UNIVERSITY OF OTTAWA

PhD THESIS

Embeddable Temporal Road-User Detection From Radar: A Hybrid CNN-MetaFormer Approach

Author:

Fahed Al Hassanat

Supervisor:

Dr. Robert Laganière

A thesis submitted to the University of Ottawa in fulfillment of the requirements for the degree of Doctorate in Philosophy in Electrical and Computer Engineering

in the

VIVA Research Lab
School of Electrical Engineering and Computer Science

Abstract

Thanks to significant breakthroughs in millimeter-wave radar technology, deep learning architectures, and edge computing capabilities, the pursuit of robust all-weather perception systems for autonomous vehicles has intensified. With various environmental challenges and safety-critical scenarios demanding reliable object detection, researchers are addressing fundamental limitations in sensor-based perception systems. One of the most pressing challenges is achieving accurate road-user detection using automotive radar while maintaining computational efficiency for embedded deployment. Given that modern vehicles require real-time processing to operate on limited computational resources, this thesis presents a hybrid deep learning framework that leverages temporal radar data through a novel CNN-MetaFormer architecture to perform efficient detection and classification of dynamic road users.

We provide a comprehensive analysis of traditional radar processing methods and their evolution toward deep learning approaches, examining both convolutional-based and transformer-based architectures for radar object detection. We also thoroughly investigate temporal modeling strategies and sensor-aware design principles specific to radar data characteristics. Furthermore, we present detailed development of our proposed CompactRADNet architecture that processes sequences of range-azimuth radar frames, introducing the Adaptive Quadratic ReLU (AQR) activation function and radar aware, multipart loss function . Our extensive experiments on the CRUW dataset demonstrate superior performance over state-of-the-art methods. The real-world deployment demonstrates the framework's implementation feasibility, highlighting the impact of hybrid architectural design, temporal sequence optimization, radar-specific adaptations, and the critical balance between detection accuracy and computational efficiency in automotive radar perception systems.

Acknowledgment

In the name of Allah, the Most Gracious, the Most Merciful, and peace and blessings be upon the last o the prophets and messengers.

Praise be to Allah for his endless blessings.

To my father and mother: Thank you for your countless sacrifices and sleepless nights you spent on making the world a better place for me.

To my wife and children: Thank you for being the roots that held me strong to weather all storms.

To my brothers: Thank you for being the whispers of hope that accompanied me throughout my journey .

To my Professor, Robert Laganière: Thank you for holding the torch that lit my path and brought me to where I am today.

To my friend J. B. Aoun "JooJ": Thank you for being there, every step of the way.

Table of Contents

Chapter 1

Introduction

The reliable detection and classification of objects in the surrounding environment is a critical component in the development of advanced driver assistance systems (ADAS) and autonomous vehicles. These systems must accurately perceive and interpret their surroundings to make informed decisions that ensure safety and efficiency. Within this domain, object detection and classification of road users represents a fundamental challenge that has garnered significant attention from both academic researchers and industrial practitioners.

Traditionally, camera based perception systems have dominated the automotive sensing landscape due to their high spatial resolution and rich semantic information. However, these systems suffer from inherent limitations including sensitivity to adverse weather conditions, poor performance in low light environments, and limited depth perception capabilities. These limitations have motivated researchers to explore alternative sensing modalities, among which millimeter wave (mmWave) radar has emerged as a particularly promising technology.

mmWave radar systems operate by transmitting electromagnetic waves in the frequency range of 30 GHz to 300 GHz and measuring the reflections from objects in the environment. The automotive industry has shown particular interest in the 76-81 GHz band for short and medium range applications due to its favorable propagation characteristics. Unlike vision based sensors, radar systems provide direct measurements

of range, relative velocity through Doppler effect, and angular information, making them inherently suitable for the dynamic nature of automotive environments. Furthermore, radar's ability to function robustly under adverse weather conditions such as fog, rain, snow, and dust, as well as in poor lighting conditions, positions it as a critical complementary sensing modality for ensuring the operational continuity of automated driving systems.

Despite these advantages, the utilization of radar for comprehensive object detection and classification tasks presents significant challenges. Radar data typically exhibits lower spatial resolution compared to camera data, especially in the angular dimension. Additionally, radar reflections are highly dependent on the material and geometric properties of objects, leading to complex and sometimes sparse representations. These characteristics make the interpretation of radar data for object detection and classification a difficult task, particularly when attempting to distinguish between different types of road users such as pedestrians, cyclists, and vehicles.

The advent of deep learning techniques has revolutionized numerous computer vision tasks, including object detection. Researchers have increasingly applied these techniques to radar data processing, aiming to overcome the aforementioned challenges and exploit the unique advantages of radar sensing. Various neural network architectures, ranging from convolutional neural networks (CNNs) to more recent transformer based models, have been adapted to process radar data in different forms, such as Range-Azimuth-Doppler (RAD) tensors, bird's eye view (BEV) projections, and point clouds.

1.1 Problem Definition

The fundamental problem addressed in this thesis concerns the accurate and robust detection and classification of road users from automotive mmWave radar data. This problem is characterized by several interrelated challenges:

Limited Angular Resolution: Automotive radar systems typically achieve angular resolutions of 1-5 degrees, which translates to lateral position uncertainties of several meters at long ranges. This limitation makes it difficult to precisely locate objects within their lanes or distinguish closely spaced objects, presenting a significant challenge for safety critical applications.

Sparse and Noisy Measurements: Radar point clouds are inherently sparser than those from LiDAR sensors, with fewer reflection points representing each object. Moreover, radar measurements often include false detections due to multipath reflections, sidelobes, and environmental clutter, complicating the detection process.

Class Imbalance and Weak Signatures: The detection of vulnerable road users (VRUs) such as pedestrians and cyclists is particularly challenging due to their smaller radar cross sections (RCS) compared to vehicles. The RCS of a pedestrian can be 100-1000 times smaller than that of a passenger car, resulting in weaker reflections that are difficult to distinguish from noise. Additionally, typical driving scenarios exhibit severe class imbalance, with many more vehicle instances than VRUs.

Real-time Processing Requirements: Automotive applications demand real-time performance with limited computational resources. This constraint necessitates efficient processing algorithms that can operate within strict latency bounds while maintaining high detection accuracy.

Cross Modal Supervision Challenge: The non-intuitive nature of radar data makes direct annotation extremely challenging, even for experts. This creates a bottleneck in

generating large scale labeled datasets necessary for training deep learning models, motivating the need for innovative supervision strategies.

1.2 Contributions

This thesis makes several significant contributions to the field of radar based object detection for automotive applications:

1. Novel Hybrid Architecture

We propose CompactRADNet, a novel deep learning architecture that synergistically combines convolutional operations with transformer based attention mechanisms. This architecture effectively addresses the unique characteristics of radar data by:

- Employing specialized 3D convolutional stems for processing multichannel radar data
- Integrating Metaformer-based blocks for capturing long range dependencies with superior computational efficiency
- Incorporating multiscale feature fusion through Feature Pyramid Networks (FPN)
- Implementing auxiliary detection heads for improved gradient flow and feature learning
- Our experimentations demonstrate that our proposed architecture achieves SOTA.

2. Advanced Loss Function Design

We developed MultiClass Detection Loss, a comprehensive multicomponent loss function that addresses the challenges of radar based detection through:

- Multiclass focal loss with class specific gamma values to handle class imbalance

- Regression loss for precise object localization
- Auxiliary losses for enhanced training stability

3. Adaptive Quadratic ReLU (AQR) Activation Function

We introduce the Adaptive Quadratic ReLU (AQR), a specialized activation function designed specifically for radar signal processing applications. Traditional activation functions, while effective for natural image processing, fail to address the unique characteristics of radar signals, particularly their sparse nature and high dynamic range. AQR incorporates an adaptive gating mechanism that provides input-dependent modulation, enabling robust processing of radar signals with varying strengths.

Our proposed mathematical formulation combines the beneficial quadratic characteristics with a learnable sigmoid gate that adapts activation strength based on input magnitude. This design enables automatic adjustment between noise suppression for weak signals and amplification for strong target returns, addressing a fundamental challenge in radar-based detection systems.

Comprehensive experimental validation demonstrates AQR's effectiveness across multiple evaluation criteria:

- In Signal Processing Performance, AQR demonstrates improvement in Signal-to-Noise Ratio for target detection scenarios compared to conventional activation functions, translating directly to enhanced detection capabilities for weak targets in challenging environmental conditions
- In terms of Architecture Integration Benefits and within our temporal processing framework, AQR delivers measurable performance improvements while maintaining computational efficiency through minimal parameter overhead
- For Radar-Specific Optimization, the adaptive characteristics prove particularly valuable for automotive radar applications where signal conditions vary dramatically across different driving scenarios, from highway environments with long-range targets to urban settings with complex clutter patterns

The development of AQR represents a significant advancement in activation function design for signal processing applications, demonstrating that domain-specific architectural innovations can yield substantial performance improvements while maintaining the computational efficiency required for real-time automotive deployment.

4. End-to-End Real-World Deployment and Validation

We demonstrate the practical viability of our CompactRADNet architecture through comprehensive real-world deployment on an instrumented vehicle, bridging the critical gap between academic research and production automotive systems. This contribution encompasses the complete development lifecycle from raw ADC data processing to real-time object detection in dynamic driving environments, validating our approach under realistic operational constraints.

- Complete Signal Processing Pipeline Implementation: We developed and deployed a comprehensive processing pipeline that transforms raw ADC data from the RFBeam V-MD3 61 GHz radar into real-time object detections. This implementation includes optimized FFT processing achieving 8-12ms latency, custom range-azimuth map generation, and temporal sequence management adapted for 7-8 Hz sensor update rates. The pipeline demonstrates the feasibility of deploying sophisticated deep learning models within the stringent computational constraints of automotive embedded systems.

- Domain Adaptation and Cross-Frequency Validation: Our deployment successfully addresses significant domain shift challenges, adapting from the CRUW dataset's 77 GHz AWR1843 configuration to the 61 GHz V-MD3 platform. This cross-frequency deployment provides initial evidence for the generalizability of our architectural innovations, with comprehensive validation pending completion of platform-specific optimization.

- Instrumented Vehicle Integration: Our instrumented vehicle platform, equipped with the Spectra2 automotive-grade edge AI system and dual RTX QUADRO A4000 GPUs, demonstrates practical deployment considerations including sensor mounting optimization, environmental robustness (IP-65 rated radar enclosure), and integration with automotive power systems. This platform validates the complete system architecture from hardware integration through software deployment.

- Production-Ready Implementation Insights: The deployment reveals critical insights for transitioning academic research to production systems, including the importance of co-designing algorithmic and system-level components, the effectiveness of flexible temporal configurations for varying sensor update rates, and the promising initial behavior of pooling-based temporal fusion across different radar platforms. These insights contribute valuable knowledge to the community regarding practical deployment challenges and solutions.

The real-world deployment implementation demonstrates that careful architectural design has the potential to achieve strong perception capabilities while respecting the resource limitations and reliability requirements of safety-critical automotive applications. This contribution demonstrates that the gap between laboratory performance and real-world deployment can be successfully bridged through careful system engineering and validation methodologies.

1.3 Thesis Structure

This thesis is organized into nine chapters that systematically present the research from theoretical foundations to practical implementation and evaluation:

Chapter 2: Background establishes the foundational elements necessary for understanding radar based object detection. It covers neural network fundamentals, object detection principles, and mmWave radar sensing technologies. The chapter provides the theoretical basis for the subsequent technical discussions.

Chapter 3: Literature Review provides a comprehensive examination of existing approaches to radar based object detection. It analyzes traditional radar processing methods, CNN based approaches, transformer architectures, and fusion techniques. The chapter identifies research gaps and positions our contributions within the broader research landscape.

Chapter 4: Methodology details the research methodology, including the datasets used for experimentation (primarily the CRUW dataset), evaluation metrics with special emphasis on Object Location Similarity (OLS), and the experimental setup. This chapter establishes the framework for evaluating our proposed approach.

Chapter 5: Solution Architecture presents the detailed design of our CompactRADNet architecture. It describes the motivation behind architectural choices, the integration of convolutional and transformer components, the loss function design, and implementation details. This chapter along with chapter 6 constitutes the core technical contribution of the thesis.

Chapter 6: Temporal Approach presents the detailed design of our temporal extension aspect of CompactRADNet architecture. It describes the motivation behind architectural choices, the integration of transformer and metaformer based components and implementation details. This chapter, along with chapter 5, constitutes the core technical contribution of the thesis.

Chapter 7: Experiments provides comprehensive experimental results and analysis. It includes ablation studies, comparative evaluations with baseline methods, performance analysis across different scenarios, and computational efficiency assessments. The

chapter demonstrates the effectiveness of our approach through rigorous empirical evaluation.

Chapter 8: Real World Deployment demonstrates our implementation and deployment of a complete end-to-end pipeline to capture data from a vehicle mounted ADC-capable radar sensor and moving the data through the parsing and restructuring, FFT processing, temporal construction, CompactRADNet model consumption and finally detections displayed on the vehicle's incabin screen.

Chapter 9: Conclusion and Future Work summarizes the key findings, discusses the implications of our work for the field of autonomous driving, acknowledges limitations, and outlines promising directions for future research. This chapter provides closure while opening avenues for continued investigation.

Through this structured exploration, the thesis aims to advance the state of the art in radar based object detection, contributing to the development of more robust and reliable perception systems for autonomous vehicles. The research presented herein demonstrates that careful architectural design, combined with appropriate loss functions and training strategies, can significantly improve the detection of all road users, particularly vulnerable ones, using automotive radar data.

Summary

In this chapter we introduced the fundamental challenges of object detection in autonomous driving systems, with particular emphasis on the limitations of camera-based perception under adverse weather conditions. The problem definition established the critical need for radar-based solutions that can maintain robust detection capabilities across varying environmental scenarios while meeting the strict computational constraints of embedded automotive platforms. The chapter outlined the four primary contributions of this thesis: the development of CompactRADNet, a novel hybrid CNN-

MetaFormer architecture; the introduction of a multi-component loss function design; the proposal of the Adaptive Quadratic ReLU (AQR) activation function specifically designed for radar signal processing; and the demonstration of end-to-end real-world deployment on an instrumented vehicle. The thesis structure was presented, outlining the systematic progression from theoretical foundations through practical implementation and evaluation across nine chapters.

Chapter 2

Background

2.1. Neural Networks

2.1.1 Fundamental Concepts and Architectures

Neural networks represent a class of machine learning algorithms inspired by the structure and function of the human brain [13]. At their core, these networks consist of interconnected processing units, or neurons, organized in layers. Each neuron receives input signals, applies a weighted transformation, and passes the result through a nonlinear activation function to produce an output [14]. The fundamental building block of a neural network is the artificial neuron, which can be mathematically represented as:

$$y = f(\sum_{i=1}^{n} w_i x_i + b)$$

where x_i are the inputs, w_i are the corresponding weights, b is a bias term, and f is a nonlinear activation function [15]. Common activation functions include the sigmoid, hyperbolic tangent (tanh), and more recently, the Rectified Linear Unit (ReLU) and its variants [16].

Recent developments in activation function design have focused on domain-specific adaptations that address the unique characteristics of specialized data modalities. For radar signal processing applications, the sparse nature of radar returns and high dynamic range measurements necessitate activation functions that can adaptively respond to varying signal magnitudes. This has led to innovations such as quadratic activation functions that naturally amplify strong signals while suppressing weak ones, and gating mechanisms that provide input-dependent modulation capabilities[134].

The architecture of a neural network is defined by the arrangement of neurons into layers and the pattern of connections between these layers. The most basic form is the feedforward neural network, where information flows in one direction from the input layer through one or more hidden layers to the output layer [17]. Each layer in this architecture is fully connected to the subsequent layer, with no connections between neurons within the same layer or feedback connections to previous layers. This structure, also known as a Multi Layer Perceptron (MLP), forms the foundation for more complex neural network architectures [18].

The training of neural networks typically involves the use of gradient based optimization algorithms, such as Stochastic Gradient Descent (SGD) and its variants, to minimize a defined loss function [19]. The backpropagation algorithm efficiently computes the gradient of the loss function with respect to the network parameters by applying the chain rule of calculus, enabling the update of these parameters in the direction that reduces the loss [20]. This process allows neural networks to learn complex mappings from inputs to outputs by adjusting their internal parameters based on the training data.

2.1.2 Evolution of Neural Networks in Computer Vision

The application of neural networks to computer vision tasks has evolved significantly over the past decades, marked by several breakthrough developments. Early attempts at using neural networks for image recognition faced limitations due to computational constraints and the lack of efficient training methods for deep architectures [21]. The introduction of Convolutional Neural Networks (CNNs) by LeCun et al. [22] represented

a significant advancement, as these networks leveraged the principles of local connectivity and weight sharing to efficiently process grid structured data such as images.

The resurgence of neural networks in computer vision can be attributed to the seminal work of Krizhevsky et al. [23] with AlexNet, which demonstrated unprecedented performance on the ImageNet classification challenge in 2012. This success catalyzed rapid developments in CNN architectures, including VGGNet [24], which explored the impact of network depth; GoogLeNet [25], which introduced the inception module to capture multiscale features; and ResNet [26], which addressed the vanishing gradient problem through residual connections, enabling the training of substantially deeper networks.

Beyond image classification, neural networks have been successfully applied to various computer vision tasks, including object detection, semantic segmentation, and instance segmentation [27]. For object detection, region based approaches such as R-CNN [28] and its faster variants [28, 30] combine region proposals with CNN based feature extraction. Single shot detectors like YOLO [31] and SSD [32] have emerged as efficient alternatives, performing detection in a single forward pass of the network. Semantic segmentation networks such as Fully Convolutional Networks (FCN) [33] and U-Net [34] have enabled pixel level classification, while instance segmentation approaches like Mask R-CNN [35] extend object detection to include pixel level masks for each detected object.

More recently, attention mechanisms and transformer architectures have gained prominence in computer vision [36]. Vision Transformer (ViT) [37] demonstrated that pure transformer architectures can achieve competitive performance on image classification tasks by treating an image as a sequence of patches. DETR [38] extended this approach to object detection by formulating it as a set prediction problem, eliminating the need for hand designed components like anchor boxes and non-maximum suppression.

2.1.3 Deep Learning Architectures Relevant to Radar Data Processing

The application of deep learning architectures to radar data processing presents unique challenges and opportunities due to the distinctive nature of radar data compared to images. Radar data can be represented in various forms, including Range-Azimuth-Doppler (RAD) tensors, bird's eye view (BEV) projections, and point clouds, each requiring specific architectural considerations [39].

Convolutional Neural Networks have been widely adapted for radar data processing due to their ability to extract spatial and spectral features from grid structured representations [40]. For RAD tensors, 3D CNNs have been employed to process the three dimensional nature of the data, capturing correlations across range, Doppler, and azimuth dimensions simultaneously, as in the work by Pallfy et al. [12], a 3D CNN architecture was proposed to process radar data cubes for automotive applications, demonstrating improved detection performance compared to 2D projections.

For radar point cloud processing, architectures inspired by PointNet [41] and PointNet++ [42] have been explored. These networks operate directly on unordered point sets, making them suitable for the sparse and irregular nature of radar point clouds [43]. The work by Wang et al.

Recurrent Neural Networks (RNNs), particularly Long Short Term Memory (LSTM) [44] and Gated Recurrent Unit (GRU) [45] variants, have been applied to radar data to model temporal dependencies across consecutive frames [46]. This temporal modeling is crucial for tracking moving objects and extracting motion patterns that can aid in classification. The work by Dong et al. [47] utilized an LSTM based approach to integrate temporal information from radar sequences, demonstrating improved classification of dynamic road users.

More recently, transformer based architectures have been adapted for radar data processing, leveraging their strong capability to model long range dependencies and capture global context [11]. The self-attention mechanism in transformers allows for flexible and adaptive feature aggregation, which is particularly valuable for the varying density and quality of radar data across different regions [49]. Mask-RadarNet [50] and

T-RODNet [48] represent notable examples of transformer based approaches for radar object detection, incorporating spatiotemporal context and multiscale feature representation.

Hybrid architectures that combine different neural network components have also shown promising results for radar data processing. For instance, RODNet [51] employs a CNN based backbone for feature extraction followed by an RNN for temporal modeling, while TransRAD [52] integrates convolutional layers with transformer blocks to leverage both local and global feature representations.

The selection and adaptation of deep learning architectures for radar data processing must consider the specific characteristics of the data, such as its dimensionality, sparsity, and the nature of the information it captures (range, velocity, angular position). Additionally, the computational constraints of automotive systems necessitate a balance between model complexity and inference efficiency. As research in this field continues to evolve, there is a growing trend towards end to end architectures that can process raw radar signals, potentially bypassing traditional radar signal processing chains and enabling more integrated and optimized detection systems [53, 54].

2.2 Object Detection Fundamentals

2.2.1 Traditional Object Detection Approaches

Object detection represents a fundamental computer vision task that involves both localizing objects within an image and classifying them into predefined categories [55]. Prior to the deep learning revolution, traditional object detection approaches relied on hand crafted features and multi stage pipelines to identify objects of interest [56].

Among the most influential traditional methods was the Viola-Jones detector [57], which employed Haar-like features and AdaBoost for rapid face detection. This approach utilized integral images for efficient feature computation and a cascade of classifiers to achieve real-time performance while maintaining acceptable accuracy. Despite its

success in face detection, the Viola-Jones framework exhibited limitations when applied to more general object categories due to its reliance on rigid templates [58].

Another significant approach was the Histogram of Oriented Gradients (HOG) combined with Support Vector Machines (SVM) for pedestrian detection, as proposed by Dalal and Triggs [59]. HOG captures local gradient orientation distributions, providing robustness to illumination changes and small deformations. The HOG-SVM detector achieved state of the art performance for its time and established a template for subsequent feature based detection methods [60].

