
│   ├── radar
│   │   ├── <timestamp>.png     # Navtech radar data
│   │   ├── ...
│   ├── velodyne_left
│   │   ├── <timestamp>.bin     # Velodyne binary sensor data
│   │   ├── <timestamp>.png     # Velodyne raw sensor data
│   │   ├── ...
│   ├── velodyne_right
│   │   ├── <timestamp>.bin     # Velodyne binary sensor data
│   │   ├── <timestamp>.png     # Velodyne raw sensor data
│   │   ├── ...
│   ├── radar.timestamps
│   ├── velodyne_left.timestamps
│   ├── velodyne_right.timestamps
│   │   # Plus the same layout of files from the original Oxford RobotCar Dataset
```

Figure 4.3. Layout of the directory after unzipping ORR dataset [10].

- The Sweep counter is stored as a uint16 in columns 9-10. This is converted to an angle using the formula: Angle of azimuth (radians) $= \left(\frac{\text{sweep counter}}{\text{encoder size}}\right) \times 2\pi$. The encoder size is consistently set to 5600.

- A Valid flag, represented as a uint8 in column 11, is included. Occasionally, a minimal number of data packets carrying azimuth returns are dropped by the Navtech radar.

Velodyne LiDAR scans: 3D Velodyne LiDAR scans are made available in this dataset in two distinct formats. The first is a raw format that encapsulates the complete sensor data for users to manipulate as needed, while the second is a binary format that presents the point cloud of each scan without motion compensation.

Raw Scans: These scans are distributed as lossless PNG files, where each column corresponds to sensor readings at specific azimuths. The file structure follows <dataset>/<laser>/

<timestamp>.png, with <laser> indicating either velodyne-left or velodyne-right and <timestamp> being the UNIX timestamp at the start of capture in microseconds. To provide comprehensive raw data, we embed azimuth-specific metadata in the PNG files in the following rows:

- Rows 1-32 contain the raw intensity of each laser, recorded as uint8.
- Rows 33-96 present the raw range of each laser as uint16, which can be converted to meters using the formula: range in meters = raw range $\times 0.02$.
- Rows 97-98 hold the sweep counter as uint16, converted to an angular measurement in radians using: Angle in radians = $\left(\frac{\text{sweep counter}}{18000}\right) \times \pi$.
- Rows 99-106 include approximate UNIX timestamps, formatted as int64.

Binary Scans: For the binary format, Velodyne data is provided as floating-point values in a binary file, akin to the Velodyne scan structure used in the KITTI dataset. The file arrangement is <laser>/<timestamp>.bin, where <laser> refers to either velodyne-left or velodyne-right and <timestamp> is the UNIX timestamp in microseconds.

Each binary scan consists of a series of values represented as (x, y, z, I) $\times$N, where x, y, z denote the 3D Cartesian coordinates of the LiDAR return relative to the sensor (in meters), and I represents the measured intensity.

5. PROPOSED ARCHITECTURE

5.1 Utilizing the Transformer Backbone

The Transformer model, pivotal in the advancement of natural language processing (NLP), is distinguished by its novel application of the attention mechanism, primarily self-attention, and its unique encoder-decoder framework. Introduced in the groundbreaking paper "Attention Is All You Need" by Vaswani and colleagues in 2017 [12], it represents a departure from traditional sequence-to-sequence processing models that relied on recurrent and convolutional neural networks.

5.1.1 Encoder-Decoder Composition

Encoder Structure: Comprising several identical layers, the encoder in the Transformer model includes a multi-head self-attention mechanism and a position-wise, fully connected feed-forward network. Each layer is equipped with residual connections and layer normalization, enhancing the stability and efficiency of the learning process.

Decoder Structure: Mirroring the encoder, the decoder also consists of multiple identical layers. Each layer, however, integrates an extra sub-layer for multi-head attention over the encoder's output, enabling targeted focus on pertinent segments of the input sequence, reminiscent of alignment in traditional sequence-to-sequence models.

Incorporating Positional Information: To account for the sequence order, which is not inherently captured due to the absence of recurrent or convolutional structures, the Transformer uses positional encodings. These are added to the input embeddings at the base of both the encoder and decoder stacks.

5.1.2 Attention Mechanism in Transformer

Self-Attention as Core: Central to the Transformer's design is the self-attention mechanism. This innovative approach allows the model to concurrently process each position of a sequence, facilitating a comprehensive understanding of the context, a significant leap from the sequential processing of earlier models. Self-attention uniquely enables each element

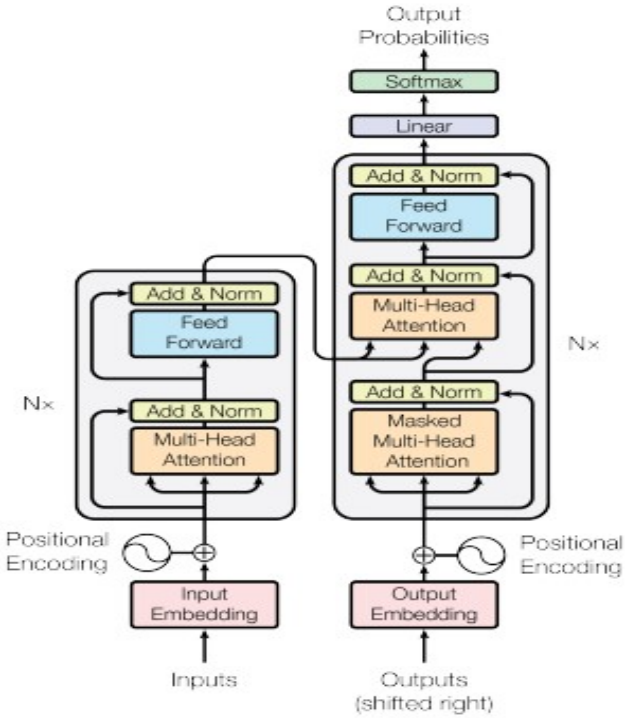

Figure 5.1. Structure of Transformer model with encoder and decoder [12].

within a sequence to be interpreted in relation to the entire sequence. In language processing, this translates to understanding each word through the lens of the entire sentence, thereby capturing context in a fluid and adaptable manner. The ability of self-attention to discern varying meanings based on context is crucial for addressing linguistic ambiguities. One of the standout features of self-attention is its capability to identify and utilize long-range dependencies within sequences, overcoming the limitations faced by traditional sequential models [36].

The self-attention mechanism involves computing relevance scores, which determine the level of focus each element of the sequence should receive in relation to others. These scores are derived from transformed representations of the input sequence, typically referred to as query (Q), key (K), and value (V) matrices. The mechanism operates through matrix multiplications, where the query matrix corresponding to a specific element is multiplied by the key matrix for all elements, yielding scores that reflect the element's relevance within the

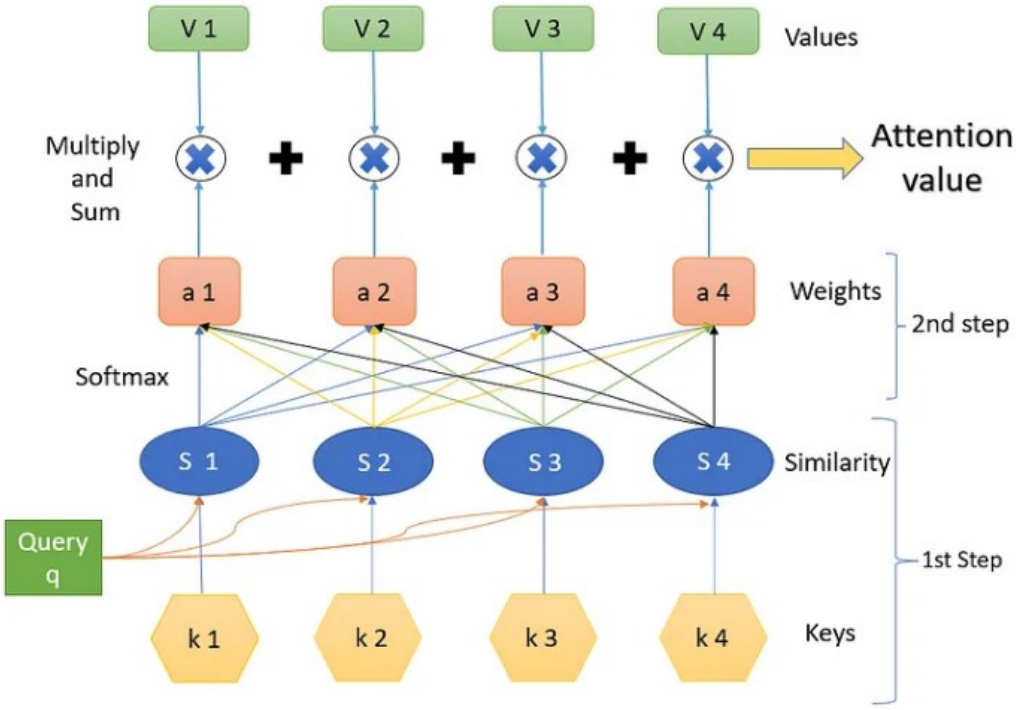

Figure 5.2. Detail illustration of self-attention mechanism [37].

sequence. Through its matrix-based approach, self-attention processes the entire sequence in parallel, significantly boosting computational efficiency, especially for extensive sequences.

The attention scores effectively create a context-weighted aggregation of the sequence, where higher scores indicate greater relevance or influence between elements. These scores are normalized, often through a softmax function, transforming them into a probabilistic distribution. This step ensures a focused and balanced allocation of attention across the sequence. The final output for each sequence element in the self-attention layer is computed as a weighted sum of these normalized attention scores and the value matrix. This process yields an output that integrates contextual information from the entire sequence.

Parallel Processing through Multi-Head Attention: The model employs multi-head attention, executing multiple attention processes in parallel. This multi-faceted approach enables the Transformer to capture varied aspects of information across different positions and representation subspaces.

5.2 Proposed Multi-Head Attention Layer

In the proposed architecture, the MVDNet's capabilities are augmented through the implementation of a Multi-Head attention mechanism. This mechanism operates by segmenting the attention process into several distinct units, referred to as 'heads'. Each head is tasked with concentrating on specific attributes of the input data. This division enables the concurrent examination of varying data characteristics, such as spatial connections or intensity levels of signals. As a result, the MVDNet is equipped to assimilate a more expansive array of information or features from the input than possible with a singular attention mechanism, which traditionally focuses on one data aspect at any given time. The efficacy

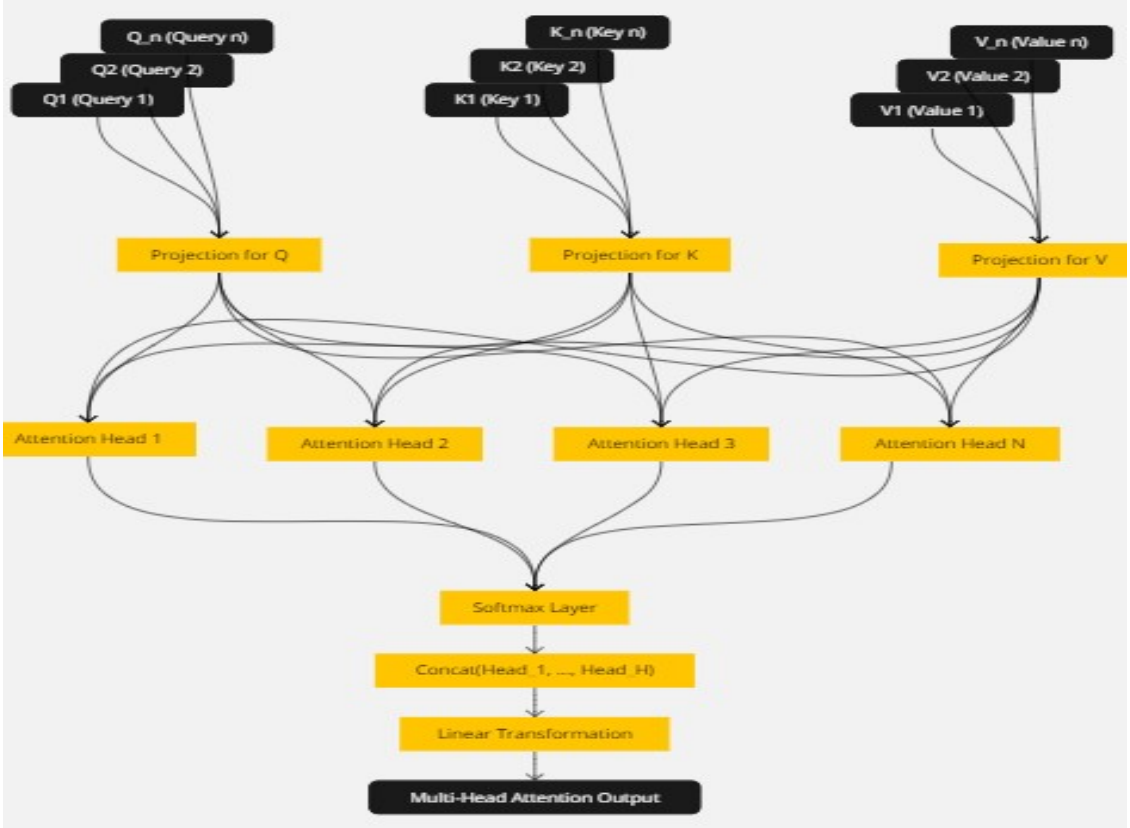

Figure 5.3. Multi-Head attention mechanism.