Deformable Part Models (DPM), introduced by Felzenszwalb et al. [61], extended the HOG-SVM approach by modeling objects as collections of parts arranged in a deformable configuration. This method addressed the limitations of rigid templates by explicitly accounting for variations in object appearance and pose. DPM represented the pinnacle of traditional object detection approaches, achieving leading performance on benchmark datasets before being surpassed by deep learning methods [62].

Traditional object detection methods typically followed a sliding window paradigm, where a classifier was applied at multiple positions and scales across an image to identify object instances [63]. This approach, while conceptually straightforward, suffered from high computational complexity and limited representational capacity. Additionally, these methods struggled with variations in object appearance, occlusions, and diverse backgrounds, leading to a significant performance gap compared to human level recognition [64].

2.2.2 Deep Learning based Object Detection Frameworks

The transition from traditional to deep learning based object detection frameworks marked a paradigm shift in the field, characterized by substantial improvements in both accuracy and generalization capabilities [65]. Deep learning approaches replaced hand crafted features with learned representations, enabling more effective modeling of complex visual patterns and hierarchical object structures [66].

Deep learning based object detection frameworks can be broadly categorized into two main approaches: two stage detectors and single stage detectors [27]. Two stage detectors first generate region proposals that potentially contain objects and then classify and refine these proposals in a second stage. In contrast, single stage detectors perform localization and classification in a single forward pass of the network, typically resulting in faster inference at the cost of some accuracy [67].

Region based Convolutional Neural Networks (R-CNN) [28] pioneered the two stage approach by combining region proposals with CNN based feature extraction. The original R-CNN framework used selective search [68] to generate region proposals, followed by a CNN to extract features from each proposal, and finally SVM classifiers for object categorization. While R-CNN demonstrated significant performance improvements over traditional methods, it suffered from computational inefficiency due to the redundant feature extraction for overlapping proposals [69].

Fast R-CNN [29] addressed this limitation by processing the entire image through a CNN to generate a feature map, from which regions of interest (RoIs) were extracted using a RoI pooling layer. This approach significantly reduced computation time by sharing features across proposals. Faster R-CNN [30] further improved efficiency by introducing the Region Proposal Network (RPN), which generates proposals directly from CNN features, creating an end to end trainable detection framework [70].

Single stage detectors emerged as an alternative approach, with YOLO (You Only Look Once) [31] and SSD (Single Shot MultiBox Detector) [32] as prominent examples. YOLO divided the image into a grid and predicted bounding boxes and class probabilities for each grid cell directly, treating detection as a regression problem. SSD extended this concept by making predictions at multiple scales using feature maps from different network layers, improving detection accuracy for objects of varying sizes [71].

The evolution of deep learning based object detection has been driven by the pursuit of higher accuracy, faster inference, and greater generalization capabilities [72]. Recent developments have explored anchor free approaches [73], which eliminate the need for

predefined anchor boxes; feature pyramid networks [74], which enhance multiscale feature representation; and attention mechanisms [75], which capture long range dependencies and context information.

2.2.3 R-CNN Family, YOLO, SSD, and Their Variants

The R-CNN family of detectors has evolved significantly since its inception, addressing various limitations and improving performance aspects [76]. Mask R-CNN [35] extended Faster R-CNN by adding a branch for predicting segmentation masks in parallel with bounding box recognition, enabling instance segmentation capabilities. Cascade R-CNN [77] introduced a multistage refinement process where a sequence of detectors trained with increasing IoU thresholds progressively improved localization accuracy. The most recent iterations, such as Sparse R-CNN [78], have explored end-to-end object detection with sparse representations, reducing the reliance on dense anchors and NMS post processing.

YOLO has undergone several iterations, each addressing limitations of its predecessors [79]. YOLOv2 [80] incorporated batch normalization, anchor boxes, and dimension clusters to improve stability and accuracy. YOLOv3 [81] introduced a feature pyramid network structure and multiple prediction scales to enhance performance on small objects. YOLOv4 [82] and subsequent versions have integrated various architectural innovations such as CSPNet (Cross-Stage Partial Network), CIOU (Complete IoU) loss, and mosaic data augmentation, pushing the boundaries of speed/accuracy tradeoffs in object detection.

SSD variants have similarly focused on improving detection performance across different scales and reducing computational requirements [83]. DSSD (Deconvolutional Single Shot Detector) [71] enhanced SSD by incorporating deconvolutional layers to provide additional context for prediction. RetinaNet [67] addressed the class imbalance problem inherent in dense detectors by introducing focal loss, which downweights the contribution of easy examples during training. Other developments include FSSD (Feature-Fused

SSD) [83], which improves feature fusion across different scales, and SSDLite [84], which employs depthwise separable convolutions to reduce computational complexity.

Beyond these established frameworks, recent research has explored transformer based object detection approaches [72]. DETR (DEtection TRansformer) [38] formulated object detection as a direct set prediction problem, using a transformer encoder/decoder architecture to predict a set of objects in parallel. This approach eliminates the need for many hand designed components like anchor generation and non-maximum suppression. Subsequent works such as Deformable DETR [85] have addressed the slow convergence and limited feature sampling of the original DETR, while maintaining its clean detection pipeline.

2.2.4 Detection Heads and Loss Functions

Detection heads refer to the network components responsible for predicting object locations and categories based on extracted features [86]. In two stage detectors like Faster R-CNN, the detection head typically consists of fully connected layers that process RoI-pooled features to output class probabilities and bounding box coordinates [87]. Single stage detectors often employ convolutional heads that make predictions at multiple spatial locations and scales, with each location responsible for detecting objects that match certain anchor configurations [88].

Various architectural designs have been explored for detection heads to improve performance [89]. For instance, the Cascade R-CNN [77] employs a sequence of detection heads with progressively increasing IoU thresholds to refine predictions iteratively. The Double-Head method [76] separates classification and localization tasks into different head structures, with fully connected layers for classification and convolution layers for localization, leveraging the strengths of each architecture for their respective tasks.

Loss functions play a crucial role in training object detectors by defining the optimization objective [90]. For classification, cross-entropy loss or its variants are commonly used to

measure the discrepancy between predicted class probabilities and ground truth labels [91]. For bounding box regression, several loss functions have been proposed, each with different properties and optimization characteristics [92].

The L1 and L2 losses are straightforward choices for bounding box regression, measuring the absolute or squared differences between predicted and ground truth coordinates [93]. However, these losses do not directly optimize for the Intersection over Union (IoU) metric commonly used to evaluate detection performance. To address this limitation, IoU loss [94] directly optimizes the overlap between predicted and ground truth boxes. Further refinements include GIoU (Generalized IoU) [90], DIoU (Distance IoU) [92], and CIoU (Complete IoU) [95] losses, which incorporate additional geometric information to guide the optimization process.

Focal loss [67], introduced in RetinaNet, addresses the class imbalance problem in dense detectors by downweighting the contribution of easy examples during training. This allows the model to focus on difficult examples, improving overall performance. Similarly, balanced L1 loss [96] modifies the standard L1 loss to assign higher weights to inliers, promoting more precise localization.

For anchor free detectors, alternative formulations have been explored [97]. FCOS (Fully Convolutional One-Stage Object Detection) represents bounding boxes as distances from each pixel to the four sides of the box, employing a combination of classification, centeredness, and regression losses. CenterNet [98] models objects as points and uses keypoint estimation techniques, with a penalty reduced focal loss for classification and L1 loss for size regression.

The design of detection heads and loss functions continues to evolve, with recent work exploring adaptive loss weighting [99], uncertainty estimation [100], and task specific optimization strategies [101]. These developments aim to address specific challenges in object detection, such as scale variation, occlusion handling, and precise localization, contributing to the overall advancement of the field.

2.3 mmWave Radar Sensing

2.3.1 FMCW Radar Principles

Frequency Modulated Continuous Wave (FMCW) radar represents the predominant technology employed in automotive mmWave radar systems [102]. Unlike pulsed radar systems that transmit short, high power pulses and measure the time delay of the return signal, FMCW radar continuously transmits a frequency modulated signal, typically in the form of a linear frequency sweep known as a chirp [103]. This approach offers several advantages for automotive applications, including lower peak power requirements, better range resolution, and the ability to simultaneously measure both range and relative velocity [5].

The fundamental operating principle of FMCW radar involves transmitting a signal whose frequency changes linearly with time over a bandwidth B during a chirp duration T_c [104]. The transmitted signal can be expressed as:

$$s_{tx}(t) = A_{tx}cos(2\pi(f_c t + \frac{\alpha}{2}t^2) + \phi_0)$$

where f_c is the carrier frequency, $\alpha = B/T_c$ is the chirp rate, and ϕ_0 is the initial phase [105]. When this signal reflects off an object at range R and returns to the receiver after a time delay $\tau = 2R/c$ (where c is the speed of light), the received signal becomes:

$$s_{rx}(t) = A_{rx}cos(2\pi(f_c(t-\tau) + \frac{\alpha}{2}(t-\tau)^2) + \phi_0)$$

By mixing the transmitted and received signals, a beat signal is produced whose frequency is proportional to the range of the target [106]:

$$f_b = \alpha\tau = \frac{2\alpha R}{c} = \frac{2BR}{cT_c}$$

For a moving target with a relative radial velocity v_r, the Doppler effect induces an additional frequency shift [107]:

$$f_d = 2v_r f_c/c$$

In a single chirp, this manifests as a phase shift, but across multiple chirps, it allows for the measurement of velocity through Fourier analysis [108].

mmWave radar systems operating in the 76-81 GHz frequency band offer significant advantages for automotive sensing applications [4]. The millimeter wavelength enables the use of compact antenna arrays, facilitating integration into vehicle structures. Additionally, the wide available bandwidth (up to 4 GHz) allows for high range resolution, typically on the order of centimeters, which is crucial for distinguishing closely spaced objects in complex traffic scenarios [109].

2.3.2 Signal Processing Chain (ADC, FFT, etc.)

The radar signal processing chain transforms raw received signals into interpretable data representations from which object detection and classification can be performed [110]. This process involves several key stages, beginning with analog-to-digital conversion and culminating in detection algorithms that identify objects in the scene.

The analog frontend of the radar system amplifies and filters the received signal before it is digitized by an Analog-to-Digital Converter (ADC) [111]. The sampling rate of the ADC must satisfy the Nyquist criterion to avoid aliasing, with typical automotive radar systems employing sampling rates in the range of 10-40 MHz [112]. The digitized samples from each chirp form the raw data for subsequent processing.

The first major processing step typically involves applying a Fast Fourier Transform (FFT) along the samples within each chirp, commonly referred to as the Range-FFT [7]. This transform converts the time domain samples into the frequency domain, where peaks correspond to the beat frequencies of targets at different ranges. The range resolution ΔR is determined by the bandwidth B of the chirp:

$$\Delta R = \frac{c}{2B}$$

For multiple input multiple output (MIMO) radar configurations, which employ multiple transmit and receive antennas, spatial processing techniques can be applied to enhance angular resolution [113]. This typically involves applying another FFT along the spatial dimension, known as the Angle-FFT, which processes the phase differences between different receive channels to estimate the angular position of targets [114]. The angular resolution $\Delta\Theta$ is related to the aperture size D and wavelength λ:

$$\Delta\Theta \approx \frac{\lambda}{D}$$

To measure velocity, a sequence of chirps is transmitted, and an FFT is performed across the same range bin in consecutive chirps, known as the Doppler-FFT [115]. This process reveals the Doppler frequency shifts associated with moving targets, with the velocity resolution Δv determined by the chirp repetition interval T_{ri} and the number of chirps N_c :

$$\Delta v = \frac{\lambda}{2T_{ri}N_c}$$

Following these transforms, detection algorithms such as Constant False Alarm Rate (CFAR) are applied to identify peaks in the transformed data that correspond to potential targets [106]. CFAR operates by adaptively setting a detection threshold based on the local noise level, maintaining a constant probability of false alarm across varying noise

conditions [116]. Common variants include Cell-Averaging CFAR (CA-CFAR), Ordered Statistics CFAR (OS-CFAR), and their two dimensional extensions for processing Range-Doppler maps [117].

The final stages of the processing chain typically involve clustering detected points to form object hypotheses, tracking these objects across frames, and classifying them into relevant categories [118]. These higher level processing tasks increasingly leverage machine learning techniques, particularly deep neural networks, to improve detection and classification performance [9].

2.3.3 Radar Data Representations (Range-Azimuth-Doppler cubes, point clouds)

Radar data can be represented in various forms, each offering different tradeoffs in terms of information content, computational requirements, and suitability for subsequent processing tasks . The choice of representation significantly impacts the design and performance of deep learning models for radar based object detection .

2.3.3.1 Traditional Processing Pipeline and Information Loss

The traditional radar processing pipeline, while computationally efficient for classical detection algorithms, introduces substantial information loss that limits the performance ceiling of detection systems. As described in Section 2.1.3.2, the conventional pipeline progresses from ADC samples through FFT processing to CFAR detection, ultimately producing sparse point clouds. Each stage of this pipeline makes hard decisions that irreversibly discard potentially valuable information.

The CFAR detection stage represents the most significant information bottleneck, typically retaining very little of the original signal energy . By applying adaptive thresholding based on local noise statistics, CFAR eliminates weak returns that may correspond to valid targets, particularly in scenarios with low signal-to-noise ratios or distributed targets like pedestrians . The subsequent clustering and tracking stages further compress the data, reducing complex extended targets to single point

representations with estimated bounding boxes . This aggressive data reduction, while enabling real-time processing on resource-constrained embedded systems, fundamentally limits the information available for higher-level reasoning about object properties, behaviors, and interactions.

2.3.3.2 Range-Azimuth-Doppler Cube Construction

The Range-Azimuth-Doppler (RAD) cube preserves the complete information content from radar signal processing before detection decisions, providing a dense three-dimensional representation of the radar scene . Construction of the RAD cube involves sequential application of Fourier transforms along three orthogonal dimensions, each extracting different physical information from the received signals.

Starting from the digitized ADC samples s[n,m,k] where n indexes fast-time samples within a chirp, m indexes chirps within a frame, and k indexes receive channels, the Range-FFT processes each chirp independently:

$$R[r, m, k] = \mathrm{FFT}_n\{w_r[n] \cdot s[n, m, k]\}$$

where $w_r[n]$ represents the range window function (typically Hamming or Hann) applied to reduce sidelobes. This transform converts time-domain samples to range bins, with each bin corresponding to a physical distance determined by the radar's bandwidth and sampling rate as established in Section 2.1.3.1.

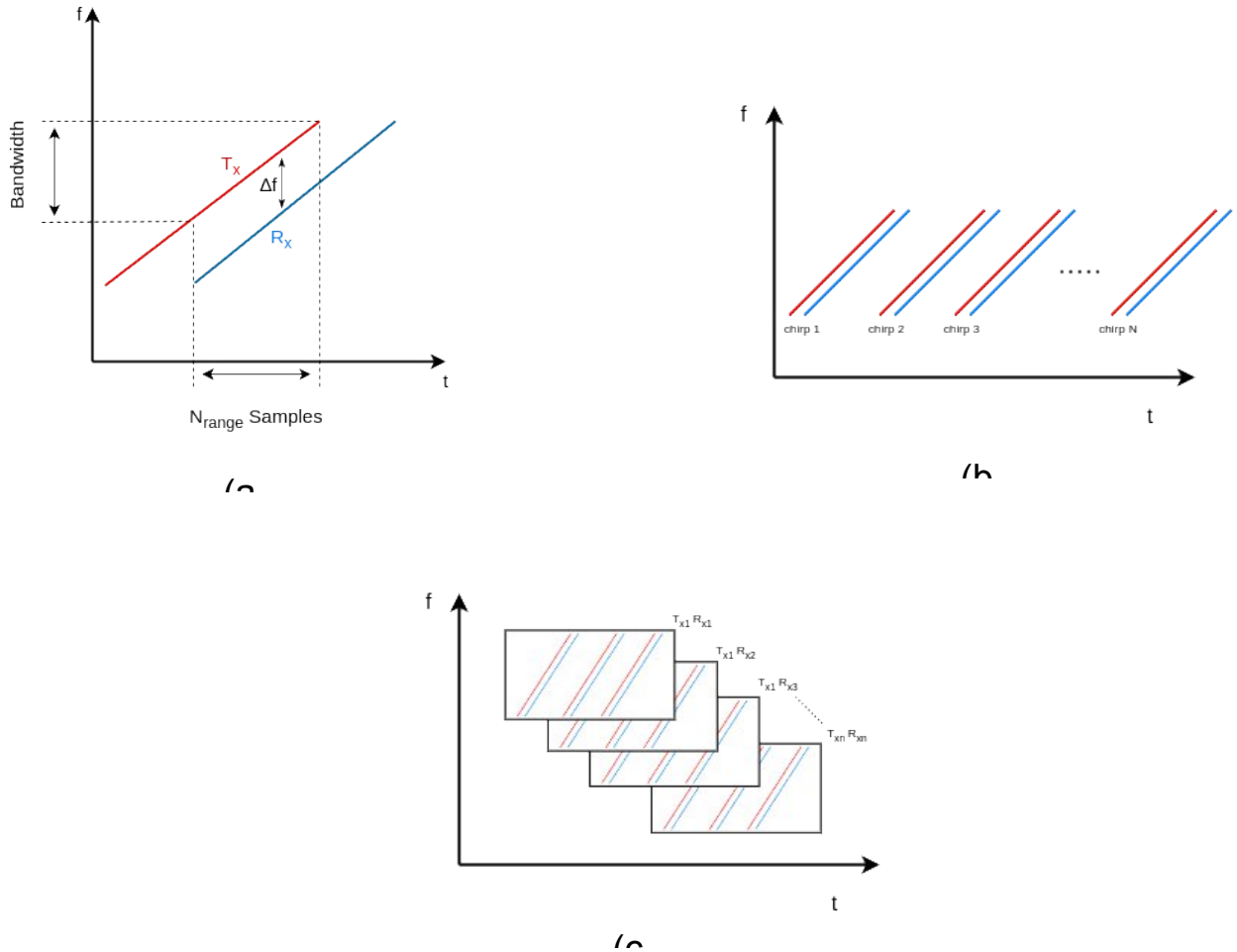


Figure 2.1: (a) FMCW chirp sequence showing fast-time sampling (source of range) and (b) slow-time stacking (source of Doppler) with (c) virtual array arrangement (source of azimuth)

The Doppler-FFT subsequently processes across the slow-time dimension for each range bin:

$$RD[r, d, k] = \mathbf{FFT}_m\{w_d[m] \cdot R[r, m, k]\}$$

where $w_d[m]$ is the Doppler window function. This transform extracts velocity information from the phase progression across chirps, with the velocity resolution dependent on the frame duration and number of chirps as described in Section 2.1.3.2.

For MIMO configurations with multiple transmit and receive antennas, the virtual array concept enables angular resolution. The received signals from N_{TX} transmitters and N_{RX} receivers create $N_{TX} \times N_{RX}$ virtual channels. The Angle-FFT processes across this virtual array dimension:

$$RAD[r, a, d] = \mathbf{FFT}_k\{w_a[k] \cdot RD[r, d, k]\}$$

where the virtual array elements k are arranged according to their spatial positions, and $w_a[k]$ represents the angular window function. Zero-padding is commonly applied before the Angle-FFT to increase the angular sampling density via interpolation, in order to produce a smoother range-azimuth representation that aids in peak localization and deep-learning-based processing, without alteration to the sensor's physical angular.

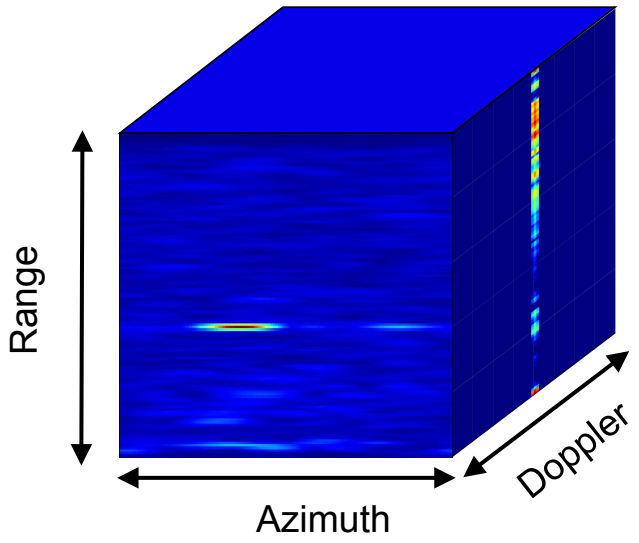


Figure 2.2: 3D visualization of RAD cube structure showing range, azimuth, and Doppler dimensions

2.3.3.3 RAD Cube Structure and Physical Interpretation

The resulting RAD cube $R \in C^{Nr \times Na \times Nd}$ represents a complete volumetric snapshot of the radar scene, where each complex-valued voxel encodes both magnitude and phase information. The magnitude $|R[r,a,d]|$ indicates the signal strength from targets at specific range r, azimuth angle a, and radial velocity d. The phase $\angle R[r,a,d]$ contains additional information about target micro-motion and can be exploited for advanced classification tasks.

Each dimension of the RAD cube maps to physical quantities with specific resolutions and ambiguity limits. The range dimension spans from zero to the maximum unambiguous range $R_{max}=c/(2 \cdot PRF)$, where PRF is the pulse repetition frequency. The azimuth dimension covers the radar's field of view, typically ±45° to ±60° for automotive applications, with resolution limited by the array aperture. The Doppler dimension extends from $-v_{max}$ to $+v_{max}$, where $v_{max}=\lambda/(4 \cdot T_{frame})$ represents the maximum unambiguous velocity.

The information density of the RAD cube varies significantly across dimensions and regions. Strong point targets like vehicles produce concentrated energy peaks, while distributed targets like pedestrians create more diffuse signatures spread across multiple voxels . Environmental clutter, multipath reflections, and interference manifest as

structured noise patterns that deep learning models can learn to distinguish from valid targets.

2.3.3.4 CRUW Dataset RAD Cube Implementation

The CRUW dataset implements a specific RAD cube configuration optimized for automotive scenarios using a 77 GHz FMCW radar with 2TX/4RX MIMO architecture [51]. The resulting 8-element virtual array undergoes processing to generate RAD cubes with dimensions of 128 (range) × 128 (azimuth) × 256 (Doppler), though typically only selected chirps (0, 64, 128, 192) are provided to reduce data volume while maintaining temporal information.

The range dimension covers 0-30 meters with approximately 0.39-meter resolution per bin, suitable for urban and highway scenarios. The azimuth dimension spans ±60° over 128 bins. The Doppler dimension represents velocities from -20 to +20 m/s (approximately ±72 km/h), with 0.156 m/s velocity resolution per bin. These specifications balance detection performance with computational tractability for real-time processing.

For practical implementation, the CRUW dataset often provides 2D projections of the full RAD cube. Range-Azimuth (RA) maps are generated by max-pooling or averaging across the Doppler dimension:

$$RA[r, a] = \max_{d} |R[r, a, d]|^2$$

These RA maps, provided as 128×128 grids after power computation and logarithmic scaling, serve as the primary input for many detection algorithms while preserving most spatial information critical for object localization .

2.3.3.5 Advantages for Deep Learning Applications

The RAD cube representation offers several fundamental advantages for deep learning based object detection compared to traditional point cloud representations . First, by preserving pre-detection information, neural networks can learn task-specific detection

thresholds that adapt to different scenarios and object types, rather than relying on fixed CFAR parameters . This adaptability proves particularly valuable for detecting weak targets in cluttered environments where traditional detection would fail.

Second, the dense volumetric structure of RAD cubes aligns naturally with convolutional neural network architectures developed for image processing . Standard 2D convolutions can process RA or RD projections, while 3D convolutions can directly operate on the full RAD volume, extracting hierarchical features that capture both local patterns and global context . The regular grid structure eliminates the need for specialized point cloud processing operations like PointNet , enabling the use of mature, optimized deep learning frameworks.

Third, the preserved phase information and weak signals in RAD cubes enable advanced capabilities beyond simple detection. Networks can learn to associate subtle Doppler signatures with specific object classes, distinguish between different pedestrian activities based on micro-Doppler patterns, and even predict future motion from current velocity distributions . These capabilities emerge naturally from end-to-end learning on RAD cube representations without explicit feature engineering.

The tradeoff for these advantages lies primarily in computational and memory requirements. Processing full RAD cubes demands substantially more resources than sparse point clouds, with memory requirements scaling as $O(N_r \times N_a \times N_d)$ compared to $O(N_{points})$ for point representations. However, modern GPU architectures and optimized deep learning frameworks increasingly mitigate these costs, making RAD cube processing viable for real-time automotive applications .

2.3.4 Advantages and Limitations in Automotive Environments

mmWave radar systems offer several distinct advantages that make them invaluable components of automotive sensing suites. Their robustness to adverse environmental conditions represents perhaps the most significant advantage, as radar performance

remains largely unaffected by rain, fog, dust, and lighting conditions [3]. This characteristic ensures reliable perception capabilities across diverse weather scenarios and time of day variations, complementing the limitations of optical sensors like cameras and LiDAR [124].

Direct measurement of radial velocity through the Doppler effect provides radar with unique capabilities for identifying and tracking moving objects [125]. This velocity information is particularly valuable for distinguishing between stationary and moving obstacles, predicting future object positions, and understanding traffic flow dynamics [126]. The extended range capabilities of radar, typically exceeding 200 meters for modern automotive systems, enable early detection of hazards and sufficient reaction time for high speed scenarios [127].

Low power consumption relative to active sensors like LiDAR makes radar an energy efficient option for both conventional and electric vehicles [128]. Additionally, the mature manufacturing ecosystem and economies of scale have driven down costs, making radar systems economically viable for mass market deployment across various vehicle segments.

Despite these advantages, mmWave radar systems face several limitations in automotive applications. The limited angular resolution, particularly in the azimuth dimension, presents challenges for accurately determining the lateral position and extent of objects [129]. For example, typical automotive radar systems achieve angular resolutions of 1-5 degrees, which translates to lateral position uncertainties of several meters at long ranges [118]. This limitation makes it difficult to precisely locate objects within their lanes or distinguish closely spaced objects [126].

Radar measurements also exhibit higher noise levels and more complex artifact patterns compared to LiDAR or camera data. Multipath reflections, where signals reach the receiver after bouncing off multiple surfaces, can create ghost targets that appear as false objects [127]. Similarly, sidelobe effects from strong reflectors can mask weaker targets

or create false detections. These phenomena necessitate sophisticated filtering and validation mechanisms to ensure reliable perception.

The radar cross section (RCS) of objects varies significantly based on their material composition, shape, and orientation relative to the radar [105]. Certain objects, particularly those with low metallic content like pedestrians and cyclists, present smaller and more variable radar signatures, making them challenging to detect consistently [119]. Additionally, the complex scattering behavior of real world environments creates difficult edge cases such as tunnels, metallic guardrails, and overhead structures that can generate confusing radar returns [130].

From a signal processing perspective, the inherent tradeoffs between range resolution, velocity resolution, and frame rate impose system level constraints [131]. Improving one aspect often requires compromising another, given fixed hardware capabilities and bandwidth allocations. For instance, achieving finer range resolution requires increased bandwidth, while better velocity resolution necessitates longer observation times, reducing the frame rate [132].

Research efforts to address these limitations have pursued several directions, including advanced signal processing techniques, sensor fusion approaches [2], and machine learning methods [8]. Signal processing advancements such as super resolution algorithms [4] and advanced MIMO configurations [113] aim to overcome the physical limitations of angular resolution. Sensor fusion strategies combine radar with complementary sensors like cameras and LiDAR to leverage the strengths of each modality. Machine learning approaches, particularly deep neural networks, attempt to extract more information from radar data by learning complex patterns and relationships that traditional processing methods might miss [9].

The integration of these approaches continues to enhance the capabilities of radar based perception systems, incrementally and progressively addressing the limitations while leveraging the inherent advantages of mmWave radar for automotive applications [131]. As these technologies mature, radar is expected to maintain its role as a critical

component in the sensor suite of advanced driver assistance systems and autonomous vehicles, providing reliable sensing capabilities across diverse operating conditions [5].

Summary

This chapter established the fundamental elements necessary for understanding radar-based object detection. The discussion of neural networks covered critical concepts and architectures, from basic artificial neurons through modern deep learning architectures including CNNs, RNNs, and transformer-based models relevant to radar data processing. Object detection fundamentals were examined, progressing from traditional approaches such as the Viola-Jones detector through deep learning-based frameworks including the R-CNN family, YOLO, and SSD architectures, along with their associated detection heads and loss functions. The chapter provided a comprehensive treatment of mmWave radar sensing, covering FMCW radar principles, the signal processing chain from ADC through FFT operations, and various radar data representations including Range-Azimuth-Doppler cubes and point clouds. The discussion concluded with an analysis of the advantages and limitations of radar sensing in automotive environments, establishing the theoretical basis for the subsequent technical contributions.

Chapter 3

Literature Review

3.1 Traditional Radar Processing Methods

Traditional radar processing methods form the foundation upon which modern deep learning approaches for radar based object detection are built. These methods typically follow a sequential pipeline that transforms raw radar measurements into interpretable object detections through a series of well established algorithms [102]. This section explores the key components of this traditional processing chain, focusing on Constant False Alarm Rate (CFAR) detection, clustering algorithms, and tracking approaches.

3.1.1 CFAR Detection

Constant False Alarm Rate (CFAR) detection represents the cornerstone of traditional radar detection pipelines, addressing the fundamental challenge of distinguishing actual targets from noise and clutter in radar measurements [106]. The central principle of CFAR is to maintain a consistent probability of false alarm by adaptively setting detection thresholds based on the local noise and clutter characteristics surrounding each cell under test (CUT) [116].

The most basic variant, Cell-Averaging CFAR (CA-CFAR), estimates the background noise level by averaging the power returns from a set of reference cells surrounding the CUT, with guard cells typically excluded to prevent target self-masking [106]. The

detection threshold is then established by multiplying this estimated noise level by a scaling factor determined based on the desired false alarm rate. Mathematically, the CA-CFAR decision rule can be expressed as:

$$x_{\mathrm{CUT}} \underset{H_0}{\overset{H_1}{\gtrless}} \alpha \cdot \frac{1}{N} \sum_{i=1}^{N} x_i$$

where x_{cut} is the power of the cell under test, xi represents the power of the ith reference cell, N is the number of reference cells, and α is the scaling factor determined by the desired probability of false alarm [106].

While CA-CFAR performs well in homogeneous noise environments, its performance degrades in scenarios involving multiple targets or clutter edges [117]. This limitation has led to the development of numerous CFAR variants, each designed to address specific operational challenges. Ordered Statistics CFAR (OS-CFAR) ranks the reference cells by power and selects a specific percentile as the noise estimate, providing robustness against interfering targets [106]. Greatest-of CFAR (GO-CFAR) and Smallest-of CFAR (SO-CFAR) divide the reference window into leading and trailing segments, using the maximum or minimum average, respectively, to estimate the noise level, with GO-CFAR demonstrating resilience to clutter edges and SO-CFAR handling multiple target scenarios more effectively [106].

For automotive radar applications, two dimensional CFAR processors are commonly employed to operate on Range-Doppler or Range-Azimuth maps, taking into account the spatial correlation of noise and clutter across both dimensions [116]. These 2D-CFAR processors typically slide a window across the radar map, applying the CFAR algorithm at each position to detect peaks that exceed the adaptive threshold [117].

Despite its effectiveness, CFAR detection introduces several limitations that affect subsequent processing stages. The binary nature of CFAR decisions results in information loss, as amplitude and shape characteristics of the original signal are

discarded during thresholding [9]. Additionally, the performance of CFAR detectors is highly dependent on parameter settings, such as the number of reference and guard cells, which may require careful tuning for specific operational environments [106]. These limitations have motivated research into deep learning approaches that can potentially bypass the CFAR detection stage entirely, working directly with pre-CFAR radar data to extract more nuanced information [9].

3.1.2 Clustering Algorithms

Following CFAR detection, the resulting point cloud typically contains numerous detection points that must be grouped into coherent object hypotheses. Clustering algorithms play a crucial role in this process, aggregating spatially proximate points that likely originate from the same physical object [118].

Density Based Spatial Clustering of Applications with Noise (DBSCAN) represents one of the most widely adopted clustering algorithms for radar point clouds due to its ability to discover clusters of arbitrary shape without requiring a predefined number of clusters [6]. DBSCAN defines clusters as dense regions separated by regions of lower density, using two key parameters: the neighborhood radius (ϵ) and the minimum number of points required to form a dense region (MinPts) [118].

The algorithm classifies points into three categories: core points (with at least MinPts within distance ϵ), border points (within distance ϵ of a core point but with fewer than MinPts neighbors), and noise points (neither core nor border) [6]. The clustering process begins by selecting an arbitrary unvisited point, identifying all points density reachable from it based on ϵ and MinPts, and marking them as belonging to the same cluster. This process repeats until all points have been processed [118].

For radar applications, DBSCAN offers several advantages, including robustness to noise, ability to discover clusters of arbitrary shape, and not requiring prior knowledge of the number of objects in the scene. However, the algorithm's performance is sensitive to parameter selection, with ϵ and MinPts typically requiring careful tuning based on the

radar's resolution and the expected characteristics of objects in the scene. Additionally, DBSCAN may struggle with clusters of varying densities, potentially merging nearby objects or fragmenting extended objects into multiple clusters [6].

Gaussian Mixture Models (GMM) provide an alternative probabilistic approach to clustering radar points [6]. Unlike DBSCAN, which assigns points to clusters deterministically, GMM represents the point cloud as a mixture of several Gaussian distributions, with each distribution corresponding to a potential object [6]. The model parameters, including the means, covariances, and mixing coefficients of the Gaussian components, are typically estimated using the Expectation Maximization (EM) algorithm [6].

In the work by Jin et al. [6], a GMM based approach was employed for radar point cloud segmentation in multimodal traffic monitoring. The authors observed that point clouds corresponding to different object types exhibited distinct Gaussian distributions with characteristic mean and variance values. By modeling these distributions, they achieved pedestrian classification with 88% precision and 61% recall, and car classification with 85% precision and 93% recall [6].

Other clustering approaches adapted for radar data include k-means, which partitions points into k clusters by minimizing the within-cluster sum of squares [118]; hierarchical clustering, which builds a hierarchy of clusters either through agglomerative (bottom-up) or divisive (top-down) strategies [118]; and grid based methods, which segment the spatial domain into a grid structure and perform clustering based on cell density [129].

3.1.3 Traditional Tracking Approaches

Tracking extends the detection and clustering processes from individual frames to sequences, establishing temporal correspondence between object detections across consecutive radar scans [118]. Traditional tracking approaches typically follow the detect-then-track paradigm, where objects are first detected in each frame independently before being associated across frames to form tracks [125].

The tracking process generally encompasses four main components: track initialization, data association, state estimation, and track management [118]. Track initialization determines when to establish a new track based on unassociated detections, often requiring multiple consecutive detections to confirm the presence of a genuine object rather than a false alarm. Data association resolves the correspondence between existing tracks and new detections, a challenging problem particularly in scenarios involving multiple closely spaced objects [125]. State estimation predicts and updates the kinematic state of tracked objects, typically using variants of Kalman filtering. Track management handles the lifecycle of tracks, including confirmation, maintenance, and termination based on detection history and track quality metrics [118].

The Hungarian algorithm represents a classical approach to data association, solving the assignment problem by finding the optimal one-to-one mapping between tracks and detections that minimizes a cost matrix, typically based on Euclidean distance or Mahalanobis distance between predicted track positions and new detections [91]. For scenarios involving varying numbers of objects, the Joint Probabilistic Data Association Filter (JPDAF) offers a probabilistic framework that considers all possible assignment hypotheses weighted by their respective probabilities [118].

Kalman filtering provides the foundation for state estimation in traditional tracking approaches, recursively predicting and updating the kinematic state (typically position, velocity, and sometimes acceleration) of tracked objects [125]. The standard Kalman filter assumes linear motion models and Gaussian noise distributions, which may not hold for complex object motions [125]. Extended Kalman Filter (EKF) addresses nonlinear motion models through linearization, while Unscented Kalman Filter (UKF) propagates a set of sigma points through nonlinear functions to better capture the transformed distribution [118].

For automotive applications, the Interacting Multiple Model (IMM) framework combines several motion models (e.g., constant velocity, constant acceleration, coordinated turn) running in parallel, with their outputs weighted based on their consistency with observed

measurements [125]. This approach handles the varying motion patterns of different road users, allowing for more accurate tracking across diverse scenarios [125].

The Multiple Hypothesis Tracking (MHT) algorithm represents another advanced approach that maintains multiple hypotheses about track detection associations over several frames, deferring hard decisions until more information becomes available [118]. While computationally intensive, MHT demonstrates robustness in challenging scenarios involving closely spaced objects, occlusions, and missed detections [118].

Recent developments in traditional tracking include the integration of object extent and shape information to enhance tracking performance for extended objects [125]. Kellner et al. [125] proposed a high resolution Doppler radar tracking system that models objects as elliptical shapes, with the shape parameters estimated alongside kinematic states. This approach improves tracking accuracy for extended objects like vehicles and enables better discrimination between different object types based on their characteristic dimensions and motion patterns [125].

Despite their widespread adoption, traditional radar processing methods face several limitations when applied to complex automotive scenarios [9]. The sequential nature of these algorithms, with each stage depending on the output of previous stages, allows errors to propagate through the pipeline [9]. The hand crafted features and manually tuned parameters used in these methods may not effectively extract the rich information contained in radar signals [9]. Additionally, these approaches typically process each frame independently or with limited temporal context, potentially missing valuable patterns that span multiple frames [9]. These limitations have motivated the exploration of deep learning approaches that can learn more effective representations directly from radar data, potentially bypassing or enhancing traditional processing stages [9].

3.2 CNN-based Approaches for Radar Object Detection

The application of Convolutional Neural Networks (CNNs) to radar data processing represents a significant paradigm shift from traditional methods, offering the potential to learn complex, hierarchical features directly from radar measurements [9]. This section explores the adaptation of CNN architectures for radar data, the extraction of meaningful features from various radar representations, and approaches to address radar specific challenges.

3.2.1 CNN Architectures for Radar Data

The adaptation of CNN architectures for radar data processing requires careful consideration of the unique characteristics of radar signals, including their dimensionality, sparsity, and physical interpretation [9]. Various network designs have been explored, each tailored to specific radar data representations and detection tasks [9, 12].

For Range-Azimuth-Doppler (RAD) tensor processing, 3D CNNs have been employed to capture correlations across all three dimensions simultaneously [9]. Major et al. [9] proposed a 3D CNN architecture operating directly on RAD tensors for vehicle detection, demonstrating improved performance compared to traditional methods. Their network comprised multiple 3D convolutional layers with small kernels (3×3×3), followed by batch normalization and ReLU activation, progressively reducing the spatial dimensions while increasing the feature channels [9]. The final layers included global average pooling and fully connected layers for classification and bounding box regression [9].

For Range-Azimuth (RA) or Range-Doppler (RD) maps, which represent 2D projections of the radar data, 2D CNN architectures similar to those used in image processing have been adapted [12]. Palffy et al. [12] employed a modified ResNet architecture for processing RA maps, introducing dilated convolutions to expand the receptive field without increasing the parameter count. Their network processed radar data at multiple resolutions, allowing it to capture both fine grained details and broader contextual information [12].

For radar point cloud processing, PointNet inspired architectures have been explored. Schumann et al. [10] adapted PointNet for processing automotive radar point clouds, incorporating Doppler velocity as an additional feature alongside the spatial coordinates. Their architecture included pointwise feature extraction using shared MLPs, followed by symmetric pooling operations (maxpooling) to achieve permutation invariance, and finally fully connected layers for classification.

Several works have explored hybrid architectures that combine different network components to leverage the strengths of each. For instance, Palffy et al. [122] introduced a multistage network that first processes radar point clouds using PointNet style layers to extract pointwise features, then projects these features onto a bird's eye view grid, and finally applies 2D convolutions to capture spatial relationships between projected points. This approach combines the flexibility of point based processing with the efficiency of convolutional operations.

U-Net and Feature Pyramid Network (FPN) architectures have been adapted for radar semantic segmentation tasks, allowing pixelwise classification of radar maps [119]. Nowruzi et al. [40] employed a U-Net variant for open space segmentation using automotive radar, with an encoder/decoder structure that preserves spatial resolution through skip connections. The network was designed to be computationally efficient for deployment on embedded automotive platforms.

3.2.2 Feature Extraction from Radar Data

Feature extraction from radar data involves identifying and learning meaningful patterns that enable effective object detection and classification [9]. Unlike image data, where features often have intuitive visual interpretations, radar features relate to physical phenomena such as reflection patterns, Doppler signatures, and spatial distributions of reflection points.

The first convolutional layers in radar CNNs typically extract low level features related to local signal characteristics [9]. For RAD tensors, these include patterns of signal strength

across range bins, Doppler shifts indicative of relative motion, and phase differences between antenna elements that encode angular information.

Intermediate layers in deeper networks aggregate these local features to form more complex representations [120]. For vehicle detection, these might include characteristic reflection patterns from vehicle corners, the distribution of reflection points along vehicle contours, and distinctive velocity profiles. For pedestrian detection, relevant features include micro Doppler signatures caused by limb movements, the relatively small spatial extent, and specific radar cross section distributions [119].

Feature learning from radar data is often complicated by the variable nature of radar reflections, which depend on object material, geometry, and orientation relative to the radar [105]. To address this challenge, several approaches have employed data augmentation techniques, including rotation, translation, and scaling of radar signals, as well as synthetic addition or removal of reflection points [12]. These augmentations help the network learn features that are robust to variations in object presentation and sensor perspective.

Multiscale feature extraction has proven particularly valuable for radar data processing [119]. Feature Pyramid Networks and atrous convolutions (dilated convolutions) enable the network to capture both fine grained details necessary for precise localization and broader contextual information that aids in classification . This multiscale approach addresses the challenge of detecting objects at various distances from the radar, where the angular resolution results in different point densities and distribution patterns [119].

For temporal feature extraction, several works have employed 3D convolutions that span both spatial and temporal dimensions. Wang et al. [51] proposed a spatiotemporal attention network that processes sequences of radar frames, extracting features that capture object motion patterns over time. The temporal dimension provides valuable context for distinguishing between object classes based on their characteristic movements.

Transfer learning has been explored as a strategy to leverage features learned from domains with abundant labeled data, such as camera images [9]. However, the significant differences between radar and image data limit the direct transferability of features, often necessitating substantial adaptation of pretrained networks. Some approaches have utilized multimodal training, where a network is trained to process both radar and camera data, allowing the radar branch to benefit from the richer semantic information available in the visual domain [126].

3.2.3 Handling Radar specific Challenges

CNN based approaches must address several challenges specific to radar data to achieve robust and accurate object detection performance [9]. These challenges include limited angular resolution, sparsity and noise in radar measurements, class imbalance, and efficient processing of radar data.