of this approach is further enhanced by the mechanism's ability to dynamically evaluate and juxtapose attention scores across different data segments. This dynamic focus adjustment facilitates a thorough and nuanced analysis, significantly contributing to the refinement and

performance enhancement of the baseline MVDNet model. The Multi-Head attention layer is implemented within the Region Fusion Network (RFN) unit of MVDNet by replacing its self-attention layer.

Within the architecture of the multi-head attention mechanism, the process begins by segmenting the input sequences, predominantly constituted of feature tensors from lidar and radar, into a series of distinct subspaces. In this context, these sequences undertake the roles of queries (Q), keys (K), and values (V), integral to the mechanism's functionality. Each specific attention head, identified as h, applies a unique linear transformation to these sequences. This is facilitated through a series of different, trainable weight matrices, each designated for a respective attention head. Such an approach, involving the subdivision and individualized processing of the input sequences, is essential in the multi-head attention mechanism. It allows for the parallel processing and examination of diverse aspects and characteristics of the lidar and radar data, thereby enriching the overall analytical capacity of the mechanism.

$$Q_h = QW_h^Q \tag{5.1}$$

$$K_h = KW_h^K \tag{5.2}$$

$$V_h = VW_h^V \tag{5.3}$$

Here, W_h^Q, W_h^K, and W_h^V are the weight matrices for the queries, keys, and values, respectively. In the multi-head attention framework, each attention head is equipped with its unique set of weight matrices. These matrices are instrumental in enabling the model to process diverse segments of the input concurrently. The computation of attention scores within each head is achieved through a scaled dot-product attention mechanism, which serves as an indicator of the alignment between queries and keys.

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{5.4}$$

The use of a scaling factor, denoted as $\sqrt{d_k}$ and and derived from the key dimensionality (d_k), plays a critical role in moderating the magnitude of the dot products. This moderation is crucial for preventing the gradients from becoming excessively small during the training phase, a phenomenon known as vanishing gradients. Subsequent to the calculation of attention outputs across the individual heads, these outputs undergo a concatenation followed by a linear transformation. This final stage effectively amalgamates the insights garnered by each attention head, culminating in a unified output that encapsulates the collective intelligence of the multi-head attention system.

$$\text{MultiHead}(Q, K, V) = \text{Concat}(head_1, \ldots, head_H)W^0 \tag{5.5}$$

The output from each attention head, denoted as headh, is generated by applying the attention function.

$$head_h = \text{Attention}(Q_h, K_h, V_h) \tag{5.6}$$

Subsequently, these outputs from the various heads are merged through concatenation. The aggregated result of this concatenation is then further processed by being multiplied with an additional trainable weight matrix, known as W^0.

6. RESULTS

6.1 Training Setup

The Oxford Radar Robotcar dataset is composed of 8,862 instances, split into two distinct groups: a training set with 7,071 entries and a testing set containing 1,791 entries, with a strict non-overlapping geographical criterion for both sets. The model's training regimen commences with a learning rate set at 0.01, employing the Stochastic Gradient Descent (SGD) optimizer. This learning rate undergoes a consistent decrement of 0.1 every 40,000 iterations throughout the training duration. This training phase extends over a total of 85,000 iterations, starting from a baseline untrained model.

The acquisition of the original ORR data entailed the deployment of a vehicle equipped with a centrally roof-mounted NavTech CTS350-X radar. This radar operated in synergy with two Velodyne HDL-32E lidars, with a fusion of their respective outputs. The synchronization strategy for the lidar and radar data diverged from the traditional method of aligning each radar scan with the closest temporal lidar scan. The adopted strategy involved the amalgamation of all lidar scans (F=5) within the time frame of an individual radar scan. More specifically, a segment of points from each lidar scan was meticulously selected so that the lidar and radar scan approximately simultaneously covered this segment. For each radar frame accompanied by its simultaneous F=5 lidar frames, the inclusion of a point 'p' from the n-th lidar frame was conditional, based on its placement within the spatial range delineated by the intervals $\left[\frac{n-1}{F+1}\boldsymbol{\pi}, \frac{n+1}{F+1}\boldsymbol{\pi}\right]$.

6.2 Evaluation Metrics

In the field of object detection, the evaluation of models predominantly hinges on four key metrics: Precision, Recall, Average Precision (AP), and Intersection over Union (IoU).

Precision: Precision is indicative of the accuracy of a model's positive predictions. It quantifies the proportion of positive identifications that were indeed correct. A high precision score implies that the model's positive predictions are generally reliable. This is crucial in scenarios where the consequences of false positives are significant. Precision is computed

by dividing the number of true positive predictions by the sum of true positives and false positives [38].

$$\text{Precision} = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)} + \text{False Positives (FP)}} \tag{6.1}$$

Recall: Recall assesses the model's capacity to identify all actual positive instances. It measures the extent to which the model captures the actual positives. A model with a high recall score efficiently identifies positive cases, minimizing the number of false negatives. This is vital in areas where missing a positive case has serious consequences. Recall is calculated as the ratio of true positives to the aggregate of true positives and false negative [38].

$$\text{Recall} = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)} + \text{False Negatives (FN)}} \tag{6.2}$$

Average Precision (AP): Average Precision (AP) stands as a vital metric in the domain of object detection, offering a comprehensive evaluation of a model's performance across a range of recall levels. This metric transcends the limitations of singular measures like precision or recall, by encapsulating a more extensive assessment of model efficacy.

AP is essentially quantified as the area under the precision-recall curve. This curve is formulated by plotting precision against recall at various threshold levels, where precision is the proportion of correct positive predictions relative to the total positive predictions, and recall is the proportion of correct positive predictions out of all actual positives.