Limited angular resolution represents a significant challenge for radar based object detection, particularly in the azimuth dimension [129]. To address this limitation, several approaches have explored super resolution techniques based on deep learning [4]. These methods attempt to recover higher resolution information by learning correlations between low resolution radar measurements and corresponding high resolution ground truth, often provided by LiDAR or stereo cameras [4]. Additionally, some approaches leverage the higher range resolution of radar to compensate for limited angular resolution, using temporal integration of measurements as the radar or objects move to build a more complete representation [129].

Sparsity and noise in radar measurements present another significant challenge [10]. Radar point clouds are typically much sparser than LiDAR point clouds, with fewer points representing each object [10]. Furthermore, radar measurements often include false detections due to multipath reflections, sidelobes, and clutter [127]. To address these challenges, CNN architectures often incorporate attention mechanisms that focus on the most reliable measurements while downweighting potential noise [51]. Robust loss functions, such as Huber loss or Focal loss, have been employed to reduce the impact

of noisy labels and outliers during training [12]. Some approaches also integrate confidence measures for each radar detection point, allowing the network to learn which measurements are more reliable [122].

Class imbalance issues arise from the uneven distribution of object classes in typical driving scenarios, with cars and other vehicles much more common than vulnerable road users like pedestrians and cyclists [119]. This imbalance can bias networks toward the majority classes, leading to poor detection performance for less represented but critically important classes. Techniques to address this challenge include weighted loss functions that assign higher importance to minority classes, oversampling of minority classes during training, and hard negative mining to focus learning on difficult examples . Additionally, some approaches employ curriculum learning strategies, starting with balanced subsets of the data and gradually introducing the natural class distribution [12].

Efficient processing of radar data is crucial for automotive applications, where real-time performance and limited computational resources are important constraints [40]. Architectures designed specifically for efficient inference, such as MobileNet inspired designs with depthwise separable convolutions, have been adapted for radar processing [122]. Additionally, some approaches employ early fusion of radar channels to reduce the input dimensionality before more computationally intensive processing stages [9].

Domain adaptation techniques have been explored to address the gap between synthetic radar data, which can be generated in large quantities with perfect annotations, and realworld radar measurements with their characteristic noise patterns and artifacts [8]. Approaches include adversarial training to align feature distributions between domains, gradually introducing realistic noise patterns during training, and mixed batches containing both synthetic and real data.

Multimodal fusion with complementary sensors represents another strategy to overcome radar specific limitations [126]. By combining radar with cameras, LiDAR, or other sensors, these approaches can leverage the strengths of each modality while mitigating their individual weaknesses [126]. Lekic and Babic [126] employed generative adversarial

networks to facilitate radar and camera fusion, learning a joint representation that preserves the semantic richness of camera data while incorporating the range and velocity information from radar. Other approaches use radar as an attention mechanism to guide processing in other modalities, focusing computational resources on regions likely to contain objects based on radar detections [126].

Despite these advances in addressing radar specific challenges, CNN based approaches still face limitations in capturing long range dependencies and modeling sequential data effectively [48]. These limitations have motivated exploration of alternative architectures, particularly transformer based models, which will be discussed in the subsequent section [48].

3.3 Transformer based Approaches

The introduction of transformer architectures has marked a paradigm shift in deep learning approaches for various perception tasks, including radar based object detection [48]. Originally developed for natural language processing [36], transformers have demonstrated remarkable capabilities in modeling long range dependencies and capturing global context information, attributes that prove valuable for processing radar data with its unique spatiotemporal characteristics [48].

The fundamental building block of transformer architectures is the self-attention mechanism, which enables each element in a sequence to attend to all other elements, capturing relationships regardless of their distance in the sequence [36]. This contrasts with the inherently local nature of convolutional operations, which are limited by their receptive field size and require deep stacking to capture long range dependencies [36]. For radar data processing, this global attention capability allows transformers to model relationships between reflection points across the entire scene, potentially identifying object patterns despite the sparse and noisy nature of radar measurements [48].

The adoption of transformers for computer vision tasks gained significant momentum with the introduction of the Vision Transformer (ViT) [37], which demonstrated that pure transformer models could achieve state of the art performance on image classification tasks. ViT divides an image into fixed size patches, linearly embeds each patch, adds position embeddings, and processes the resulting sequence with a standard transformer encoder [37]. This approach treats image patches similarly to word tokens in language processing, allowing direct application of transformer architectures to visual data [37].

For object detection, DETR (DEtection TRansformer) [38] pioneered a transformer based approach that formulates detection as a direct set prediction problem. DETR employs a transformer encoder/decoder architecture to process image features extracted by a CNN backbone, with the decoder generating a fixed set of object predictions in parallel [38]. This approach eliminates the need for many hand designed components in traditional detection pipelines, such as anchor generation and non-maximum suppression, offering a cleaner and potentially more effective detection framework [38].

The adaptation of transformer architectures for radar data processing presents both opportunities and challenges [48]. On one hand, the global attention mechanism is well suited for capturing the complex relationships between radar reflection points that may represent the same physical object [48]. On the other hand, the computational complexity of self-attention, which scales quadratically with sequence length, poses efficiency challenges for processing high resolution radar data [48]. Additionally, radar data lacks the rich semantic information found in images, potentially limiting the effectiveness of direct application of vision transformer approaches [48].

Recent research has explored various modifications to the transformer architecture to better suit radar data processing [48, 50]. These adaptations include specialized embedding layers for radar features, attention mechanisms that incorporate radar specific properties such as range and Doppler information, and hybrid architectures that combine convolutional operations for local feature extraction with transformer modules for global context modeling [48, 50]. The following sections analyze key papers in this domain that

influenced the research in this literature, focusing on their architectural innovations and contributions to radar based object detection.

3.4 Analysis of Key Papers for Radar Object Detection

3.4.1 RADDet: Range-Azimuth-Doppler based Radar Object Detection for Dynamic Road Users [123]

This pioneering work by Zhang et al. represents a foundational contribution to deep learning based radar object detection, establishing both methodological frameworks and dataset standards that would influence subsequent research. The paper addresses a critical gap in autonomous driving perception by focusing specifically on dynamic road users through Range-Azimuth-Doppler (RAD) tensor analysis, marking a significant departure from traditional radar point cloud approaches.

The authors' decision to work with full RAD tensors rather than sparse point clouds demonstrates sophisticated understanding of radar signal processing. While radar point clouds typically contain fewer than 5 points per nearby vehicle, RAD tensors preserve rich information about object motion patterns, surface texture variations, and micro Doppler signatures. This preservation of raw signal information proves particularly valuable for distinguishing between different types of road users; pedestrians exhibit complex micro Doppler patterns due to limb movement, while vehicles show more coherent Doppler signatures. The paper effectively argues that this additional information justifies the increased computational complexity of processing 3D tensors.

The instancewise auto annotation methodology represents an innovative solution to one of radar perception's most challenging problems. The enhanced radar preprocessing pipeline, which connects discrete patterns on Range-Doppler spectrums and applies DBSCAN clustering, addresses the traditional limitations of 2D OS-CFAR detection. By recognizing that rigid bodies create linear patterns in RD spectrums regardless of viewing angle, the authors develop a more robust detection framework. The integration with stereo vision for category labeling demonstrates practical problem solving, while radar

excels at detection and localization, visual sensors remain superior for classification tasks.

The RadarResNet architecture shows thoughtful adaptation of computer vision techniques to radar's unique characteristics. The dual detection head design is particularly innovative, simultaneously predicting 3D bounding boxes in RAD space ([x, y, z, w, h, d]) and 2D boxes in Cartesian bird's-eye-view. This dual approach acknowledges that different downstream applications may prefer different output formats while enabling the model to leverage both representations during training. The coordinate transformation layer using channelwise fully connected networks represents a learnable approach to the nonlinear polar-to-Cartesian transformation, potentially capturing complex radar specific distortions that simple geometric transformations might miss.

The experimental design reveals both strengths and limitations. The K-means clustering for anchor box generation, resulting in 6 anchors for both 3D and 2D heads, follows established object detection practices while adapting to radar specific object distributions. The choice of loss functions, Complete-IOU for box regression, Focal Loss for objectness, and Cross Entropy for classification, demonstrates understanding of different task requirements. The $\beta = 0.1$ weighting for box loss prevents regression from dominating training, a crucial insight for radar where classification is inherently more challenging than localization.

However, several limitations constrain the work's impact. The dataset's 10,158 frames, while substantial for initial research, fall short in comparison to modern perception datasets. More critically, collection exclusively during sunny weather conditions undermines radar's primary advantage: all weather operation. The severe class imbalance (cars dominating over motorcycles, buses, and bicycles) reflects real world distributions but poses significant training challenges. The 75% auto annotation coverage, requiring manual completion, suggests scalability limitations for larger datasets.

The reported performance metrics, 56.3% AP@0.3 for 3D detection and 51.6% AP@0.5 for 2D detection, establish important baselines but highlight substantial room for

improvement. The performance drop at higher IOU thresholds indicates localization precision challenges, possibly due to radar's inherent range-azimuth resolution limitations. The inference time of 75.2ms for RadarResNet, while real-time capable, suggests computational optimization opportunities.

The paper's discussion of architecture exploration provides valuable insights. VGG based alternatives performing worse than ResNet variants confirms that deeper residual connections benefit radar feature extraction. The superiority of maxpooling over strided convolution for downsampling might relate to radar's sparse meaningful features, since maxpooling preserves peak responses while strided convolution might miss critical sparse activations. The exploration of self-attention mechanisms (SAGAN and SAUNet), while not improving performance, presages later transformer based approaches and highlights that simple application of image based techniques doesn't guarantee success for radar.

This work's primary contribution lies not in achieving state of the art performance but in establishing feasibility and frameworks. By demonstrating that end-to-end learning can extract meaningful features from RAD tensors, creating an annotated dataset (despite being limited), and establishing evaluation protocols, Zhang et al. created the foundation upon which subsequent research would build. The paper's transparency about limitations and failure modes, misclassification between cars and trucks, false positives in complex scenes, provides valuable guidance for continuing research directions. While the constrained dataset and weather conditions limit immediate practical application, this work opened the door for more ambitious radar perception research.

3.4.2 RODNet: A Real-Time Radar Object Detection Network Cross-Supervised by Camera-Radar Fused Object 3D Localization[51]

RODNet proposes a groundbreaking framework for real-time radar based object detection that bypasses traditional reliance on visual sensors during inference. At its core lies a cross modal supervision strategy, Camera-Radar Fusion (CRF),that enables radar only perception by leveraging vision based annotations exclusively during training. This dual

modality design addresses one of the central challenges in radar deep learning: the absence of large scale, high quality labeled radar datasets. Rather than depending on labor intensive and error prone manual labeling of radar signals, RODNet's supervision pipeline systematically generates radar pseudo labels by fusing high resolution camera based detections with radar reflections using probabilistic modeling. The result is a student-teacher framework that trains radar perception systems under favorable visual conditions while operating independently of visual input during deployment.

The CRF supervision framework constitutes the backbone of this innovation. A visual teacher network uses Mask R-CNN for object detection, self-supervised monocular depth estimation, adaptive ground plane fitting, and multi object tracking (based on TrackletNet) to recover accurate 3D object locations. These locations are projected into radar's native range-azimuth coordinate system and aligned with radar signal peaks obtained via the CFAR (Constant False Alarm Rate) algorithm. Both radar and visual sources are modeled using multivariate Gaussian probability maps with class dependent spatial uncertainty parameters. Camera localization uncertainty is modeled using depth confidence and azimuthal spread, while radar uncertainty accounts for its high range precision and angular resolution variation. By computing the elementwise product of radar and camera probability maps, the system fuses complementary sensor strengths and suppresses noise, resulting in robust pseudo label generation. These fused labels train RODNet without requiring direct human annotation of radar data.

The architectural design of RODNet is equally innovative. Built upon a 3D convolutional autoencoder, the base architecture is further enhanced using an hourglass (HG) backbone with skip connections to capture multiscale spatiotemporal features. Three custom modules are introduced to handle radar specific data properties: M-Net, Temporal Deformable Convolution (TDC), and temporal inception. M-Net is designed to merge chirp level information from multiple radar chirps within each frame. Using learnable temporal convolutions followed by maxpooling, M-Net effectively replicates and generalizes Doppler processing traditionally handled via FFTs. TDC, inspired by deformable convolutions in computer vision, adapts the receptive field across time to

account for object motion in radar data. Unlike traditional 3D convolutions, which operate over rigid spatiotemporal grids, TDC modifies the spatial sampling dynamically, preserving temporal alignment and better capturing object displacement across frames. Temporal inception layers, added at deeper levels, allow the network to capture patterns over varying time scales, which is critical in dynamic driving environments.

To support and evaluate this architecture, the authors introduce the CRUW (Camera-Radar of the University of Washington) dataset. Comprising ~400,000 synchronized radar and RGB frames, CRUW is the first large scale dataset offering radar RF images rather than sparse radar points. The dataset covers 464 sequences from a variety of scenarios, parking lots, urban roads, highways, and campus areas, with 260,000 labeled objects. Data is collected using a stereo camera system and two perpendicular 77GHz FMCW radars, with parameters like 0.23 m range resolution and ~15° azimuth resolution. Although stereo vision aids CRF supervision during validation, training annotations require only monocular input, making the system efficient and scalable. The dataset includes “hard” sequences, scenes with overexposure, nighttime conditions, and occlusions, allowing rigorous evaluation of radar’s robustness where vision based systems typically fail.

RODNet introduces a novel evaluation metric tailored for radar based object detection: Object Location Similarity (OLS). Analogous to Intersection over Union (IoU) in image based detection, OLS measures spatial closeness between predicted and ground truth object centers, weighted by range and class dependent size priors. This allows for meaningful comparisons in radar’s native polar coordinate system. OLS is also used to implement a location based non-maximum suppression (L-NMS) algorithm, which filters redundant detections based on similarity, rather than overlapping bounding boxes which radar does not naturally produce.

Experimental results are compelling. RODNet achieves 86% average precision (AP) and

88% average recall (AR) on the CRUW test set, outperforming multiple radar only

baselines by wide margins. In challenging scenarios, such as low light and cluttered backgrounds, RODNet maintains 67.28% AP, compared to <11% from CFAR+CNN approaches. Ablation studies confirm the additive contributions of each module: CRF supervision improves AP by 8%, the hourglass backbone by 5%, and the combined use of M-Net, TDC, and inception layers yield an additional 3-5%. Moreover, RODNet proves efficient: using just 8 out of 255 chirps (∼3% of a full frame), it sustains high performance, showing strong data compression capabilities. Real-time feasibility is also addressed, models with simpler backbones can achieve <100ms inference times, suitable for deployment in real-world ADAS or AV systems.

The authors further analyze feature activations within the trained RODNet, showing that deeper layers learn class discriminative semantic patterns, despite radar's low spatial resolution. Qualitative results highlight RODNet’s advantages in handling occlusions, low light environments, and clutter, situations that typically challenge camera based detectors. For instance, RODNet correctly detects partially visible or occluded pedestrians and cyclists missed by vision methods.

While RODNet demonstrates outstanding performance, it is not without limitations. Its reliance on camera derived supervision, while effective, may restrict training in scenarios where camera data is unavailable or uncalibrated. The system also demands considerable training time (4–10 days depending on architecture) and high end GPUs. However, these tradeoffs are offset by its ability to extract rich semantic understanding from radar data, shifting radar perception from mere detection toward structured scene understanding.

3.4.3 T-RODNet: Transformer for Vehicular Millimeter-Wave Radar Object Detection [48]

T-RODNet marks a crucial evolution in radar object detection by being the first to successfully integrate transformer architectures, addressing fundamental limitations of CNN based approaches while achieving dramatic improvements in computational efficiency. This work demonstrates how architectural innovations from natural language processing and computer vision can be thoughtfully adapted to radar's unique challenges.

The paper's motivation stems from two critical observations about existing radar perception methods. First, 3D CNNs, while effective at capturing local spatiotemporal features, require extreme depth to model global dependencies, leading to computational complexity that makes real-time deployment challenging. RODNet-HG, for instance, requires 65 hours of training and achieves only borderline real-time inference. Second, CNNs' limited receptive fields struggle to capture long range dependencies in radar sequences, missing crucial temporal correlations that could improve detection accuracy. The authors recognize that transformers' self-attention mechanisms naturally address both limitations.

The architectural design of T-RODNet addressed radar requirements with transformer capabilities. Rather than simply replacing CNNs with transformers, the authors develop a hybrid architecture leveraging each component's strengths. The encoder combines modified Dimensional Apart Modules (DAM) using dilated convolutions with 3D Swin Transformer blocks. This design acknowledges that local feature extraction remains important for radar, the modified DAM with 1×5×1, 1×1×5, and 1×5×5 dilated convolutions expands receptive fields without computational penalty, capturing multiscale local patterns essential for radar target characterization.

The adaptation of Swin Transformer's windowed attention mechanism proves particularly well suited for radar data. The 4×4×4 window size balances computational efficiency with sufficient context, while the alternating window based and shifted window attention enables information flow across window boundaries. The use of Group Normalization

instead of Batch Normalization addresses practical training constraints, radar networks' large parameter counts often necessitate small batch sizes where GN performs superior to BN.

The T-W-MSA/T-SW-MSA modules in the decoder represent novel architectural innovation inspired by NLP transformers. Unlike typical vision transformers that use self-attention in encoders only, T-RODNet implements cross-attention in the decoder, fusing encoder keys/values with decoder queries. The learnable ratio parameter weighting encoder/decoder feature fusion allows the network to adaptively balance global context from the encoder with local refinement in the decoder. This design particularly benefits radar where global motion patterns inform local detection decisions.

The comprehensive experimental validation across both CRUW and CARRADA datasets strengthens the findings' generalizability. On CRUW, T-RODNet achieves comparable detection performance (83.27% AP) to RODNet while reducing GFLOPs to just 8.5% of RODNet-HG. This significant efficiency improvement transforms radar perception feasibility for embedded deployment. The CARRADA results (IoU improvement of 2.2 over TMVA-Net) confirm effectiveness across different radar configurations and annotation styles.

The robustness evaluation with additive Gaussian noise provides crucial insights into transformers' advantages for radar. T-RODNet maintains over 50% AP at 6dB SNR where CNN based methods drop below 5%. This dramatic difference suggests that self-attention's ability to dynamically focus on salient regions while suppressing noise provides fundamental advantages for radar's low-SNR regime. The gradual degradation curve indicates graceful degradation rather than catastrophic failure, essential for safety critical applications.

The ablation studies reveal nuanced insights into component contributions. The Swin Transformer backbone alone provides 2.69% AP improvement, validating the global modeling hypothesis. The T-W/SW-MSA decoder modules add 1.68%, confirming that transformer based decoding benefits radar. The modified DAM contributes 2.24%,

emphasizing that local feature extraction remains important even with transformers. The GN versus BN comparison (significant performance drop with BN) highlights practical considerations often overlooked in academic research.

The paper acknowledges limitations and challenges. The parameter count remains high due to transformers' inherent complexity, potentially limiting deployment on resource constrained platforms. The training time of 15 hours, while improved from predecessors, still poses practical challenges for iterative development. The observation that transformer based models exhibit different overfitting characteristics than CNNs suggests that regularization strategies need rethinking for radar applications.

The qualitative analysis provides valuable insights into failure modes and success cases. The reduction in "ghosting" artifacts (duplicate detections from temporal aliasing) demonstrates transformers' superior temporal modeling. Misclassification examples reveal that global context helps but doesn't eliminate radar's fundamental classification challenges. The improved performance in cluttered scenarios suggests that attention mechanisms effectively suppress background interference.

T-RODNet's contribution extends beyond performance metrics to demonstrate architectural feasibility. By showing that transformers can dramatically reduce computational requirements while maintaining accuracy, the work opens new research directions. The successful hybrid CNN-transformer design provides a template for balancing local and global feature extraction. The decoder innovations suggest that radar might benefit from architectures diverging further from standard vision models.

The paper's impact on subsequent research is evident. Later works like TransRAD, Mask-RadarNet and the work presented in this literature build upon T-RODNet's transformer foundation while addressing its limitations. By breaking the CNN-only paradigm and demonstrating transformers' unique advantages for radar, T-RODNet represents a pivotal moment in radar perception research, enabling a new generation of efficient, robust detection systems.

3.4.4 TransRAD: Retentive Vision Transformer for Enhanced Radar Object Detection [52]

TransRAD introduced a transformer based radar object detection architecture that redefines the processing of Range-Azimuth-Doppler (RAD) radar data through radar specific innovations in transformer design. Unlike most prior efforts that adapt vision transformers or 3D CNNs with minimal consideration for radar's spatial structure, TransRAD is purposefully built framework that tightly couples radar characteristics with deep transformer based modeling, delivering significant gains in both performance and computational efficiency.

At the core of TransRAD is its Retentive Manhattan Self-Attention (MaSA) mechanism, which departs from the uniform attention weights of traditional self-attention and instead uses explicit spatial decay priors modeled with Manhattan distance. This choice aligns with the grid structured nature of RAD data and radar target characteristics, where target reflections exhibit a high intensity core that fades toward the edges. MaSA's bidirectional and decomposed implementation (horizontal and vertical attention computed separately) allows the network to model spatial importance while reducing computational complexity. The use of decomposed MaSA likely reduces the computational burden compared to full 2D attention, although the paper does not provide explicit complexity analysis. Unlike the original Retentive Network, TransRAD integrates a Softmax nonlinearity into MaSA, enhancing convergence and stability; an essential component given the noisy and imbalanced nature of radar datasets.

TransRAD's backbone is based on a Retentive Vision Transformer (RMT), chosen for its ability to efficiently capture both global context and local details. RMT uses a combination of full MaSA blocks in later stages and decomposed MaSA in earlier layers. Additionally, Local Context Enhancement (LCE) via depthwise convolutions compensates for transformers' limited local receptive fields. The resulting architecture has a lightweight configuration (2-2-8-2 MaSA blocks with embedding dims 32-64-128-256), optimized for radar's lower-resolution, lower-diversity data. This yields a small model (5.78M parameters) capable of real-time performance (4.37 ms/frame).

For multi-scale feature learning, TransRAD employs a radar specific adaptation of the Feature Pyramid Network (FPN) as the neck. FPN is well suited for radar due to the predominance of small, irregularly shaped targets which can be lost in deeper layers. By combining high resolution features with semantic richness from deeper layers, FPN in TransRAD enables effective small object detection without sacrificing semantic abstraction.

TransRAD's anchor free detection heads are another radar-aware innovation. Traditional anchor based methods, derived from camera based object detection, perform poorly with radar's shape variability. Instead, TransRAD uses four decoupled heads: objectness, class, RA 2D bounding box, and Doppler regression. This separation reduces interference between tasks and allows specialized loss design per task. Notably, the Doppler head regresses $[z_1, z_2]$ extents, sidestepping full 3D box prediction while preserving motion information. This enables 3D detection in RAD space ($[x_1, y_1, z_1, x_2, y_2, z_2]$) using only 2D projections and 1D extents, eliminating the need for memory heavy 3D convolutions.

The loss function is meticulously designed with nine terms, each reflecting a different aspect of radar detection. CIoU loss governs 2D RA and RD bounding boxes, accounting for overlap, center distance, and aspect ratio. A Distribution Focal Loss (DFL) accounts for spatial uncertainty. Focal Loss handles class imbalance, using class weighting based on inverse frequency (addressing severe imbalance in the RADDet dataset). Smooth L1 loss is used for Doppler regression to minimize sensitivity to outliers. Center loss is included to stabilize small object localization. An optional 3D IoU loss supplements the overall loss, and each term is weighted to ensure balanced training. Task aligned learning (TAL) is used for label assignment, favoring samples with strong alignment across localization and classification. A key innovation is retaining unnormalized task alignment weights (t rather than $\hat{t}$) to preserve learning signal under radar's low IoU characteristics.

Perhaps the most elegant solution in TransRAD is the Location-Aware NMS (LA-NMS). Traditional NMS assumes high classification confidence, which radar lacks. LA-NMS

filters overlapping boxes even across different predicted classes based on IoU overlap. This leverages radar's high localization accuracy while sidestepping its classification weakness. Combined with the low probability of true object overlap in radar scenes, LA-NMS delivers substantial gains in post-processing effectiveness; contributing ~2.7% AP gain in ablation studies.

Quantitative results on the RADDet dataset are strong. TransRAD achieves 61.9% AP@0.3 for 3D detection, outperforming RODNet-CDC, RAMP-CNN, RadarResNet, and even transformer-heavy RadarFormer and T-RODNet, some by a factor of 2–6× on higher IoU thresholds. For 2D detection on the RA plane, it achieves 55.90% AP@0.5, with consistent performance at higher thresholds (up to AP@0.9), validating precise localization. Notably, inference runs at 4.37 ms/frame on a single V100 GPU, enabling true real-time deployment. Despite its transformer backbone, TransRAD remains lightweight and outpaces models like YOLOv8 and RadarFormer in both inference speed and accuracy.

Ablation studies reveal the modularity and additive impact of each innovation. Removing the FPN Neck reduces performance by 6%, and replacing decoupled heads with a coupled design reduces performance by ~2%. Disabling 2D losses and center loss affects localization at higher thresholds. Removing LA-NMS introduces cross-class duplicates and causes performance to drop to 59.14% AP@0.3.

In qualitative results, TransRAD consistently matches ground truth, while RadarResNet fails on small or misclassified targets and produces duplicate boxes. LA-NMS eliminates overlapping class predictions; decoupled heads improve classification objectness separation; and the full model maintains strong detection under clutter and ambiguity. One noted failure case is a missed detection of a motorcycle, highlighting an opportunity to improve performance on underrepresented classes.

3.4.5 Mask-RadarNet: Enhancing Transformer With Spatial-Temporal Semantic Context [50]

Mask-RadarNet [50] represents a comprehensive synthesis of recent advances in radar perception, introducing spatiotemporal semantic context through innovative architectural components while maintaining computational efficiency. This work demonstrates mature understanding of radar signal processing, deep learning, and the specific challenges of radar based object detection in autonomous driving applications.

The paper's core innovation, the Class Masking Attention Module (CMAM), addresses a fundamental limitation in previous radar perception approaches: the lack of explicit semantic reasoning. While previous methods focused on geometric feature extraction and motion modeling, CMAM introduces semantic space projection through class embedding layers. The query and key features undergo classwise projection, creating C×H×W×class tensors that explicitly represent class conditional spatial distributions. This design enables the network to learn class specific radar signature patterns, potentially capturing physical differences in how pedestrians, cyclists, and vehicles reflect radar signals.

The mathematical formulation of CMAM reveals sophisticated design choices. The similarity computation in class embedded space creates attention weights informed by semantic compatibility. The value features remain in the original feature space, preventing information loss while being modulated by semantic attention. The learnable scalar β for residual weighting allows smooth optimization from identity initialization. This design ensures that semantic enhancement augments rather than replaces geometric features.

The patch shift mechanism demonstrates elegant efficiency in temporal modeling. Unlike computationally expensive 3D convolutions or memory intensive recurrent architectures, patch shift achieves temporal fusion with zero computational cost. The exploration of different shift patterns (3-frame, 4-frame, and 9-frame temporal fields) shows systematic investigation of temporal context requirements. The finding that Pattern C (9-frame field)

performs best, particularly for cyclists, suggests that longer temporal context helps disambiguate complex motion patterns.

The architectural integration shows thoughtful system design. PatchShift 3D SwinTransformer modules handle local spatiotemporal features while CMAM provides global semantic context. The U-shaped architecture with skip connections ensures multiscale feature preservation. The auxiliary decoder aggregating prior maps from different stages enables deep supervision without cluttering the main detection pathway. This separation of semantic prior generation and final detection allows each component to specialize.

The experimental setup on the CRUW dataset enables direct comparison with previous methods. Achieving 84.29% AP and 87.36% AR [50] places Mask-RadarNet among top performers while using only 176.91 GFLOPs[50], demonstrating that semantic understanding doesn't require excessive computation. The particularly strong performance on cyclists (85.06% AP versus 82.28% for T-RODNet) validates that semantic context helps with challenging non-rigid objects.

The ablation studies provide crucial insights into component contributions. Patch shift alone improving AP by 2.94% confirms temporal modeling's importance. The modified DAM contributing 2.24% validates that local receptive field expansion remains valuable even with global attention. CMAM adding 2.69% improvement demonstrates semantic context's value. The exploration of auxiliary loss weights reveals $\alpha=0.4$ as suitable, suggesting semantic supervision should guide but not dominate training.

The feature visualization comparing pre/post CMAM activations provides rare insight into semantic enhancement mechanisms. The enhanced focus on object regions while suppressing background demonstrates that CMAM successfully learns to distinguish targets from clutter based on semantic consistency. This visualization validates the theoretical motivation: semantic context helps identify coherent objects within noisy radar returns.

The paper addresses practical deployment considerations often overlooked in academic work. The inference time of 14 hours training on consumer hardware (RTX 3080Ti) makes the approach accessible to researchers without massive computational resources. The parameter count of 32.12M [50], while not minimal, remains deployable on modern embedded platforms. The robustness to various driving scenarios (parking lots to highways) demonstrates generalization capability.

Several design decisions merit particular attention. Using Group Normalization throughout the architecture acknowledges radar's typical small batch training requirements. The 9×5×5 convolutional kernels in early stages provide large receptive fields crucial for sparse radar features. The auxiliary loss weight scheduling suggests that semantic supervision benefits from careful balancing with geometric objectives.

The limitations discussed reveal a clear assessment and future directions. The auxiliary decoder, while enabling semantic supervision, adds architectural complexity and parameters used only during training. The reliance on CRUW dataset, while enabling comparison, may limit generalization assessment. The semantic projection to fixed class embeddings assumes closed set detection, potentially limiting adaptation to new object types.

Mask-RadarNet's contribution lies not in revolutionary new concepts but in thoughtful integration of proven techniques with radar specific innovations. By showing that semantic context significantly improves detection performance without prohibitive computational costs, the work demonstrates that radar perception can benefit from high-level reasoning previously thought exclusive to vision based methods. The comprehensive experimental validation, careful ablation studies, and practical performance achievements establish Mask-RadarNet as a mature contribution that advances radar perception toward deployment readiness.

The paper's impact extends beyond specific technical contributions to demonstrate how the field has matured. Early works focused on proving feasibility; Mask-RadarNet assumes feasibility and optimizes for practical deployment. The creation of efficient

temporal modeling, semantic understanding, and computational efficiency represents the kind of holistic system design necessary for real-world autonomous driving applications.

3.4.6 E-RODNet: Lightweight Approach to Object Detection by Vehicular Millimeter-Wave Radar [134]

E-RODNet represents a significant advancement in addressing the computational efficiency challenges that have limited the practical deployment of radar-based object detection systems. While previous works demonstrated the feasibility of transformer architectures for radar perception, they often resulted in models with substantial parameter counts and computational requirements that hindered real-world implementation. E-RODNet tackles this critical gap by introducing a lightweight architecture that achieves superior performance with dramatically reduced computational complexity.

The paper's central innovation lies in its adoption of the ConvFormer block as the primary feature extraction unit, adapting it specifically for spatiotemporal radar data processing. The ConvFormer architecture, originally developed for vision tasks, represents a hybrid approach that combines the efficiency of convolutional operations with the global modeling capabilities of transformers. This choice reflects a strong understanding of the trade-offs between computational complexity and modeling capacity. The authors' modification of the SepConv operator within the ConvFormer block to accommodate 3D spatiotemporal modeling demonstrates thoughtful adaptation rather than naive application of existing architectures.

The MetaFormer architecture philosophy underlying ConvFormer proves particularly innovative for radar applications. Unlike traditional vision transformers that rely heavily on self-attention mechanisms, MetaFormer abstracts the transformer concept to its essential components: token mixing and channel mixing operations. This abstraction allows for more flexible implementations where attention can be replaced with more efficient operations while maintaining the beneficial structural properties of transformers.

For radar data, this approach is especially valuable given the sparse and noisy nature of radar returns, where the full complexity of self-attention may not always be necessary.

E-RODNet's adoption of MetaFormer principles enables the network to capture long-range dependencies crucial for radar object detection while maintaining computational efficiency. The ConvFormer blocks leverage separable convolutions for token mixing, which provides local receptive field expansion with reduced parameter overhead. This design choice acknowledges that radar data often contains strong local spatial correlations that can be effectively captured through convolutional operations, while still allowing for global context through the overall transformer structure.

The Global Feature Fusion (GFF) module represents another key innovation addressing the inherent limitations of ConvFormer's local modeling capabilities. The module adds global context information to every pixel during the encoding process, effectively addressing the challenge that pure convolutional approaches struggle with long-range spatial dependencies. The mathematical formulation of GFF, incorporating global average pooling followed by projection layers, creates a mechanism for broadcasting global context across all spatial locations. The inclusion of a projection layer, as validated through ablation studies, proves crucial for effective global feature integration, showing 0.34% AP improvement over methods without this component.

The enhanced short-time sequence fusion module demonstrates sophisticated temporal modeling adapted from RODNet-CDC. The authors recognize that effective multi-chirp feature extraction within radar frames significantly impacts network performance. Unlike previous approaches that treated temporal fusion as an afterthought, E-RODNet integrates temporal modeling as a core component of the architecture. This integration allows the network to capture both intra-frame temporal dynamics (across chirps) and inter-frame temporal relationships, crucial for robust object detection in dynamic environments.

The experimental results on the CRUW dataset reveal E-RODNet's high efficiency achievements. With only 6.1 million parameters and 33.25 GFLOPs, the model achieves

85.46% AP and 89.19% AR, surpassing T-RODNet's performance while using merely 3.8% of its parameters and 18.2% of its computational requirements. This dramatic efficiency improvement represents a paradigm shift from previous research that focused primarily on accuracy improvements often at the expense of computational complexity. The results demonstrate that sophisticated architectural design can achieve better performance with significantly fewer resources.

The ablation studies provide crucial insights into component contributions and validate the design choices. The short-term sequence fusion module contributes 8.75% AP improvement, confirming the importance of temporal modeling for radar applications. The Global Feature Fusion module adds 3.55% AP, demonstrating that global context enhancement provides meaningful benefits even when combined with local feature extraction. The systematic evaluation of different GFF configurations, including the impact of the projection layer, shows thorough empirical validation of architectural decisions.

The comparison with medical image segmentation models (3D UX-Net, Swin UNETR, PMFSNet3D) reveals interesting insights about domain adaptation. While these models perform well in static scenarios, they struggle with multi-object and cluttered environments, highlighting the importance of motion characteristics in radar data. E-RODNet's superior performance across all scenarios demonstrates that radar-specific architectural adaptations are essential for robust performance.

The literature’s discussion of computational efficiency extends beyond mere parameter counting to practical deployment considerations. The inference speed improvements, combined with maintained or enhanced accuracy, address real-world constraints faced by autonomous driving systems. The authors' emphasis on achieving practical deployability while maintaining state-of-the-art performance represents a mature approach to research that considers both academic metrics and industrial requirements.

Several design decisions merit particular attention. The U-shaped encoder-decoder structure provides multi-scale feature processing essential for detecting objects at various ranges and sizes. The integration of ConvFormer blocks at different scales allows the

network to capture both fine-grained local features and broader spatial patterns. The careful balance between local and global feature extraction through ConvFormer blocks and GFF modules demonstrates sophisticated architectural design.

The limitations acknowledged by the authors reveal honest assessment and future research directions. The reliance on visual sensor supervision for generating radar annotations, while practical, may limit the exploration of radar-specific detection capabilities. The authors' recognition of this limitation and suggestion for future work demonstrates scientific rigor and awareness of broader research challenges.

E-RODNet's contribution extends beyond specific technical innovations to demonstrate that efficient radar perception is achievable without sacrificing performance. By showing that MetaFormer-based architectures can dramatically reduce computational requirements while improving accuracy, the work enables practical deployment of sophisticated radar perception systems. The successful integration of global feature fusion with efficient local feature extraction provides a template for future efficient radar architectures.

The paper's impact on the field lies in shifting the focus from pure performance optimization to practical deployment considerations. Early transformer-based radar works proved feasibility but often resulted in computationally intensive models. E-RODNet demonstrates that careful architectural design can achieve the best of both worlds: superior accuracy with practical computational requirements. This contribution is particularly valuable for the autonomous driving industry, where computational resources are constrained and real-time performance is mandatory.

3.5 Activation Functions for Radar Signal Processing

While most radar-based deep learning approaches employ standard activation functions such as ReLU [16], recent work has begun exploring specialized activation functions designed for radar signal characteristics. Traditional activation functions, optimized for

natural image processing, may not be suitable for the sparse, high-dynamic-range nature of radar measurements.

The E-RODNet architecture [134] demonstrated improved performance using StarReLU activation, which applies quadratic scaling to positive inputs. However, this approach lacks adaptive characteristics necessary for handling the varying signal strengths typical in automotive radar applications, indicating a need for more sophisticated activation strategies tailored to radar signal processing requirements.

3.6 Advanced Topics in Radar Object Detection

3.6.1 Temporal Information Processing

The exploitation of temporal information represents a critical advancement in radar based object detection, enabling more robust performance through the integration of data across multiple radar frames [46]. Unlike static approaches that process each frame independently, temporal information processing leverages the coherence and evolution of radar signatures over time, providing valuable cues for distinguishing between different object types and tracking their motion [46]. This section explores approaches for sequence modeling in radar data, methods for exploiting temporal coherence, and techniques for dynamic object tracking.

3.6.1.1 Sequence Modeling in Radar Data

Sequence modeling techniques for radar data aim to capture the temporal dependencies and patterns that characterize different objects and their movements [46]. These approaches recognize that the temporal dimension contains valuable information that complements the spatial features available in individual frames [46].

Recurrent Neural Networks (RNNs), particularly Long Short Term Memory (LSTM) [44] and Gated Recurrent Unit (GRU) [45] variants, have been widely employed for radar sequence modeling due to their ability to maintain internal states that capture temporal

dependencies [46]. Kim et al. [46] proposed a convolutional recurrent neural network for moving target classification in automotive radar systems, which combined convolutional operators with LSTM and GRU layers for spatiotemporal feature extraction. Their architecture processed sequences of radar range-velocity images derived from FMCW radar measurements, with the convolutional recurrent units learning the dynamics of moving targets across consecutive frames to distinguish between pedestrians, cyclists, and vehicles [46].

Convolutional LSTM (ConvLSTM) networks represent a specialized architecture that has shown particular promise for radar sequence modeling [51]. ConvLSTM extends the standard LSTM by replacing the fully connected operations with convolutional operations, allowing it to capture spatio-temporal patterns while preserving the spatial structure of the data [51]. Wang et al. [51] employed ConvLSTM as the core component of RODNet, enabling the network to maintain a memory of object appearances and movements through sequences of radar frames. This approach proved especially valuable for distinguishing between object classes with similar spatial signatures but different motion patterns, such as pedestrians and cyclists [51].

Transformer based approaches have recently emerged as powerful alternatives for sequence modeling in radar data [48]. The self-attention mechanism in transformers can capture dependencies regardless of temporal distance, potentially modeling longer range temporal patterns than recurrent architectures [48]. [48] introduced a spatiotemporal transformer for radar sequence processing that employed separate attention mechanisms for spatial and temporal dimensions. This factorized approach reduced computational complexity while effectively capturing both spatial relationships within frames and temporal patterns across frames.

Three dimensional convolutional networks (3D CNNs) offer another approach to joint spatiotemporal modeling [51]. By applying 3D convolutions across both spatial dimensions and time, these networks can directly learn spatio-temporal features from radar sequences [51]. Wang et al. [51] employed 3D CNNs for processing sequences of

radar Range-Azimuth-Doppler tensors, enabling the network to capture complex motion patterns across consecutive frames. This approach was particularly effective for objects with distinctive dynamic signatures, such as pedestrians with their characteristic limb movements [51].

The choice of sequence length represents an important consideration in radar temporal modeling [46, 51]. Short sequences may fail to capture sufficient temporal context for distinguishing between object classes with similar spatial characteristics but different motion patterns. Conversely, excessively long sequences increase computational requirements and may introduce irrelevant information when object behavior changes. Wang et al. [51] conducted ablation studies on sequence length for radar object detection, finding that performance generally improved with longer sequences up to a certain point, beyond which returns diminished. They determined that 16 frames, corresponding to approximately 0.5 seconds of radar data, provided an effective balance for their application.

3.6.1.2 Exploitation of Temporal Coherence

Temporal coherence in radar data refers to the consistency and predictable evolution of radar signatures over time [125]. Exploiting this coherence enables more robust detection and classification by integrating information across frames, potentially compensating for noise or occlusions in individual frames [125].

Multiframe integration represents a fundamental approach to leveraging temporal coherence [115]. By coherently combining radar measurements across multiple frames, this technique can enhance the signal-to-noise ratio, revealing objects that might be barely detectable in individual frames [115]. Traditional radar signal processing employs coherent integration through techniques such as Moving Target Indication (MTI) filters, which enhance returns from moving targets while suppressing stationary clutter [115]. In deep learning approaches, multiframe integration can be implemented through architectural components such as 3D convolutions or recurrent layers that aggregate features across the temporal dimension [51].

Temporal consistency constraints have been employed to improve the stability and accuracy of radar detections [118]. These approaches enforce smoothness in detection results across consecutive frames, reducing false detections that appear and disappear rapidly [118]. Schumann et al. [118] implemented a temporal consistency mechanism that penalized sudden changes in object detection confidence between frames, encouraging more stable predictions. This approach proved particularly valuable for challenging scenarios with partial occlusions or varying radar cross-sections, where object visibility might fluctuate across frames [118].

Doppler information provides direct measurements of radial velocity, offering a powerful cue for temporal coherence [108]. Objects maintain relatively consistent velocity profiles over short time intervals, enabling the association of detections across frames based on Doppler signatures [108]. Hyun et al. [108] exploited this property by incorporating Doppler coherence into their pedestrian detection system, tracking the characteristic Doppler patterns of walking humans across multiple frames. Their approach demonstrated improved detection performance compared to methods that treated each frame independently, particularly for distinguishing between pedestrians and static objects with similar spatial signatures [108].

Micro Doppler analysis exploits the fine grained temporal patterns caused by moving components of an object, such as the limbs of a pedestrian or the wheels of a vehicle [117]. These patterns create distinctive signatures in the time frequency domain that can aid in object classification [117]. Cao et al. [117] developed a system for human identification based on radar micro Doppler signatures, using deep convolutional networks to classify the temporal patterns. Their approach captured the unique gait patterns of different individuals, demonstrating the rich information available in detailed temporal analysis of radar returns [117].

Motion consistency tracking leverages the predictable nature of object movements to improve detection robustness [125]. By modeling the expected trajectory of objects based on physical motion constraints, these approaches can maintain tracking through brief

occlusions or detection failures [125]. Kellner et al. [125] employed a physics based motion model for tracking extended objects with high resolution Doppler radar, incorporating constraints such as maximum acceleration and turning rates based on the object class. This approach demonstrated improved tracking continuity through challenging scenarios such as partial occlusions and crossing trajectories [125].

3.7 Domain Specific Challenges

Radar based object detection in automotive environments faces several domain specific challenges that distinguish it from other perception tasks and modalities. These challenges stem from both the inherent limitations of radar technology and the specific characteristics of the automotive context. This section explores the key challenges of low angular resolution, multipath effects and ghost targets, weather and environmental impacts, and class imbalance in the detection of vulnerable road users.