The precision and recall values are computed by altering the threshold for a positive prediction. These variations influence the precision and recall, thereby shaping the precision-recall curve. The calculation of AP involves determining the area under this curve. A larger area signifies a superior performance of the object detection model.

Intersection over Union (IoU): IoU is a critical metric in computer vision, predominantly employed in tasks like object detection and segmentation, to evaluate the precision of an object detector on a specific dataset. It quantifies the extent of overlap between the predicted bounding box and the actual ground truth bounding box. IoU calculates the ratio of the area of overlap (intersection) between the predicted and ground truth bounding boxes to their combined coverage area (union) [39]. The IoU score directly reflects the precision

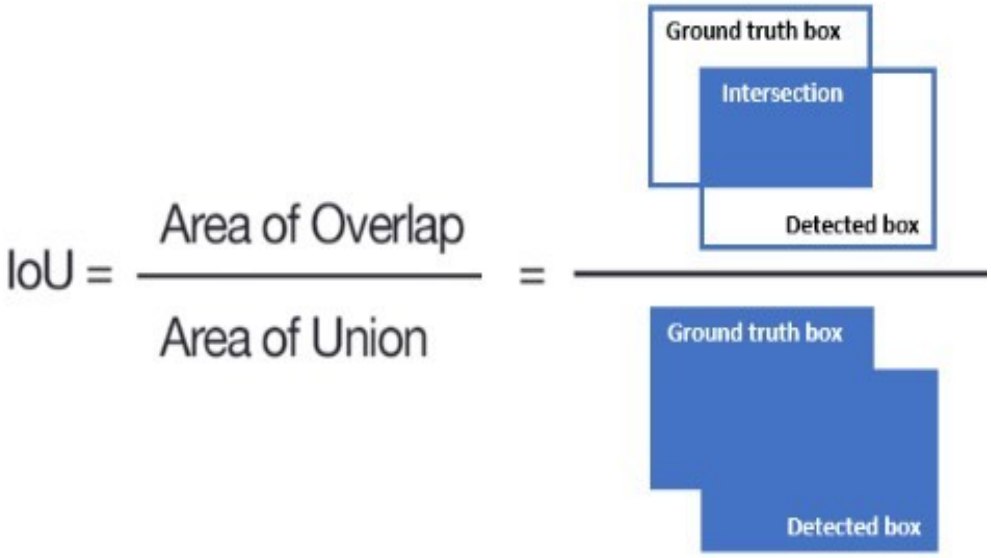

Figure 6.1. Intersection over Union (IoU) for object detection [40].

with which a model localizes an object. A higher IoU value signifies a closer alignment between the model's prediction and the actual object location. Common practice involves setting an IoU threshold (such as 0.5) to categorize a detection as either a true positive or a false positive, which standardizes the assessment of detection correctness.

6.3 Selection of Multi-Head Number

Determining the appropriate quantity of heads within a multi-head attention module of a neural network framework is a pivotal choice that significantly influences the efficacy of the model. An increase in the number of heads augments the model's capability to assimilate intricate characteristics. Each head assimilates a unique interpretation of the input data, enabling the model to encapsulate a diverse array of data facets.

The dimensions of the key, query, and value vectors within each head are typically a fraction of the overall dimensionality, divided equally by the total number of heads. As a result, a greater number of heads corresponds to a reduced dimensionality for each head, potentially impacting the model's proficiency in comprehending complex patterns. The

requirement for enhanced computational resources escalates with an increase in the number of heads. Under scenarios of restricted computational resources, a reduction in the number of heads may be advisable. Occasionally, amplifying the number of layers within the model can be more beneficial than increasing the number of heads. Enhanced layering facilitates a more profound extraction of features, whereas a greater number of heads enables a broader scope of feature extraction.

Within the framework of a multi-head attention mechanism, it is imperative for the dimensionality of the input, which encompasses the dimensions of the key, query, and value vectors, to be divisible by the designated number of heads. This divisibility criterion is essential as it guarantees an equitable distribution of the input dimension across each head, thereby facilitating uniform processing by all heads. Consequently, in the proposed architecture the selection of heads was confined to numbers such as 2, 4, 7, 14, 21, and 49, which are compatible with the divisibility requirement of the baseline model's input dimension. Following extensive experimental analysis, we narrowed our focus to comparing the effectiveness of using 4, 7, and 14 heads. This comparison aimed to ascertain the optimal number of heads for the proposed model. Table- 6.1 presents a comparative analysis of these configurations, delineating the variations in average precision across different values of Intersection over Union (IoU) for 4, 7, and 14 heads. This comparison forms the basis for the decision on the most suitable number of heads for the proposed Multi-Head Vehicle Detection model's architecture.

Table 6.1. Comparison of Average Precision (AP) in terms of IoU for different numbers of heads in multi-head attention.

No. of Head	**IoU = 0.5**	**IoU = 0.65**	**IoU = 0.8**	**IoU = 0.5 : 0.05 : 0.95**
4	89.90%	88.20%	73.90%	67.60%
7	91.20%	88.90%	74.10%	67.90%
14	89.30%	88.10%	73.80%	67.60%

In this research, an experiment was conducted to evaluate the optimal number of heads in a model, focusing on the choices of 4, 7, and 14 heads. This experiment aimed to assess

the Average Recall (AR) to varying levels of maximum detections set at 1, 10, and 100. The results, as detailed in Table- 6.2, indicate a notable outcome: the configuration employing 7 heads demonstrated a consistently higher Average Recall (AR) across all specified maximum detection categories. This empirical evidence underscores the significance of the number of heads in the attention mechanism, particularly emphasizing the effectiveness of 7 heads in enhancing recall performance within the model, regardless of the set threshold for maximum detections.

Table 6.2. Comparison of Average Recall (AR) in terms of maximum detection for different numbers of heads.

No. of Head	**Max. Detection 1**	**Max. Detection 10**	**Max. Detection 100**
4	12.80%	78.20%	89.40%
7	12.80%	78.60%	90.30%
14	12.80%	77.80%	89.20%

Upon conducting a comparative analysis of both Average Precision (AP) with respect to Intersection over Union (IoU) and Average Recall (AR) considering various levels of maximum detection for head numbers 4, 7, and 14, the selection of head number 7 was determined to be the most suitable for implementation within the MVDNet fusion network unit. This decision was based on the observed performance metrics, where head number 7 consistently showed optimal results in enhancing both the precision and recall capabilities of the network.

6.4 Performance Analysis

Figure- 6.2 presents a comparative evaluation of the loss metrics between the Multi-Head Vehicle Detection Network and the baseline MVDNet, charted over the iterations. This comparison, exemplified through the course of 1000 iterations within a single epoch, offers insights into the performance dynamics of the two network configurations.

This research rigorously evaluates the efficacy of the proposed Multi-Head MVDNet across diverse meteorological conditions, particularly focusing on scenarios with fog and clear weather. The assessment encompasses a comparative analysis involving the Multi-

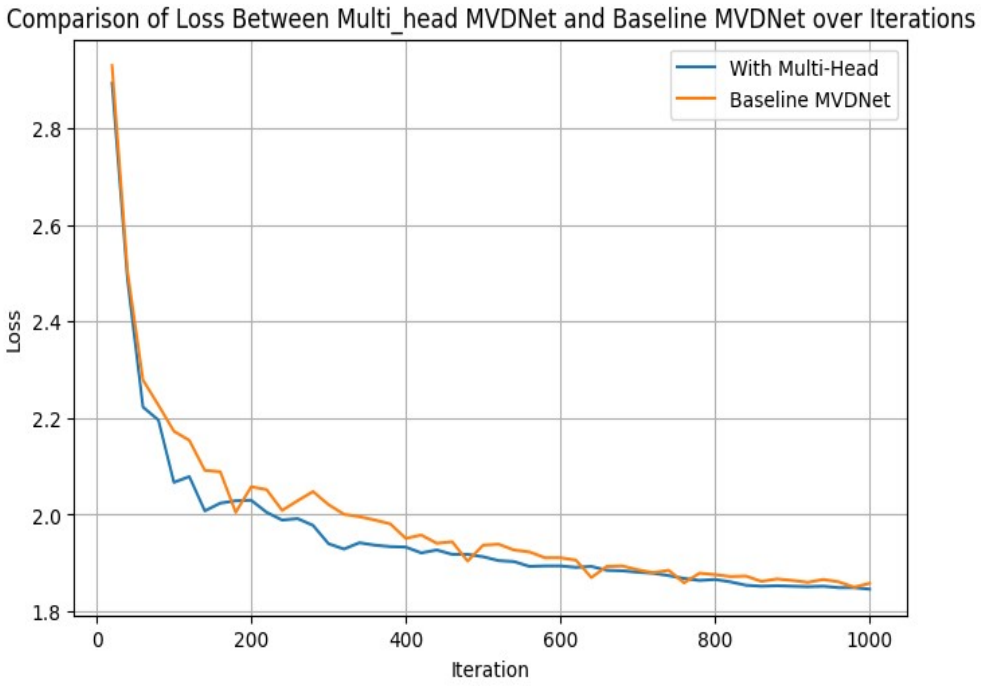

Figure 6.2. Example of one epoch: loss graphs of Multi-Head and baseline MVDNet over first 1000 iterations.

Head MVDNet and various benchmark systems, including the baseline MVDNet, the DEF lidar-radar fusion method, and systems utilizing solely lidar or radar. Findings, as detailed in Table- 6.3, 6.4, 6.5 demonstrate that the Multi-Head MVDNet consistently surpasses the aforementioned methods in performance, as evidenced across a spectrum of Intersection over Union (IoU) values.

Table 6.3. Comparison of Average Precision (AP) for different methods when IoU = 0.5.

Method	**IoU = 0.5**
No Sensor Fusion	87.89%
No Self Attention	88.19%
DEF	84.02%
Radar-Only	73.04%
Lidar-Only	82.28%
MVDNet (Base)	89.15%
Multi-Head MVDNet (Proposed)	91.20%

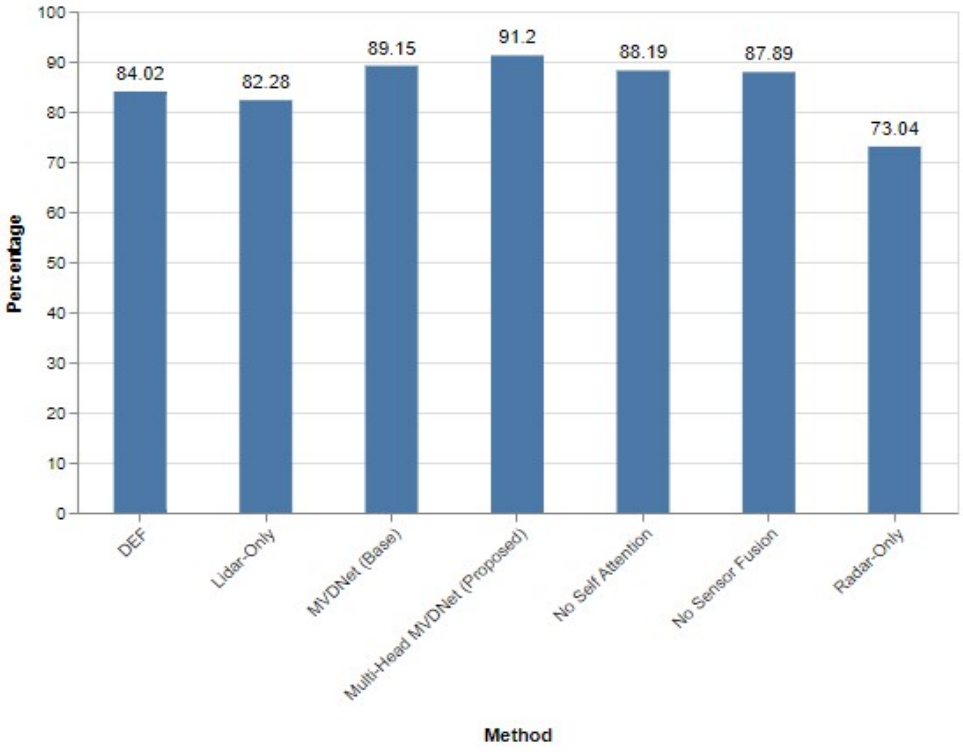

Figure 6.3. Performance evaluation of different methods when IoU = 0.5.

Table- 6.3 highlights that when the Intersection over Union (IoU) threshold is established at 0.5, the Multi-Head MVDNet demonstrates an enhancement in average precision (AP) by 2.05% and 3.01% relative to the baseline MVDNet and the conventional MVDNet without self-attention mechanism, respectively.

Table 6.4. Average Precision (AP) for different methods when IoU = 0.65.