3.7.1 Low Angular Resolution Problems

Low angular resolution represents one of the most significant limitations of automotive radar systems, particularly impacting object detection and classification performance [129]. While radar provides excellent range resolution due to the wide bandwidth available in the 76-81 GHz band, its angular resolution is typically limited by physical constraints related to antenna array size and design [129].

The fundamental physics of radar dictates that angular resolution is proportional to the wavelength divided by the aperture size [114]. For automotive radars operating at 77 GHz with typical aperture dimensions of 5-10 cm, this results in angular resolutions of approximately 1-5 degrees [114]. This contrasts sharply with camera systems that can achieve angular resolutions of 0.01-0.1 degrees [114]. The limited angular resolution becomes particularly problematic at longer ranges, where a few degrees of angular uncertainty can translate to several meters of lateral position ambiguity [129].

This resolution limitation manifests in several ways that complicate object detection [129]. Extended objects like vehicles often appear as collections of scattered points rather than coherent shapes, making it challenging to determine object boundaries [129]. Closely spaced objects may become indistinguishable, appearing as a single merged reflection [129]. The precise lateral positioning of objects becomes increasingly uncertain with distance, complicating tasks such as lane assignment and trajectory prediction [129].

Various approaches have been developed to address the challenge of limited angular resolution [4, 130]. Super resolution techniques attempt to exceed the theoretical resolution limits through sophisticated signal processing [4]. Methods such as MUSIC (Multiple Signal Classification) and ESPRIT (Estimation of Signal Parameters via Rotational Invariance Techniques) can achieve enhanced angular resolution by modeling the signal subspace structure, though they typically require high signal-to-noise ratios to perform reliably [4]. Bilik et al. [4] demonstrated that such techniques could improve angular resolution by a factor of 2-3 compared to conventional beamforming, enabling better separation of closely spaced objects.

MIMO (Multiple-Input Multiple-Output) radar configurations represent another approach to enhancing angular resolution [113]. By employing multiple transmit and receive antennas with appropriate spacing, MIMO radars can create a virtual array with effectively larger aperture, improving angular resolution without increasing physical dimensions [113]. Sun et al. [113] explored various MIMO waveform designs for automotive applications, demonstrating that carefully designed MIMO configurations could achieve angular resolutions approaching 0.5 degrees, substantially better than conventional single-input multiple-output (SIMO) designs.

Temporal integration methods leverage vehicle or object motion to build a more complete picture over time [129]. As either the radar platform or the objects move, different perspectives become available, potentially allowing for improved localization through multiframe integration [129]. Werber et al. [129] developed a radar gridmap representation that accumulated radar measurements over time, compensating for the

limited instantaneous angular resolution through temporal information fusion. Their approach demonstrated improved object boundary definition and separation of closely spaced objects compared to single frame processing.

Deep learning approaches have shown promise in addressing angular resolution limitations by learning to extract more information from limited resolution data [9]. These methods may recognize patterns in the radar returns that suggest object boundaries even when they're not explicitly resolved [9]. Major et al. [9] trained a 3D CNN on Range-Azimuth-Doppler tensors to detect vehicles despite limited angular resolution, demonstrating that neural networks could effectively leverage subtle patterns in the radar data that might be missed by conventional processing.

Sensor fusion represents a complementary strategy, combining radar with sensors that offer better angular resolution, such as cameras or LiDAR [126]. In these approaches, radar provides reliable range and velocity information, while the other sensors contribute more precise angular positioning [126]. Lekic and Babic [126] demonstrated a GAN based fusion approach that aligned radar and camera features, effectively compensating for the radar's angular resolution limitations with the camera's superior lateral resolution.

Despite these advances, low angular resolution remains a fundamental challenge for automotive radar object detection, particularly for applications requiring precise lateral positioning or separation of closely spaced objects [129]. This limitation underscores the importance of complementary sensing modalities and sophisticated processing techniques in automotive perception systems [129].

3.7.2 Multipath Effects and Ghost Targets

Multipath effects and ghost targets represent significant challenges for radar-based object detection in automotive environments [127]. In real-world scenarios, electromagnetic waves returning from target vehicles can reflect off surfaces such as parked cars, metallic road barriers, or large street signs, creating non-existent "ghost targets" that could lead to false decisions by driver assistance systems [127].

Ghost targets occur when radar signals follow indirect paths involving reflections from both the actual target and environmental surfaces [127]. Roos et al. [127] developed a mathematical model to describe ghost target parameters, showing that when a real target at position P is detected via reflection from a perfect electric wall at distance a, the ghost target appears at a different range and angle determined by geometric relationships. These ghost detections possess realistic radar signatures, making them difficult to identify using conventional filtering methods [127].

The key insight from Roos et al. [127] is that ghost targets exhibit a fundamental inconsistency between their apparent orientation and motion state. For a real vehicle, the orientation estimated from radar reflections should align with the angle of its motion vector. However, for ghost targets created by multipath reflections, this alignment breaks down due to the altered Doppler characteristics resulting from the indirect signal path [127].

To exploit this characteristic, Roos et al. [127] proposed a ghost identification method based on analyzing the Doppler distribution across multiple scattering centers on the target. Using high-resolution radar with 5.4 cm/s velocity resolution, they could detect multiple scattering points on each vehicle and estimate both the vehicle's orientation (using a box model fitting algorithm) and its motion vector [127]. For linear motion, a single radar sensor suffices, but for turning vehicles with non-zero yaw rate, two radar sensors at different positions are required to accurately estimate the complete motion state [127].

The mathematical framework accounts for both linear and rotational motion. For a turning vehicle with yaw rate ω, each point Q on the vehicle has velocity altered according to its distance from the rotation center (typically the rear axle) [127]. By solving an overdetermined system of equations using Doppler measurements from multiple scattering centers observed by two sensors, the complete motion state can be recovered [127].

Simulation results demonstrated the effectiveness of this approach. For a real target with 75° orientation, the estimated orientation (74°) closely matched the motion vector angle

(76°) [127]. In contrast, a ghost target showed significant discrepancy, with 98° orientation but only 59° motion vector angle [127]. This mismatch persisted even for turning vehicles, where real targets maintained consistency between orientation and motion while ghost targets exhibited large deviations [127].

The approach is particularly relevant for next-generation automotive radar systems operating at 76.45 GHz with 2 GHz bandwidth, which provide sufficient resolution to resolve multiple scattering centers per vehicle [127]. Unlike approaches requiring prior knowledge of reflecting surface positions (common in through-the-wall imaging applications), this method works without environmental maps, making it suitable for dynamic automotive scenarios [127].

While Roos et al. [127] focused specifically on the motion-orientation mismatch criterion, they noted that ghost target identification remains challenging in complex environments. The presence of reflecting surfaces cannot be avoided in automotive scenarios, making robust ghost detection essential for safe operation of driver assistance systems [127]. Their physics-based approach provides a principled method for ghost identification that complements other radar signal processing techniques.

3.7.3 Weather and Environmental Impacts

One of radar's key advantages for automotive applications is its robustness to adverse weather conditions compared to optical sensors such as cameras and LiDAR [3]. However, while radar performance is less degraded by challenging environmental conditions, it is not entirely immune to their effects. Understanding these impacts is crucial for developing reliable detection systems that perform consistently across diverse operating conditions.

Precipitation in the form of rain, snow, or hail can influence radar performance through several mechanisms [3]. Water droplets and ice particles create additional reflections that appear as clutter in the radar returns, potentially masking legitimate targets. Heavy precipitation can also attenuate the radar signal, reducing the effective detection range.

The specific impact depends on the radar frequency, with higher frequencies experiencing greater attenuation. Bijelic et al. [3] conducted comparative experiments across sensing modalities in adverse weather, finding that while 77 GHz automotive radar experienced some performance degradation in heavy precipitation, it maintained considerably better detection capabilities than cameras or LiDAR, which suffered severe degradation or complete failure.

Fog and mist present minimal challenges for radar detection due to the small size of water droplets relative to the radar wavelength. Unlike optical sensors that experience significant scattering and absorption in foggy conditions, radar signals at 77 GHz propagate through fog with negligible attenuation. This advantage makes radar particularly valuable for maintaining perception capabilities in foggy environments where visibility for human drivers and optical sensors is severely limited [3]. Bijelic et al. demonstrated that radar detection performance remained virtually unchanged in dense fog conditions that rendered camera and LiDAR sensors effectively blind beyond a few meters.

Temperature variations can affect radar performance through changes in atmospheric refraction and the dielectric properties of materials. Extreme temperature conditions may also impact the radar hardware itself, affecting sensitivity and potentially introducing calibration drift. Modern automotive radar systems incorporate temperature compensation mechanisms, but significant thermal gradients or extreme conditions can still influence detection performance. Chadwick et al. [128] observed subtle changes in radar detection patterns across different ambient temperatures, noting that very cold conditions (-20°C and below) resulted in slightly reduced detection ranges for distant targets.

Road surface conditions represent another environmental factor affecting radar performance. Wet roads can create strong specular reflections that increase ground clutter, potentially masking low height objects. Ice and snow accumulation can change the reflectivity of the road surface and roadside objects, altering their radar signatures.

Schumann et al. [118] found that wet road surfaces increased the false positive rate for low height object detection by approximately 12% compared to dry conditions, necessitating more sophisticated clutter filtering approaches for reliable operation.

Seasonal variations in vegetation and ground cover can impact radar returns from the environment [118]. Dense foliage may attenuate radar signals and create additional clutter, while seasonal changes in ground moisture affect ground reflectivity. These variations can alter the background against which objects must be detected, potentially requiring adaptive processing approaches. Schumann et al. [118] noted seasonal variations in detection performance for objects partially obscured by roadside vegetation, with denser summer foliage creating more challenging conditions than winter scenarios with bare branches.

To address these environmental challenges, several approaches have been developed [3, 131]. Adaptive CFAR techniques adjust detection thresholds based on estimated clutter conditions, providing more consistent performance across varying environments [118]. Weather specific processing modes employ different parameter settings or algorithms tailored to particular conditions, such as increased clutter filtering during heavy rain [118]. Multisensor fusion leverages the complementary strengths of different sensing modalities, with radar providing reliable detection in conditions challenging for optical sensors, and cameras or LiDAR contributing higher resolution in favorable conditions [3].

Machine learning approaches have shown promise for developing detection systems robust to environmental variations. By training on diverse datasets that include various weather conditions and environments, these models can learn to extract relevant features despite environmental differences. Bijelic et al. [3] demonstrated a deep fusion approach that maintained consistent detection performance across normal and adverse conditions by learning modality specific reliability patterns and adjusting the fusion weights accordingly.

While radar generally maintains better performance than optical sensors in adverse conditions, the degree of robustness varies across different radar systems and

processing approaches [3]. High end automotive radars with greater transmit power, sophisticated antenna designs, and advanced signal processing typically demonstrate better environmental robustness than simpler systems [3]. This variability highlights the importance of comprehensive testing across diverse conditions and the development of adaptive methods that can maintain reliable performance regardless of environmental challenges [3].

3.7.4 Class Imbalance and Detection of Vulnerable Road Users

The detection of vulnerable road users (VRUs), including pedestrians and cyclists, represents a critical capability for automotive perception systems, yet it presents significant challenges for radar based approaches [119]. These challenges stem from both the radar characteristics of VRUs and the class imbalance inherent in typical driving scenarios.

Vulnerable road users typically present weaker and more variable radar signatures compared to vehicles [119]. The radar cross-section (RCS) of a pedestrian is approximately 100-1000 times smaller than that of a passenger car, resulting in weaker reflections that may be difficult to distinguish from noise or clutter . Furthermore, the RCS of pedestrians and cyclists varies significantly with aspect angle, posture, and clothing, creating inconsistent detection patterns. Prophet et al. [119] characterized the RCS variations of pedestrians across different postures and orientations, finding variations of up to 10 dB depending on whether the pedestrian was facing toward, away from, or perpendicular to the radar. These variations complicate the development of reliable detection algorithms specific to VRUs.

Micro Doppler signatures provide valuable information for identifying VRUs but require specialized processing to extract effectively. The movement of limbs during walking or cycling creates characteristic patterns in the time frequency domain that can help distinguish VRUs from static objects or vehicles. However, these signatures are often subtle and can be masked by stronger returns from the main body or nearby objects . Cao et al. [117] demonstrated that deep learning approaches could effectively extract

and classify these micro Doppler patterns, achieving over 90% accuracy in distinguishing between pedestrians, cyclists, and vehicles based solely on their motion signatures.

Class imbalance in training data presents another significant challenge [119]. In typical driving scenarios, the radar observes far more vehicles than pedestrians or cyclists, leading to datasets dominated by vehicle examples [119]. This imbalance can bias learning based detection systems toward the majority class, resulting in poor performance for the under represented VRU classes [119]. The imbalance is further exacerbated by the fact that vehicles generally present stronger and more consistent radar signatures, making them easier to detect and thus potentially reinforcing the bias during training.

Several approaches have been developed to address these challenges and improve VRU detection performance [132, 145, 133]. Specialized radar configurations with higher angular resolution and sensitivity can provide more detailed measurements of VRUs. Prophet et al. [119] employed a 79 GHz radar with 1° angular resolution to capture more detailed spatial patterns of pedestrians, enabling better discrimination from other object types. Their system achieved 85% detection recall for pedestrians at ranges up to 50 meters, significantly better than conventional automotive radars with lower resolution.

Temporal integration methods leverage the distinctive motion patterns of VRUs over time. By tracking radar returns across multiple frames, these approaches can identify the characteristic gait patterns of pedestrians or the pedaling patterns of cyclists, providing additional discriminative features beyond the spatial signature of a single frame. Hyun et al. [108] developed a coherent phase difference method that tracked the periodic motion patterns of pedestrians across radar frames, achieving improved detection performance particularly for partially occluded individuals.

Sampling strategies to address class imbalance include oversampling minority classes, undersampling majority classes, or generating synthetic examples of under represented classes. These approaches aim to create more balanced training datasets that give appropriate weight to all object categories. Prophet et al. [119] implemented a weighted sampling approach for their pedestrian classification system, ensuring that the network

received a similar number of pedestrian and non-pedestrian examples during each training epoch despite the imbalance in the original dataset. This approach improved pedestrian detection recall from 68% to 82% compared to training with the unbalanced dataset.

Loss function modifications represent another approach to addressing class imbalance . Focal loss, introduced by Lin et al. [67], down weights the contribution of well classified examples to the overall loss, focusing the learning process on the more challenging cases. This approach is particularly valuable for radar based VRU detection, where the network might otherwise focus on the easier vehicle detection task at the expense of VRU performance. Several studies have applied variants of focal loss to radar object detection, reporting significant improvements in recall for minority classes without substantial degradation in performance for majority classes.

Data augmentation techniques specific to radar VRU signatures can help enrich training datasets and improve generalization. These techniques include adding synthetic noise, varying the amplitude of reflections, and simulating different aspect angles and postures. Prophet et al. [119] employed a combination of geometric transformations and signal level augmentations to expand their pedestrian radar signature dataset, demonstrating improved robustness to variations in pedestrian presentation and radar viewing angle.

Despite these advances, the detection of vulnerable road users with radar alone remains challenging compared to vehicle detection [119]. Performance metrics from recent studies indicate that while vehicle detection with radar can achieve recall rates of 90-95% at ranges up to 100 meters, pedestrian detection typically achieves 70-85% recall at ranges of 30-50 meters [119]. This performance gap highlights the importance of multisensory approaches for comprehensive VRU protection, combining radar's all weather capability with the higher resolution and richer semantic information provided by cameras or LiDAR.

Research in this area continues to evolve, with increasing focus on specialized architectures and training methodologies tailored to the unique challenges of VRU detection with radar. As radar technology advances toward higher resolution and

sensitivity, and as deep learning techniques for handling class imbalance and weak signatures mature, the performance gap between vehicle and VRU detection is expected to narrow, enhancing the overall safety capabilities of automotive perception systems [119].

3.8 Emergence of End-to-End Detection

Recent years have witnessed a paradigm shift in radar based object detection with the emergence of end-to-end approaches that fundamentally reimagine the radar processing pipeline [53, 54]. These methods challenge the conventional wisdom of sequential radar signal processing, exploring the potential of learning based approaches to extract more information directly from minimally processed radar data [53]. This section examines the trend toward bypassing traditional radar processing chains, approaches for processing raw Analog-to-Digital Converter (ADC) data, and the development of learned representations for radar data.

3.8.1 Bypassing Traditional Radar Processing Chains

Traditional radar processing follows a sequential pipeline comprising multiple distinct stages, including range-FFT, Doppler-FFT, angle estimation, CFAR detection, clustering, and tracking [110]. While this approach is well established and theoretically grounded, it suffers from several limitations. Each processing stage makes decisions based on limited information, potentially discarding valuable signal components that fall below thresholds or don't conform to expected patterns. Furthermore, errors propagate through the pipeline, with early stage misdetections or false alarms affecting all subsequent processing [110]. The parameter tuning required for strong performance at each stage becomes increasingly complex for diverse operating environments.

End-to-end detection approaches seek to bypass some or all of these traditional processing stages, replacing them with learned models that map directly from less processed radar data to detection outputs [53]. Rather than applying fixed algorithmic

transformations with hand tuned parameters, these approaches learn the most effective transformations directly from annotated data [53]. This paradigm shift is analogous to the evolution in computer vision from hand crafted feature extractors like HOG [59] and SIFT to learned CNN features [23], which dramatically improved performance across various vision tasks [53].

Rebut et al. [53] demonstrated one of the pioneering efforts in this direction, developing a deep neural network architecture that operated directly on range-Doppler maps before CFAR thresholding. Their approach eliminated the need for CFAR detection, clustering, and hand designed feature extraction, instead learning to map from the continuous valued range-Doppler map to object detections. Experiments on a highway dataset showed improved detection of distant vehicles compared to a traditional pipeline, with the network learning to extract weak but consistent patterns that would have been eliminated by standard CFAR thresholding.

The potential advantages of bypassing traditional processing chains include more effective information extraction, adaptive processing based on context, and joint optimization across tasks that were previously treated separately [53, 54]. By learning from annotated data, these approaches can discover signal patterns that might be overlooked by conventional processing, particularly for challenging object classes like pedestrians with weak and variable radar signatures [53]. Additionally, a unified model can potentially adapt its processing based on the specific environment and task requirements, rather than applying the same fixed transformations in all scenarios.

However, these approaches also face significant challenges [53]. They typically require large amounts of annotated data, which can be difficult and expensive to collect for radar. The black box nature of deep learning models makes it challenging to analyze failure cases or provide performance guarantees, which are important considerations for safety critical automotive applications. Additionally, computational requirements may be higher than for traditional processing, potentially complicating deployment on resource constrained automotive platforms.

3.8.2 Raw ADC Data Processing Approaches

The most radical end-to-end approaches operate directly on raw or minimally processed ADC data, the direct output of the radar receiver before any significant signal processing. This approach represents the earliest possible intervention point in the radar processing chain, potentially enabling the extraction of information that might be lost in conventional processing.

Raw ADC data from automotive radar systems typically consists of complex valued samples collected from multiple receive channels across multiple chirps [110]. For a typical configuration with 4 receive antennas, 512 samples per chirp, and 128 chirps per frame, this results in a data tensor of dimension 4×512×128, with each element being a complex number. This high dimensional, complex valued data presents both opportunities and challenges for direct processing approaches.

Rebut et al. [53] introduced one of the first deep learning systems for processing raw high definition radar data for multitask perception. Their approach employed a complex valued CNN that operated directly on the ADC samples, learning to extract relevant features for object detection, classification, and velocity estimation. The network architecture incorporated specialized layers for processing complex valued data, including complex convolutions and complex batch normalization. This approach demonstrated superior performance compared to traditional processing for detecting small objects like pedestrians and cyclists, particularly in cluttered environments where conventional CFAR detection struggled to distinguish target returns from background.

Several key challenges must be addressed for effective raw ADC data processing [53]. The high dimensionality and complex valued nature of the data require specialized network architectures and significantly more computational resources than processing at later stages. The limited availability of annotated raw radar data presents another obstacle, as most existing datasets provide preprocessed radar data or detections. Additionally, raw data processing must contend with various hardware specific effects and calibration issues that are typically handled in the traditional processing chain.

To address these challenges, researchers have explored various approaches [53]. Complex valued neural networks that explicitly handle the real and imaginary components of radar signals have shown promise for preserving phase information that is crucial for angular resolution. Domain adaptation techniques help bridge the gap between simulation, where unlimited training data can be generated, and real-world deployment. Efficient architectural designs, including factorized convolutions and progressive downsampling, help manage the computational demands of processing high dimensional raw data.

Despite these challenges, raw ADC data processing offers several compelling advantages [53]. It provides access to the complete signal information, enabling the potential discovery of subtle patterns that might be lost in conventional processing [53]. It allows for adaptive processing that can optimize differently for various object types or environments, rather than applying the same fixed transformations to all data [53]. Perhaps most significantly, it offers the potential for super resolution capabilities that exceed the theoretical limits of conventional processing by learning to exploit signal characteristics specific to the automotive context.

As research in this area continues to advance, raw ADC data processing is likely to become increasingly practical for automotive applications, potentially offering significant performance improvements for challenging perception tasks. The evolution toward earlier intervention points in the radar processing chain represents a fundamental rethinking of radar signal processing, shifting from a physics based paradigm with fixed algorithms to a data driven approach that learns suitable transformations directly from examples.