Method	**IoU = 0.65**
No Sensor Fusion	85.59%
No Self Attention	85.88%
DEF	75.32%
Radar-Only	68.27%
Lidar-Only	80.72%
MVDNet (Base)	86.72%
Multi-Head MVDNet (Proposed)	88.90%

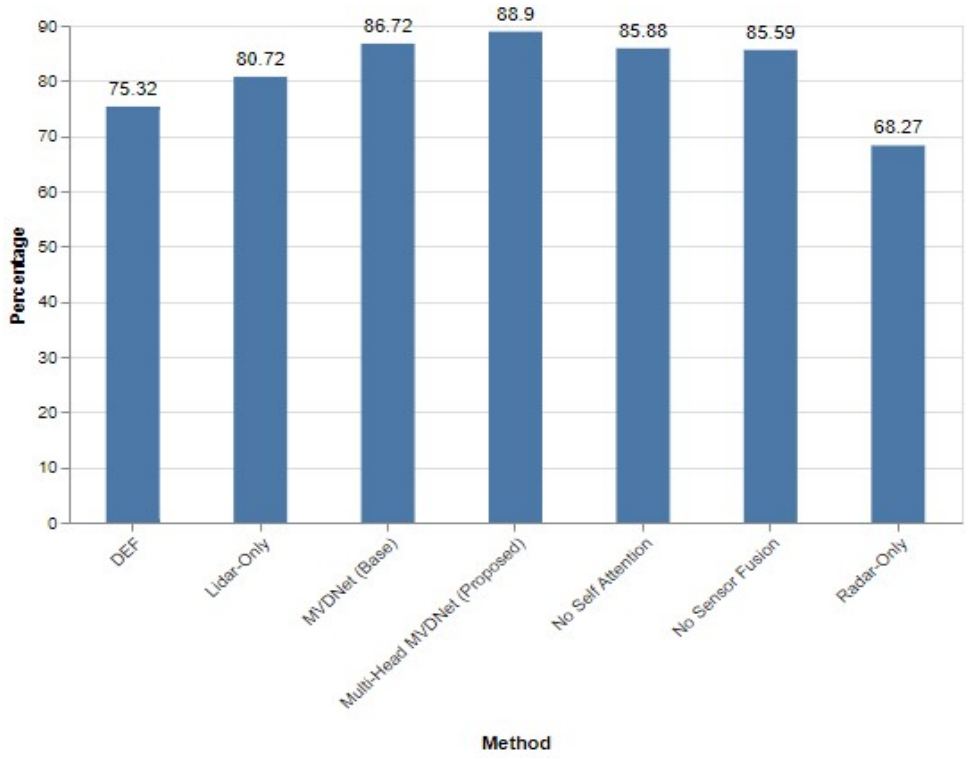

Figure 6.4. Performance evaluation of different methods when IoU = 0.65.

Incorporating the multi-head attention concept into the MVDNet has led to its superior performance over other related methodologies. Furthermore, when evaluated at IoU thresholds of 0.65 and 0.8, the Multi-Head MVDNet model exhibits increments in average precision (AP) by 2.18% and 2.34%, respectively, in comparison to the baseline MVDNet model.

Table 6.5. Comparison of Average Precision (AP) for different methods when IoU = 0.8.

Method	**IoU = 0.8**
No Sensor Fusion	70.61%
No Self Attention	71.41%
DEF	43.62%
Radar-Only	43.25%
Lidar-Only	67.83%
MVDNet (Base)	71.76%
Multi-Head MVDNet (Proposed)	74.10%

7. CONCLUSION

The development of the Multi-Head Vehicle Detection Network represents a significant milestone in the realm of vehicle detection technologies, offering enhanced performance, particularly in challenging environmental conditions such as snow and fog. This advanced model is an extension of the pre-existing MVDNet framework. The foundational strength of the original MVDNet lies in its strategic utilization of the synergistic capabilities of lidar and radar technologies, achieved by fusing their extracted feature tensors at a deeper stage in the processing sequence. The Multi-Head Vehicle Detection Network introduces a critical and innovative change: the implementation of a multi-head attention layer within its fusion algorithm. This new layer is a departure from the standard self-attention mechanism used in the earlier MVDNet version, effectively dividing the attention mechanism into multiple segments for more refined processing.

Through rigorous experimental analysis involving various configurations of attention heads, the model incorporating seven heads was identified as the most effective. This specific setup has become the standard for the multi-head attention component of the model. In the experimental validation of the Multi-Head MVDNet, the ORR dataset, noted for its high-resolution radar and lidar data, was utilized. The experimental results were compelling, demonstrating that the Multi-Head Vehicle Detection Network consistently outperforms not only the original MVDNet but also other models such as those based solely on lidar and the DEF model, in terms of vehicle detection accuracy.

A notable aspect of these findings is the improvement in average precision (AP). When the Intersection over Union (IoU) threshold was set to 0.5, the Multi-Head MVDNet exhibited a **2.05%** increase in AP over the baseline MVDNet. This trend of enhanced performance was also evident at higher IoU thresholds of 0.65 and 0.8, where the Multi-Head Vehicle Detection Network showed increases in AP by **2.18%** and **2.34%**, respectively, compared to the baseline model. This evidences the superiority of the multi-head attention approach over existing methods in the field.

Looking forward, there is an ambitious plan to expand the capabilities of the system further. This includes experimentation with different types of attention mechanisms to

potentially improve the model's efficiency and accuracy. Additionally, there is a keen interest in exploring the application of this technology for real-time vehicle detection, which could have significant implications for various practical applications in the field of autonomous vehicle navigation and traffic management.

8. FUTURE SCOPE

Although the proposed architecture has demonstrated substantial improvements over the foundational model, there exists an opportunity for further exploration to enhance its accuracy. The following are some proposed methodologies for future implementation:

- **Advanced Attention Mechanisms:** Building upon the successful application of the multi-head attention mechanism, future studies could delve into exploring more complex or innovative attention-based models, such as those based on transformer architectures, to potentially enhance both accuracy and computational efficiency.
- **Real-Time Application and Optimization:** Future research could focus on adapting the Multi-Head Vehicle Detection Network for real-time applications. This involves refining the model to achieve a balance between rapid processing capabilities and maintaining high accuracy, which is vital for real-time systems like autonomous vehicles and traffic surveillance.
- **Diverse Environmental Testing:** Expanding the testing environment for the Multi-Head MVDNet to include a broader spectrum of conditions, such as intense rainfall, varied lighting environments, and different geographic settings, could provide valuable insights into the models adaptability and resilience.
- **Integration with Varied Sensory Inputs:** Augmenting lidar and radar data with other types of sensory information, such as visual imagery or ultrasonic signals, could further refine vehicle detection. Research into effective methods for integrating these disparate data sources would be beneficial.
- **Utilization in Autonomous Vehicle Navigation:** Applying the Multi-Head Vehicle Detection Network within autonomous vehicle navigation systems presents a practical and significant application area. Research could concentrate on how this model can enhance navigation precision and obstacle detection in autonomous driving systems.

- **Scalability and Real-World Deployment:** Investigating the scalability of the model and challenges associated with its deployment in diverse hardware settings or vehicle types is crucial for practical applications.
- **Advanced Data Fusion Methods:** Expanding upon the current data fusion methods, research into more sophisticated techniques, such as early, intermediate or hierarchical fusion approaches, could provide insights into the most effective ways to combine sensor data.

REFERENCES

[1] *J3016_202104: Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles - SAE International*. [Online]. Available: https://www.sae.org/standards/content/j3016_202104/.

[2] M. Bijelic, T. Gruber, F. Mannan, *et al.*, *Seeing Through Fog Without Seeing Fog: Deep Multimodal Sensor Fusion in Unseen Adverse Weather*, arXiv:1902.08913 [cs], Jun. 2020. DOI: 10.48550/arXiv.1902.08913. [Online]. Available: http://arxiv.org/abs/1902.08913.

[3] H. Caesar, V. Bankiti, A. H. Lang, *et al.*, *nuScenes: A multimodal dataset for autonomous driving*, arXiv:1903.11027 [cs, stat], May 2020. DOI: 10.48550/arXiv.1903.11027. [Online]. Available: http://arxiv.org/abs/1903.11027.

[4] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? The KITTI vision benchmark suite," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, ISSN: 1063-6919, Jun. 2012, pp. 3354–3361. DOI: 10.1109/CVPR.2012.6248074. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/6248074.

[5] X. Chen, H. Ma, J. Wan, B. Li, and T. Xia, *Multi-View 3D Object Detection Network for Autonomous Driving*, arXiv:1611.07759 [cs], Jun. 2017. DOI: 10.48550/arXiv.1611.07759. [Online]. Available: http://arxiv.org/abs/1611.07759.

[6] J. Ku, M. Mozifian, J. Lee, A. Harakeh, and S. Waslander, *Joint 3D Proposal Generation and Object Detection from View Aggregation*, en, Dec. 2017. [Online]. Available: https://arxiv.org/abs/1712.02294v4.

[7] C. R. Qi, W. Liu, C. Wu, H. Su, and L. J. Guibas, *Frustum PointNets for 3D Object Detection from RGB-D Data*, arXiv:1711.08488 [cs], Apr. 2018. DOI: 10.48550/arXiv.1711.08488. [Online]. Available: http://arxiv.org/abs/1711.08488.

[8] M. Bijelic, T. Gruber, and W. Ritter, "A Benchmark for Lidar Sensors in Fog: Is Detection Breaking Down?" In *2018 IEEE Intelligent Vehicles Symposium (IV)*, arXiv:1912.03251 [cs], Jun. 2018, pp. 760–767. DOI: 10.1109/IVS.2018.8500543. [Online]. Available: http://arxiv.org/abs/1912.03251.

[9] Y. Golovachev, A. Etinger, G. Pinhasi, and Y. Pinhasi, "Millimeter Wave High Resolution Radar Accuracy in Fog ConditionsTheory and Experimental Verification," en, *Sensors*, vol. 18, no. 7, p. 2148, Jul. 2018, ISSN: 1424-8220. DOI: 10.3390/s18072148. [Online]. Available: http://www.mdpi.com/1424-8220/18/7/2148.

[10] D. Barnes, M. Gadd, P. Murcutt, P. Newman, and I. Posner, "The Oxford Radar RobotCar Dataset: A Radar Extension to the Oxford RobotCar Dataset," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*, ISSN: 2577-087X, May 2020, pp. 6433–6438. DOI: 10.1109/ICRA40945.2020.9196884. [Online]. Available: https://ieeexplore.ieee.org/document/9196884.

[11] D. Konstantinidis, I. Papastratis, K. Dimitropoulos, and P. Daras, "Multi-Manifold Attention for Vision Transformers," *IEEE Access*, vol. 11, pp. 123 433–123 444, 2023, ISSN: 2169-3536. DOI: 10.1109/ACCESS.2023.3329952. [Online]. Available: https://ieeexplore.ieee.org/document/10305583.

[12] A. Vaswani, N. Shazeer, N. Parmar, *et al.*, *Attention Is All You Need*, arXiv:1706.03762 [cs], Aug. 2023. DOI: 10.48550/arXiv.1706.03762. [Online]. Available: http://arxiv.org/abs/1706.03762.

[13] K. Qian, S. Zhu, X. Zhang, and L. E. Li, "Robust Multimodal Vehicle Detection in Foggy Weather Using Complementary Lidar and Radar Signals," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, ISSN: 2575-7075, Jun. 2021, pp. 444–453. DOI: 10.1109/CVPR46437.2021.00051. [Online]. Available: https://ieeexplore.ieee.org/document/9578621.

[14] J.-B. Cordonnier, A. Loukas, and M. Jaggi, *Multi-Head Attention: Collaborate Instead of Concatenate*, en, Jun. 2020. [Online]. Available: https://arxiv.org/abs/2006.16362v2.

[15] S. M. Group, *What's Best for Autonomous Cars: LiDAR vs Radar vs Cameras*, en, Sep. 2020. [Online]. Available: https://www.techbriefs.com/component/content/article/37699-what-s-best-for-autonomous-cars-lidar-vs-radar-vs-cameras.

[16] D. Xu, D. Anguelov, and A. Jain, "PointFusion: Deep Sensor Fusion for 3D Bounding Box Estimation," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, ISSN: 2575-7075, Jun. 2018, pp. 244–253. DOI: 10.1109/CVPR.2018.00033. [Online]. Available: https://ieeexplore.ieee.org/document/8578131.

[17] A. H. Lang, S. Vora, H. Caesar, L. Zhou, J. Yang, and O. Beijbom, *PointPillars: Fast Encoders for Object Detection from Point Clouds*, arXiv:1812.05784 [cs, stat], May 2019. DOI: 10.48550/arXiv.1812.05784. [Online]. Available: http://arxiv.org/abs/1812.05784.