Summary

In this chapter we provided a comprehensive examination of existing approaches to radar-based object detection. Traditional radar processing methods were analyzed, including CFAR detection algorithms, clustering approaches such as DB-SCAN, and

tracking methods based on Kalman filtering and the Hungarian algorithm. The evolution of CNN-based approaches for radar object detection was examined from early ResNet-based architectures through specialized designs such as RODNet, which introduced cross-supervised learning frameworks for radar detection. Transformer-based approaches were investigated, including T-RODNet, TransRAD, and Mask-RadarNet, which leverage self-attention mechanisms to capture long-range dependencies in radar data. The chapter also reviewed MetaFormer architectures such as E-RODNet, which demonstrated that the macro-level architectural patterns of transformers contribute as significantly to their effectiveness as the specific attention mechanism. The literature review done in this chapter identified critical research gaps in achieving balance between detection performance with computational efficiency, setting the contributions of this thesis in a crucial position within the broader research landscape.

Chapter 4

Methodology

4.1 Radar Datasets for Road User Detection

The development and evaluation of radar based object detection algorithms require comprehensive, well annotated datasets that capture the diversity of scenarios encountered in real-world driving conditions. Public datasets serve as benchmarks for comparing different approaches and accelerate research by providing common ground for experimentation without the need for expensive data collection campaigns. This section examines key radar datasets for road user detection, beginning with the principal dataset used in this research, the CRUW dataset, and continuing with other datasets, RADDet and CARRADA, on which future work will continue due to availability of the necessary raw radar data representations.

4.1.1 CRUW Dataset

The Camera-Radar of University of Washington (CRUW) dataset, introduced by Wang et al. [51], represents one of the first publicly available synchronized camera-radar datasets specifically designed for object detection research in autonomous driving. This dataset has become a valuable resource for developing and evaluating radar based detection algorithms, particularly those leveraging cross-modal supervision.

4.1.1.1 Data Collection Methodology

The CRUW dataset was collected using a sensor platform consisting of a stereo camera pair and a 77 GHz FMCW radar [51]. The radar employed was an AWR1843 BOOST automotive radar from Texas Instruments, configured with two transmit antennas and four receive antennas in a MIMO configuration, providing a total of eight virtual channels. This setup achieved a range resolution of approximately 0.15 meters and an angular resolution of about 15 degrees. The stereo camera system consisted of two synchronized global shutter cameras with a baseline of 0.3 meters, providing both visual information and depth estimation capabilities.

The data collection vehicle was driven through various environments in Seattle, Washington, covering a diverse range of scenarios [51]. The collection routes were carefully selected to include different road types (urban streets, residential areas, highways), lighting conditions (daytime, nighttime, dawn/dusk), and weather states (clear, rainy, foggy). This diversity was deliberately incorporated to ensure the dataset captured the variability encountered in real-world driving situations.

Data was recorded in sequence format rather than as isolated frames, with each sequence typically lasting 10-20 seconds. This approach preserved the temporal context necessary for developing algorithms that exploit motion information. The sensors were precisely synchronized using hardware triggers, ensuring that camera images and radar frames were captured simultaneously with minimal temporal offset [51]. This synchronization is crucial for cross modal supervision approaches that transfer labels between heterogeneous sensing systems.

The collection methodology also included careful calibration between the camera and radar sensors [51]. A specialized calibration procedure was developed using corner reflectors placed at known positions, enabling precise mapping between the camera image space and the radar's range-azimuth coordinates. This calibration information was provided alongside the dataset, facilitating research on sensor fusion and cross-modal learning.

4.1.1.2 Format and Annotations

The CRUW dataset is organized into sequences, with each sequence containing synchronized camera images and radar frames [51]. The camera data consists of stereo image pairs in standard image format (PNG), with a resolution of 1920x1200 pixels and a frame rate of 30 Hz. The radar data is provided in two formats: the raw ADC data (complex I/Q samples) and preprocessed range-azimuth (RA) maps. The RA maps are generated by performing range-FFT and angle-FFT on the raw data, followed by noncoherent integration across chirps to produce 2D power maps with dimensions of 128×128, corresponding to a field of view of approximately 50 meters in range and 90 degrees in azimuth. The radar operates at a frame rate of 30 Hz, matching the camera frame rate for straightforward synchronization.

The annotation approach in CRUW is notable for its cross modal methodology [51]. Rather than directly annotating the radar data, which is challenging even for human experts due to its low resolution and unintuitive nature, the annotations are primarily performed in the image domain [51]. Objects are labeled in the camera images with bounding boxes and object class information, covering three main categories: pedestrian, cyclist, and vehicle. These image domain annotations are then projected into the radar domain using the calibration parameters and depth information from the stereo camera system].

This projection results in what the authors term "conformal annotations" for radar [51]. Each labeled object is associated with a region in the range-azimuth map, with additional confidence weighting based on factors such as visibility, occlusion, and distance. This confidence weighted approach acknowledges the inherent uncertainty in transferring labels between modalities with different physical characteristics and resolution capabilities.

The dataset includes training, validation, and test splits, with approximately 50,000 frame pairs in total [51]. The training set comprises sequences that cover various scenarios and conditions, while the validation and test sets are designed to evaluate generalization

across different locations and conditions. Alongside the sensor data and annotations, the dataset provides metadata for each sequence, including GPS coordinates, timestamps, weather conditions, and location types.

4.1.1.3 Strengths and Limitations

The CRUW dataset offers several notable strengths that have contributed to its widespread adoption in radar based detection research [51]. First, the synchronized multimodal nature of the dataset, combining camera and radar data, facilitates research on sensor fusion and cross modal learning approaches. The precise calibration between modalities enables accurate projection of information between the camera and radar domains, supporting the development of cross supervised learning methods.

The diversity of collection scenarios represents another key strength [51]. By including various environments, lighting conditions, and weather states, the dataset enables the development and evaluation of algorithms that must operate robustly across diverse situations. The inclusion of challenging weather conditions, such as rain and fog, is particularly valuable for highlighting radar's advantages in adverse environments where optical sensors may struggle.

The temporal sequence format of CRUW preserves important motion context that is lost in datasets consisting of isolated frames [51]. This temporal information enables the development of algorithms that exploit the coherence of object movements over time, which is particularly valuable for distinguishing between object types based on their characteristic motion patterns.

The provision of both raw ADC data and preprocessed RA maps offers flexibility for researchers exploring different entry points into the radar processing chain [51]. Those focused on end-to-end approaches can work directly with the minimally processed ADC data, while others can build upon the preprocessed RA maps for higher level detection algorithms.

Despite these strengths, the CRUW dataset has several limitations that should be considered when interpreting results based on this data [51]. The angular resolution of the radar system (approximately 15 degrees) is relatively low compared to more recent automotive radar systems, which can achieve resolutions of 1-5 degrees. This limitation makes fine grained object discrimination and precise localization more challenging, potentially limiting the performance achievable with this dataset. It is worth noting that the sequences were selected or choreographed in such fashion to minimize the effect of the low angular resolution, distancing itself from real world type data.

The annotation methodology, while innovative, introduces certain uncertainties [51]. The projection from camera to radar space relies on depth estimation from stereo vision, which can contain errors, particularly for distant or partially occluded objects. Additionally, the fundamental differences in what each sensor measures mean that not all radar reflections correspond cleanly to visible objects in the camera view, and vice versa.

The dataset's size, while substantial in the number of frames, is technically smaller than inspired due to the high frame rate, 30 Hz, which exhibits minimal change across neighboring frames, potentially limiting the training of very deep models that require large amounts of data to avoid overfitting [51]. Additionally, the object class distribution exhibits an imbalance typical of real-world driving scenarios, with large variation in the number of vehicle instances, pedestrians and cyclists, which can bias learning based approaches toward the majority class without appropriate countermeasures.

Another limitation lies in the evaluation methodology [51]. The dataset uses a center based detection criterion, where a detection is considered correct if it falls within a certain distance of the ground truth object center. This approach may not fully capture the performance of methods that provide more detailed object extent information, such as oriented bounding boxes or segmentation masks.

4.1.1.4 Notable Results Achieved with this Dataset

The CRUW dataset has served as the foundation for several influential works in radar based object detection [48, 51, 121]. The original RODNet paper by Wang et al. [51] established baseline performance on the dataset using their proposed convolutional recurrent architecture. Their system achieved average precision (AP) scores of 0.77 for vehicles, 0.58 for pedestrians, and 0.40 for cyclists, demonstrating the feasibility of radar only detection while highlighting the challenges of detecting smaller road users with weaker radar signatures.

The temporal modeling component of RODNet proved particularly valuable, with experiments showing a significant performance drop when processing frames independently rather than in sequence [51]. This finding underscored the importance of motion information for disambiguating between object types in radar data, where spatial resolution alone may be insufficient.

Building upon the RODNet framework, Chen et al. [48] introduced T-RODNet, which replaced the convolutional recurrent architecture with a transformer based design. This approach achieved improved performance across all object categories, with particularly notable gains for pedestrians and cyclists. The AP scores increased to 0.83 for vehicles, 0.67 for pedestrians, and 0.52 for cyclists, representing relative improvements of 8%, 16%, and 30% respectively compared to the original RODNet. These results demonstrated the effectiveness of transformer architectures in capturing the complex spatiotemporal patterns in radar data.

Wang et al. [51] explored multitask learning on the CRUW dataset, jointly performing object detection, classification, and velocity estimation. Their spatiotemporal attention network achieved further improvements in detection performance while also providing accurate velocity estimates that could benefit downstream tracking and prediction task. Their approach demonstrated that learning multiple related tasks simultaneously could enhance the representations learned from radar data, improving performance across all tasks.

Several works have investigated fusion approaches using the CRUW dataset [8, 127]. These studies typically used the radar as a complementary sensor to enhance detection robustness in challenging conditions such as low light or adverse weather . The synchronized nature of the dataset made it particularly valuable for developing and evaluating different fusion strategies, from early to late fusion.

More recent work has explored end-to-end approaches on the CRUW dataset, working with the provided raw ADC data rather than the preprocessed RA maps [53]. These approaches have shown promise for extracting more information from the radar signals, particularly for challenging object classes like pedestrians and cyclists. However, they typically require more computational resources and larger training datasets to achieve better performance.

The CRUW dataset continues to serve as an important benchmark in radar based detection research, providing a common ground for comparing different approaches [60, 64, 121]. Its inclusion of synchronized camera data alongside radar measurements has proven particularly valuable for developing cross-modal learning techniques that leverage the strengths of both sensing modalities [48, 51, 121].

4.1.2 RADDet Dataset

The RADDet (Range-Azimuth-Doppler based Radar Object Detection) dataset, introduced by Zhang et al. [123], provides a substantial advancement in radar based perception, aimed at enabling deep learning models to detect dynamic road users using full resolution radar data. It offers a three dimensional representation of radar measurements that incorporates Doppler velocity in addition to traditional range and azimuth information. This dataset is among the first to include full Range-Azimuth-Doppler (RAD) tensors with corresponding annotations including 3D bounding boxes and object class labels, bridging a significant gap in public radar datasets available to the research community.

4.1.2.1 Range-Azimuth-Doppler Representation

The dataset is built upon a 3d RAD representation, capturing radar returns in a dense 3D tensor with dimensions (256, 256, 64), corresponding to range bins, azimuth bins, and Doppler velocity bins, respectively [123]. This approach retains all available spatial and motion information, preserving a richer signal structure than traditional Range-Azimuth (RA) maps or radar point clouds. Each voxel in the RAD tensor contains the magnitude of the complex radar signal after a standard sequence of signal processing transformations.

The RAD tensor is generated from raw ADC data through a typical FFT based signal chain. The first stage is a Range-FFT across the fast time samples of each chirp, transforming the time domain signal into distance information. A second stage applies Doppler-FFT across chirps for each range bin to reveal velocity characteristics. Finally, a zero-padded FFT is applied across the antenna channels to extract angular location using beamforming [123]. This process produces a full 3D tensor encoding range, azimuth, and Doppler information in a coherent format, offering deep learning models access to both motion and spatial patterns within the radar's field of view.

Inclusion of the Doppler dimension enables separation of closely spaced objects with different velocities, helps identify motion signatures characteristic of different object classes, and supports suppression of clutter centered around zero Doppler. It also offers robustness in scenes with overlapping objects, such as a pedestrian walking beside a vehicle, or a cyclist overtaking another vehicle. These distinctions are often lost in 2D projections or in sparse point cloud formats. As the RAD tensor retains raw signal magnitude values, both strong reflections from vehicles and weaker ones from pedestrians or cyclists remain visible for learning based methods [123].

The dataset focuses exclusively on dynamic road users; stationary objects are intentionally excluded. Labels include not only the position and class of each object but also motion information extracted from the Doppler signature. This labeling strategy supports tasks such as motion prediction and tracking, as well as static object

suppression. The dataset is thus designed to enable end-to-end models that can simultaneously localize, classify, and infer motion from radar input.

4.1.2.2 Data Preparation and Preprocessing

Data for RADDet was collected using a static sensor platform set up along sidewalks facing active roadways, rather than a moving vehicle. The radar used is the Texas Instruments AWR1843-BOOST, a 77 GHz FMCW radar featuring a 4 transmitter by 4 receiver antenna configuration. This MIMO setup provides 16 virtual channels, allowing improved angular resolution through spatial diversity. Alongside the radar, a stereo camera system (DFK 33UX273) was used to provide visual data for annotation and calibration purposes [123].

The raw ADC data captured by the radar has a base shape of (256, 8, 64). After applying zero padding along the azimuth axis, the processed RAD tensors reach the final dimensionality of (256, 256, 64). This preprocessing includes the FFT stages described earlier and additional steps such as calibration to correct hardware induced imbalances or phase misalignments. The result is a clean, normalized 3D tensor that retains the spatial and Doppler fidelity of the radar signal.

A central feature of RADDet's preparation is its novel auto annotation pipeline. First, object detection candidates are extracted using 2D Ordered-Statistic CFAR (OS-CFAR) applied to the Range-Doppler (RD) slices, followed by DBSCAN clustering to group detection candidates [123]. These radar only object proposals are used to locate dynamic objects based on Doppler coherence, which is especially strong for vehicles.

To assign class labels and refine object boundaries, the dataset leverages stereo vision. Depth maps are computed from rectified image pairs. Instance segmentation masks are generated via a pretrained Mask R-CNN applied to the left image. These masks are projected onto the depth maps to generate 3D point clouds for each object instance.

These point clouds are then transformed into the radar's bird's eye view using a calibrated projection matrix obtained through extrinsic sensor alignment with a trihedral corner reflector [123].

This fusion of radar and stereo vision allows RADDet to annotate approximately 75% of all objects appearing in the radar field of view. Remaining annotations are added or corrected manually by experts, particularly in cases where stereo vision fails due to field of view limitations or segmentation inaccuracies. The final annotations include 3D bounding boxes in RAD space, object type (vehicle, pedestrian, cyclist), and associated motion information inferred from Doppler data.

Importantly, RADDet does not apply filtering to remove stationary clutter or background reflections. This preserves the full complexity of the radar environment, enabling models to learn to distinguish relevant targets from background structures. Such flexibility supports the development of more robust radar only detection models, especially in cluttered urban scenes [123].

A total of 10,158 annotated radar frames are included in the dataset, covering over 40,000 labeled object instances. The dataset is divided into training, validation, and testing splits using classwise sampling to maintain a representative distribution of object classes across scenes. As is common in real-world driving data, the class distribution is imbalanced, with vehicles comprising the majority of annotations, followed by pedestrians and cyclists [123].

4.1.2.3 Benchmark Results

The dataset was released in tandem with a tailored object detection architecture, also named RADDet, designed specifically to exploit the 3D nature of RAD tensors. The proposed model uses a customized ResNet based backbone referred to as RadarResNet, which processes RAD tensors using 2D convolutions across the Range-Azimuth plane while treating the Doppler axis as input channels [123].

Unlike vision based detection systems, the RADDet model omits a neck stage for multiscale fusion, as radar does not exhibit scale changes due to perspective. Instead, the model features two detection heads: a 3D detection head operating in RAD space and a 2D detection head that predicts bounding boxes in Cartesian coordinates after a nonlinear coordinate transformation of feature maps. The coordinate transform layer maps polar coordinates (r, θ) to Cartesian (x, z) using fully connected layers and residual blocks. Both detection heads follow a YOLO style architecture with anchor based predictions tailored to radar box geometries [123].

Evaluation shows strong results. Using RadarResNet with maxpooling layers, the 3D detection head achieves average precision (AP) of 0.251 at IoU 0.5, with 0.563 AP at IoU 0.3. The 2D detection head performs even better, achieving 0.516 AP at IoU 0.5 and up to 0.727 at IoU 0.3 [123]. The model significantly outperforms projection limited baselines that use only RA slices, especially for vulnerable road users such as pedestrians and cyclists, demonstrating the critical role of Doppler in detecting objects with weak spatial signatures.

In summary, the RADDet dataset and its associated model present a powerful benchmark for learning based radar object detection. With full 3D RAD input, rigorous annotation, and a novel dual head architecture, RADDet enables a new level of fidelity and performance in radar perception research.

4.1.3 CARRADA Dataset

The CARRADA (Camera and Automotive Radar with Range-Angle-Doppler Annotations) dataset, introduced by Ouaknine et al. [121], addresses a critical gap in radar based perception by providing synchronized camera and radar recordings with multilevel annotations. It is specifically designed for the development of object detection, segmentation, and tracking algorithms using raw radar data, and includes annotations across both range-angle and range-Doppler representations.

4.1.3.1 Data Acquisition and Sensor Setup

The dataset was recorded using a stationary acquisition setup consisting of a 77 GHz FMCW radar with a 2 × 4 MIMO antenna configuration (yielding 8 virtual channels) and a high resolution RGB camera. The radar offers a 4 GHz sweep bandwidth, 0.2 m range resolution, and 0.7° angular resolution, covering up to 50 meters and 180° field of view.

Data was collected in a controlled test track environment, with scenarios involving cars, pedestrians, and cyclists performing varied trajectories (e.g., approaching, receding, lateral motion). In total, the dataset contains 30 sequences comprising over 12,000 frames, of which 7,193 are annotated. All frames are synchronized between camera and radar, and data is captured in diverse lighting conditions, though weather is kept consistent.

4.1.3.2 Annotation Methodology

CARRADA's unique strength lies in its multilevel annotation strategy, which includes:

- Sparse points: direct reflections annotated in the radar tensor
- Segmentation masks: per object masks in both range-angle and range-Doppler domains
- Object level bounding boxes: oriented boxes with class labels and instance IDs

Annotation is performed using a semiautomatic pipeline that combines vision based instance detection (via Mask R-CNN and SORT) with geometric projection and Doppler informed radar feature extraction. Camera derived positions and velocities are mapped to radar space and refined using clustering on Direction of Arrival (DoA)-Doppler point clouds. Robust cluster selection is guided by probabilistic modeling and tracking is performed across frames. Manual verification ensures high annotation quality.

4.1.3.3 Applications in Research

CARRADA has been instrumental in advancing radar based deep learning. Its annotations have enabled:

- Point level radar detection alternatives to CFAR
- Instance segmentation via U-Net and multiscale methods
- Transformer based object detection architectures like TransRAD.

The dataset's structure supports multitask learning and domain adaptation studies, particularly for training on clean radar data and transferring to complex environments. A semantic segmentation baseline using FCNs demonstrates promising performance, particularly in the range-Doppler domain where higher resolution is available.

4.1.4 Comparison of Dataset Characteristics

The radar datasets discussed in this section exhibit diverse characteristics that make them suitable for different research objectives. Table 2.1 provides a comparative overview of key attributes across these datasets.

Dataset	Representation	Size	Annotations	Reference Sensor	Unique Strengths	Limitations
CRUW	RA maps, ADC data	~50,000 frames	Cross-modal centroid labels	Camera	Temporal sequences, Cross-modal supervision	Limited angular resolution
RADDet	RAD tensors	~10,000 frames	3D boxes with velocity	Camera reference	Comprehensive RAD representation, Dynamic focus	Smaller size, Controlled scenarios

CARRADA	RA/RD maps, RAD tensors	~7,000 frames	Multilevel: points, masks, boxes	Camera reference	Multilevel annotations, Clean scenarios	Limited environmental diversity

4.2 Evaluation Metrics

The assessment of radar based object detection algorithms requires rigorous evaluation metrics that quantify performance across various aspects of the detection task. Standardized metrics enable objective comparison between different approaches and track progress in the field. This section examines the key metrics used to evaluate radar based object detection, beginning with fundamental object detection metrics.

4.2.1 Object Detection Metrics

4.2.1.1 Precision, Recall, F1-score

The evaluation of object detection algorithms begins with three fundamental metrics: precision, recall, and F1-score [55]. These metrics quantify different aspects of detection performance and provide complementary insights into algorithm behavior.

Precision measures the accuracy of positive predictions, calculating the proportion of detected objects that correspond to actual objects in the scene [55]. Mathematically, precision is defined as:

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$$

In the context of radar based detection, a high precision indicates that the algorithm generates few false alarms, which is particularly important for automotive applications where false detections could trigger unnecessary interventions [67]. However, precision

alone is insufficient for comprehensive evaluation, as a trivial algorithm that makes very few but confident detections could achieve high precision while missing most objects [67].

Recall measures the algorithm's ability to detect all relevant objects, calculating the proportion of actual objects that are successfully detected [55]. Mathematically, recall is defined as:

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

In radar based detection, high recall indicates that the algorithm successfully detects most objects present in the scene, which is crucial for safety-critical applications where missing objects could have severe consequences [67]. However, recall can be trivially maximized by generating numerous detection hypotheses, many of which would be false positives [67].

The F1-score combines precision and recall into a single metric, providing a balanced assessment of detection performance [55]. It is calculated as the harmonic mean of precision and recall:

$$\text{F1-score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

The harmonic mean emphasizes balanced performance, penalizing algorithms that achieve high performance on one metric at the expense of the other [55]. In radar detection, the F1-score helps identify algorithms that provide a good trade off between minimizing false alarms and maximizing the detection of relevant objects [67].