[18] R. Q. Charles, H. Su, M. Kaichun, and L. J. Guibas, "PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, ISSN: 1063-6919, Jul. 2017, pp. 77–85. DOI: 10.1109/CVPR.2017.16. [Online]. Available: https://ieeexplore.ieee.org/document/8099499.

[19] Y. Li, P. Duthon, M. Colomb, and J. Ibanez-Guzman, *What happens to a ToF LiDAR in fog?* en, Mar. 2020. [Online]. Available: https://arxiv.org/abs/2003.06660v4.

[20] B. Yang, W. Luo, and R. Urtasun, "PIXOR: Real-time 3D Object Detection from Point Clouds," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, ISSN: 2575-7075, Jun. 2018, pp. 7652–7660. DOI: 10.1109/CVPR.2018.00798. [Online]. Available: https://ieeexplore.ieee.org/document/8578896.

[21] M. Parker, "Chapter 20 - Automotive RadarWith contributions by Ben Esposito.," in *Digital Signal Processing 101 (Second Edition)*, M. Parker, Ed., Newnes, Jan. 2017, pp. 253–276, ISBN: 9780128114537. DOI: 10.1016/B978-0-12-811453-7.00020-2. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B9780128114537000202.

[22] *Radar Principle - Radartutorial*, en. [Online]. Available: https://www.radartutorial.eu/01.basics/Radar%20Principle.en.html.

[23] B. Yang, R. Guo, M. Liang, S. Casas, and R. Urtasun, *RadarNet: Exploiting Radar for Robust Perception of Dynamic Objects*, arXiv:2007.14366 [cs], Jul. 2020. DOI: 10.48550/arXiv.2007.14366. [Online]. Available: http://arxiv.org/abs/2007.14366.

[24] M. Liang, B. Yang, S. Wang, and R. Urtasun, *Deep Continuous Fusion for Multi-Sensor 3D Object Detection*, arXiv:2012.10992 [cs], Dec. 2020. DOI: 10.48550/arXiv.2012.10992. [Online]. Available: http://arxiv.org/abs/2012.10992.

[25] H. Kuang, X. Liu, J. Zhang, and Z. Fang, *Multi-Modality Cascaded Fusion Technology for Autonomous Driving*, en, Feb. 2020. [Online]. Available: https://arxiv.org/abs/2002.03138v1.

[26] I. H. Sarker, "Deep Learning: A Comprehensive Overview on Techniques, Taxonomy, Applications and Research Directions," en, *SN Computer Science*, vol. 2, no. 6, p. 420, Aug. 2021, ISSN: 2661-8907. DOI: 10.1007/s42979-021-00815-1. [Online]. Available: https://doi.org/10.1007/s42979-021-00815-1.

[27] J. Schmidhuber, "Deep learning in neural networks: An overview," *Neural Networks*, vol. 61, pp. 85–117, Jan. 2015, ISSN: 0893-6080. DOI: 10.1016/j.neunet.2014.09.003. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0893608014002135.

[28] S. Saha, *A Comprehensive Guide to Convolutional Neural Networks — the ELI5 way*, en, Nov. 2022. [Online]. Available: https://towardsdatascience.com/a-comprehensive-guide-to-convolutional-neural-networks-the-eli5-way-3bd2b1164a53.

[29] R. Yamashita, M. Nishio, R. K. G. Do, and K. Togashi, "Convolutional neural networks: An overview and application in radiology," en, *Insights into Imaging*, vol. 9, no. 4, pp. 611–629, Aug. 2018, ISSN: 1869-4101. DOI: 10.1007/s13244-018-0639-9. [Online]. Available: https://doi.org/10.1007/s13244-018-0639-9.

[30] R. Jayawardana and T. Sameera Bandaranayake, *ANALYSIS OF OPTIMIZING NEURAL NETWORKS AND ARTIFICIAL INTELLIGENT MODELS FOR GUIDANCE, CONTROL, AND NAVIGATION SYSTEMS*, Apr. 2021.

[31] K. Simonyan and A. Zisserman, *Very Deep Convolutional Networks for Large-Scale Image Recognition*, arXiv:1409.1556 [cs] version: 6, Apr. 2015. DOI: 10.48550/arXiv.1409.1556. [Online]. Available: http://arxiv.org/abs/1409.1556.

[32] S. Ren, K. He, R. Girshick, and J. Sun, *Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks*, en, Jun. 2015. [Online]. Available: https://arxiv.org/abs/1506.01497v3.

[33] Y. Wu, A. Kirillov, F. Massa, W.-Y. Lo, and R. Girshick, *Detectron2*, https://github.com/facebookresearch/detectron2, 2019.

[34] P. Sun, H. Kretzschmar, X. Dotiwalla, *et al.*, *Scalability in Perception for Autonomous Driving: Waymo Open Dataset*, arXiv:1912.04838 [cs, stat], May 2020. DOI: 10.48550/arXiv.1912.04838. [Online]. Available: http://arxiv.org/abs/1912.04838.

[35] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? The KITTI vision benchmark suite," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, ISSN: 1063-6919, Jun. 2012, pp. 3354–3361. DOI: 10.1109/CVPR.2012.6248074. [Online]. Available: https://ieeexplore.ieee.org/document/6248074.

[36] *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale*. [Online]. Available: https://research.google/pubs/an-image-is-worth-16x16-words-transformers-for-image-recognition-at-scale/.

[37] A. Sarkar, *All you need to know about Attention and Transformers In-depth Understanding Part 1*, en, Feb. 2022. [Online]. Available: https://towardsdatascience.com/all-you-need-to-know-about-attention-and-transformers-in-depth-understanding-part-1-552f0b41d021.

[38] D. M. W. Powers, *Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation*, en, Oct. 2020. [Online]. Available: https://arxiv.org/abs/2010.16061v1.

[39] H. Rezatofighi, N. Tsoi, J. Gwak, A. Sadeghian, I. Reid, and S. Savarese, *Generalized Intersection over Union: A Metric and A Loss for Bounding Box Regression*, en, Feb. 2019. [Online]. Available: https://arxiv.org/abs/1902.09630v2.

[40] A. Rosebrock, *Intersection over Union (IoU) for object detection*, en-US, Nov. 2016. [Online]. Available: https://pyimagesearch.com/2016/11/07/intersection-over-union-iou-for-object-detection/.