These metrics depend critically on the classification of detections as true positives, false positives, or false negatives, which requires a matching criterion between detections and ground truth objects [55]. For radar based detection, this matching typically employs

spatial overlap criteria, such as Intersection over Union (IoU), or distance based criteria that consider the separation between detection and ground truth centers [67]. The choice of matching criterion can significantly influence the reported metrics and should be considered when comparing different approaches [67].

Several factors complicate the application of these metrics to radar based detection [67]. The inherent uncertainty in radar measurements, particularly in angular dimensions, makes exact spatial matching challenging [67]. Class imbalance in typical driving scenarios, with many more vehicles than vulnerable road users, may skew aggregate metrics without appropriate stratification [67]. Additionally, the relevance of detections may vary with distance and object type, suggesting the need for weighted metrics that prioritize critical detections [67].

To address these challenges, researchers often report precision, recall, and F1-scores separately for different object classes, distance ranges, and occlusion levels [67]. This layered evaluation provides more targeted insights into algorithm performance across various operating conditions. Additionally, confidence weighted versions of these metrics may be employed to assess not only the correctness of detections but also the reliability of the associated confidence scores.

4.2.1.2 mAP and IoU Thresholds

Mean Average Precision (mAP) has emerged as a standard metric for evaluating object detection algorithms across confidence levels and object classes [65]. Unlike precision, recall, and F1-score, which are typically calculated at a specific confidence threshold, mAP summarizes detection performance across the entire precision-recall curve.

The calculation of mAP begins with the determination of Average Precision (AP) for each object class [65]. AP is computed by sampling the precision-recall curve at different recall levels and averaging the resulting precision values. Mathematically, AP can be defined as:

$$AP = \frac{1}{11} \times \sum_{r \in \{0, 0.1, \ldots, 1.0\}} P_{\text{interp}}(r)$$

where $P_{interp}(r)$ is the interpolated precision at recall level r [65]. The interpolation addresses the zigzag nature of the raw precision-recall curve, replacing each precision value with the maximum precision achieved at that recall level or any higher recall level [65]:

$$P_{\text{interp}}(r) = \max_{\tilde{r} \geq r} P(\tilde{r})$$

Once AP is calculated for each class, mAP is obtained by averaging these values across all classes [65]:

$$\text{mAP} = \frac{1}{N} \sum_{i=1}^{N} \text{AP}_i$$

where N is the number of object classes [65]. This approach gives equal weight to each class regardless of its frequency in the dataset, preventing dominant classes from overwhelming the metric [65].

The Intersection over Union (IoU) threshold plays a crucial role in the calculation of detection metrics, defining the criterion for matching detections to ground truth objects [93]. IoU measures the overlap between two bounding boxes as the ratio of their intersection area to their union area [93]:

$$\text{IoU} = \frac{\text{Area of Intersection}}{\text{Area of Union}}$$

A detection is typically considered a true positive if its IoU with a ground truth object exceeds a specified threshold, commonly set at 0.5 or 0.7 for conventional object detection [93]. Higher IoU thresholds demand more precise localization, making the evaluation more stringent.

For radar data, the application of IoU involves several considerations specific to radar representations [64, 167]. In Range-Azimuth (RA) or Range-Doppler (RD) maps, IoU can be calculated directly in the 2D space, similar to image based detection [51]. For Range-Azimuth-Doppler (RAD) tensors, IoU may be calculated in 3D space, incorporating the velocity dimension alongside spatial dimensions [123]. For radar point clouds, IoU calculation may involve generating bounding boxes around point clusters or using alternative metrics like chamfer distance that directly compare point sets.

The resolution differences between radar dimensions complicate IoU calculation. The fine range resolution contrasts with the coarser angular resolution, creating asymmetric uncertainty that may not be well captured by standard IoU . To address this, some researchers employ modified IoU formulations that account for the different resolutions across dimensions, or use separate thresholds for range and angular dimensions.

Several variations of mAP have been developed for comprehensive evaluation [65]. COCO style mAP averages AP values across multiple IoU thresholds (typically from 0.5 to 0.95 in steps of 0.05), providing a more robust assessment of localization accuracy [65]. Average Precision for different object sizes (small, medium, large) helps evaluate performance across the detection range [65]. Class specific AP values identify strengths and weaknesses for different object types, which is particularly important for radar detection where performance can vary significantly between vehicles and vulnerable road users [65].

For radar based detection, researchers have proposed domain specific extensions to mAP [51, 124]. Some approaches incorporate velocity accuracy alongside spatial localization, computing a joint score that considers both position and velocity estimates [123]. Others weight detections based on range, giving greater importance to distant

objects that are typically more challenging to detect with radar [51]. These adaptations aim to align the evaluation metrics with the specific challenges and requirements of radar based detection in automotive environments [51, 124].

4.2.1.3 Detection Evaluation Protocols

Standardized evaluation protocols ensure fair and meaningful comparison between different detection algorithms by specifying the complete procedure for computing metrics [93]. These protocols define aspects such as matching criteria, confidence thresholds, and handling of edge cases that might otherwise vary between implementations [93].

The PASCAL VOC protocol, one of the earliest standardized evaluation frameworks for object detection, established several conventions that remain influential [93]. It employs an IoU threshold of 0.5 for matching detections to ground truth, requires confidence scores for all detections to generate precision-recall curves, and handles duplicate detections by considering only the highest confidence detection that matches a ground truth object [93]. AP is calculated by averaging precision at 11 equally spaced recall levels from 0 to 1 [93].

The COCO evaluation protocol introduced several refinements to provide a more comprehensive assessment [93]. It calculates AP across multiple IoU thresholds (0.5 to 0.95 in steps of 0.05) and reports both the average across thresholds and specific values like AP at IoU 0.5 (AP50) and 0.75 (AP75) [93]. It also layers evaluation by object size, reporting separate AP values for small, medium, and large objects [93]. Additionally, it employs 101 recall points for AP calculation rather than the 11 points used in PASCAL VOC, providing finer granularity [93].

For radar based detection, several domain specific considerations influence evaluation protocols [51, 124, 122]. The inherent uncertainty in radar measurements, particularly in angular dimensions, has led some researchers to adopt adaptive matching criteria that vary with range [51]. For instance, the matching threshold might be tighter for nearby

objects and more relaxed for distant objects, reflecting the range dependent angular resolution of radar [51].

The representation format affects the evaluation approach [51, 124, 122]. For dense representations like Range-Azimuth maps, protocols similar to image based detection can be applied with appropriate modifications for radar characteristics [51]. For point cloud representations, evaluation may employ distance based matching rather than IoU, considering the separation between detection and ground truth centers relative to object size [123]. For segmentation masks in radar data, metrics like mean IoU or Dice coefficient may be employed, similar to image segmentation evaluation [121].

Several radar detection datasets have established their own evaluation protocols [51, 124, 122]. The CRUW dataset employs a center based detection criterion, where a detection is considered correct if its center falls within a specified radius of the ground truth center [51]. This radius adapts based on object class and distance, acknowledging the varying sizes of different road users and the range dependent resolution of radar [51]. The RADDet dataset uses a 3D IoU threshold of 0.5 for spatial matching, with an additional criterion for velocity accuracy that requires estimated velocity to be within 20% of the ground truth [123]. The CARRADA dataset provides multilevel evaluation protocols corresponding to its multilevel annotations, including point level, segmentation, and object level metrics [121].

Cross-dataset evaluation remains challenging due to these protocol differences [8]. A detection algorithm may perform well under one evaluation protocol but poorly under another, complicating fair comparison across research that uses different datasets. This challenge is further amplified by differences in radar hardware, configurations, and preprocessing across datasets. Some researchers have attempted to address this by evaluating their approaches on multiple datasets using the respective native protocols, providing a more comprehensive assessment of generalization capabilities.

Time dependent evaluation protocols extend beyond single frame detection to assess temporal consistency [118]. These protocols track detection performance across

sequences, considering aspects such as ID switches, track fragmentation, and detection stability. Metrics like Multiple Object Tracking Accuracy (MOTA) and Multiple Object Tracking Precision (MOTP) quantify tracking performance alongside detection, providing a more complete assessment of algorithms intended for continuous operation.

The development of standardized evaluation protocols for radar based detection continues to evolve as the field matures [8]. Efforts to establish common frameworks that account for radar specific characteristics while enabling fair comparison across approaches represent an important direction for future research. These protocols will likely need to balance general applicability with the flexibility to accommodate diverse radar configurations and application requirements.

4.2.2 OLS and Its Significance

Object Location Similarity (OLS) [51] represents an alternative evaluation metric specifically designed for radar based object detection, particularly in the context of the CRUW dataset and related research . Unlike traditional metrics based on Intersection over Union (IoU), OLS addresses the unique characteristics of radar data, providing a more appropriate measure of detection performance for automotive radar applications.

4.2.2.1 Detailed Explanation of OLS Methodology

Object Location Similarity (OLS) was introduced by Wang et al. alongside the CRUW dataset to address the limitations of conventional IoU based metrics when applied to radar detection. OLS employs a center based evaluation approach that considers the distance between detected objects and ground truth objects relative to their size and type .

The fundamental principle of OLS is to determine whether a detection correctly identifies an object based on the distance between their centers rather than their overlap . This approach aligns with the characteristics of radar data, where precise boundary determination is challenging due to limited resolution, but center locations can be estimated more reliably .

Mathematically, for a detection d and ground truth object g, the OLS metric is defined as :

$$OLS(d,g) = exp(-\frac{\left|\left|P_d - P_g\right|\right|^2}{2\sigma^2})$$

where P_d and P_g are the center positions of the detection and ground truth respectively, $||\ ||$ represents the Euclidean distance, and σ is a parameter that controls the sensitivity of the similarity measure to distance. This parameter σ is class dependent, reflecting the different sizes and radar signatures of various road users and it is defined as follows :

$$\sigma = SK_{cls}$$

Where S represents the distance of the object from the radar, and K_{cls} is a per class constant representing the error tolerance for each specific class. K_{cls} is derived from the size of the object, i.e. vehicles vs pedestrian, and in the context of the studies that use OLS, the value per class has been determined empirically. In essence, for larger objects like vehicles, the threshold is set higher than for smaller objects like pedestrians, reflecting the different sizes and radar signatures of these road users . Additionally, the threshold may increase with distance from the sensor to account for the decreasing angular resolution of radar at longer ranges .

In the original implementation for the CRUW dataset, the thresholds were set as 0.5m, 0.3m, and 0.3m for vehicles, pedestrians, and cyclists respectively at close range (under 20m), with gradual increases for objects at greater distances . These values were determined empirically based on the typical sizes of the objects and the resolution characteristics of the radar system used in the dataset .

A key aspect of OLS is its handling of multiple detections corresponding to the same ground truth object . When multiple detections fall within the threshold distance of a

ground truth center, only one is counted as a true positive (typically the one with highest confidence), while the others are considered false positives . Similarly, if multiple ground truth objects are within the threshold distance of a detection, the detection is associated with only one ground truth (typically the closest), with the others counted as false negatives .

4.2.2.2 Application to Radar Detection Tasks

The OLS metric is used in the research described by this literature and has been applied in radar detection research, particularly in works based on the CRUW dataset and its descendants [48, 51]. Its center based approach addresses several radar specific challenges that make traditional IoU based metrics less suitable for radar data.

In the seminal RODNet paper, Wang et al. employed OLS as the primary evaluation metric for their radar object detection network. The authors demonstrated that OLS provided a more meaningful assessment of detection performance than IoU based metrics, particularly for smaller objects like pedestrians and cyclists that generate fewer radar reflection points. Their detection system achieved OLS based Average Precision (AP) scores of 0.77 for vehicles, 0.58 for pedestrians, and 0.40 for cyclists, establishing baseline performance on the CRUW dataset [51].

Jiang et al. [48] utilized the OLS metric to evaluate their transformer based T-RODNet architecture on the CRUW dataset. Their approach achieved improved OLS-AP scores of 0.83 for vehicles, 0.67 for pedestrians, and 0.52 for cyclists, representing significant advances over the original RODNet baseline. The consistent use of OLS across these studies enabled direct performance comparison and tracking of research progress [48].

4.2.2.3 Advantages over Traditional Metrics

OLS offers several significant advantages over traditional evaluation metrics like IoU for radar based detection tasks [48, 51]. These advantages stem from its alignment with the physical characteristics of radar measurements and the practical requirements of automotive perception.

One key advantage is OLS's suitability for sparse radar data . Unlike camera images, radar typically provides sparse measurements with few reflection points per object, making boundary determination challenging and IoU calculation potentially unstable . The center based approach of OLS avoids this issue by focusing on the more reliably estimated center positions, providing a more robust evaluation metric for sparse radar detections .

OLS naturally accommodates the anisotropic resolution of radar sensors . Automotive radars typically have fine resolution in the range dimension but coarser resolution in azimuth, creating asymmetric uncertainty that isn't well captured by standard IoU . By using distance thresholds that can be adjusted based on sensor characteristics, OLS better reflects the actual capabilities and limitations of the radar system .

The class dependent thresholds in OLS provide a more nuanced evaluation across different object types . Smaller objects like pedestrians generate weaker radar signatures with fewer reflection points than larger objects like vehicles, making precise localization more challenging . By employing different distance thresholds based on object class, OLS accommodates these differences and provides a fair evaluation across diverse road users .

OLS offers computational simplicity compared to IoU, particularly for non-rectangular object representations . For radar data represented in polar coordinates or as point clouds, calculating IoU can be complex and computationally intensive . In contrast, center distance calculation is straightforward regardless of the coordinate system or object representation, facilitating efficient evaluation of large detection sets .

The range dependent thresholds in OLS reflect the practical requirements of automotive applications . At longer ranges, precise lateral positioning becomes less critical for decision making than simply detecting the presence of an object . By allowing larger distance thresholds for distant objects, OLS aligns the evaluation with the actual needs of downstream planning and control systems .

4.2.2.4 Statistical Interpretation

The continuous nature of the OLS metric offers rich statistical interpretations that provide deeper insights into detection performance beyond simple binary assessments . Understanding these interpretations helps explain the metric's behavior and guides its application in radar detection evaluation.

From a statistical perspective, OLS implements a form of radial basis function similarity, specifically using a Gaussian kernel to measure the similarity between detection and ground truth positions . This approach has strong connections to kernel density estimation and Gaussian processes in statistical learning theory . The exponential form of OLS means that the similarity drops off gradually with distance rather than abruptly, providing a fine tuned measure of localization accuracy.

The parameter σ in the OLS formula can be interpreted as the standard deviation of a Gaussian distribution centered at the ground truth position . This creates an implicit probabilistic model of localization uncertainty, where the likelihood of a detection being considered correct decreases with distance according to a Gaussian profile . Larger σ values create a more forgiving metric that allows greater positional discrepancy, while smaller values enforce stricter localization requirements .

The class dependent σ values in the CRUW implementation reflect an implicit statistical model of the expected localization accuracy for different object types . The larger σ for vehicles acknowledges their greater spatial extent and the resulting higher uncertainty in determining their exact center . Conversely, the smaller σ for pedestrians and cyclists enforces stricter localization requirements for these more compact objects where precise positioning is more critical for downstream tasks like path planning .

The OLS values can be directly interpreted as confidence scores for localization accuracy . A high OLS value indicates high confidence that the detection is correctly localized, while lower values indicate increasing uncertainty . This continuous nature

allows for more sophisticated analyses than binary metrics, such as examining the distribution of localization errors across different detection scenarios .

When calculating average precision using OLS (OLS-AP), the continuous similarity scores provide a more graduated evaluation of detection quality . Rather than counting detections as either correct or incorrect based on a threshold, each detection contributes according to its OLS value . This approach rewards methods that achieve precise localization while still acknowledging the value of detections that are slightly offset but still reasonably close to the ground truth .

4.3 Experimental Setup and Configuration

The experimental framework for this research was designed to enable comprehensive evaluation of radar-based object detection algorithms while maintaining reproducibility and scientific rigor. This section details the hardware and software infrastructure, data preprocessing methodologies, and training configurations that form the foundation of our experimental approach.

4.3.1 Hardware and Software Environment

The computational infrastructure for this research was carefully selected to balance performance requirements with accessibility for the broader research community. All experiments were conducted on systems equipped with NVIDIA RTX 5090 GPU featuring 32GB of GDDR7 memory, providing sufficient computational resources for training deep neural networks on radar data while remaining representative of hardware accessible to academic researchers. The choice of RTX 5090 over higher-end alternatives like A100 or V100 reflects the practical consideration that most research institutions operate within budget constraints, and our goal was to demonstrate that state-of-the-art performance could be achieved without requiring prohibitively expensive hardware.

The software environment was built upon PyTorch 1.12.1 with CUDA 11.6 support, providing a stable and well-documented deep learning framework with extensive

community support. This combination ensures reproducibility across different systems while leveraging GPU acceleration for efficient training and inference. The choice of PyTorch over alternatives like TensorFlow was motivated by its dynamic computation graph capabilities, which proved valuable during the iterative development of our hybrid CNN-Transformer architecture, allowing for more flexible model modifications and debugging during the research process.

Experiment tracking and model management were handled through MLflow 1.28.0, providing comprehensive logging of hyperparameters, metrics, and model artifacts. MLflow's integration with our training pipeline enabled systematic comparison of different architectural variants and hyperparameter configurations, facilitating the extensive ablation studies that form a crucial component of this research. The MLflow tracking server was configured to store all experimental data locally, ensuring complete control over sensitive research data while maintaining the ability to share results with collaborators when appropriate.

Real-time monitoring and visualization of training progress were also achieved through MLFlow integration, allowing for immediate assessment of model convergence, loss dynamics, and performance metrics for multiple experiments/runs. This capability proved particularly valuable during the development of our multi-component loss function, where visualizing the individual contributions of different loss terms enabled fine-tuning of the weighting parameters. Figure 4.1 illustrates the comprehensive monitoring dashboard that tracks training metrics, validation performance, and computational efficiency throughout the training process.

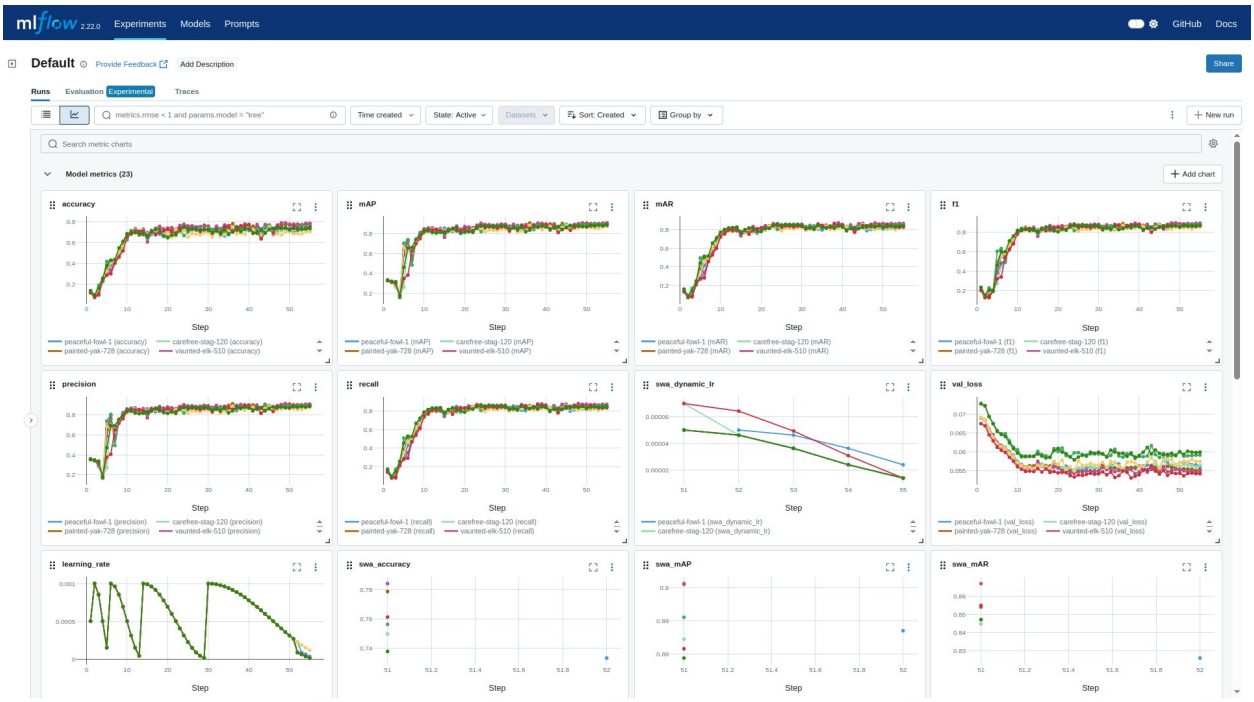


Figure 4.1: Training monitoring dashboard showing loss curves, performance metrics, and computational statistics

The training infrastructure was designed with fault tolerance and checkpoint management in mind. Automatic checkpoint saving every epoch ensures that training progress is preserved even in the event of system failures, while the implementation of warm restart capabilities allows for efficient exploration of different hyperparameter configurations without starting training from scratch. The checkpoint system stores not only model weights but also optimizer states, learning rate scheduler states, and random number generator seeds, ensuring perfect reproducibility of training runs.

Memory management optimizations were crucial for handling the high-dimensional radar data tensors efficiently. The implementation employs mixed-precision training using PyTorch's automatic mixed precision (AMP) capabilities, reducing memory consumption by approximately 40% while maintaining numerical stability. This optimization enables training with larger batch sizes, in order to explore the possibility of improving gradient estimation quality and training stability.

The distributed training capabilities were implemented using PyTorch's DistributedDataParallel (DDP) framework, enabling efficient utilization of multiple GPUs when available. While our primary experiments were conducted on single-GPU systems to ensure broad accessibility, the distributed training infrastructure provides scalability for future research requiring larger computational resources. The implementation includes proper handling of batch normalization statistics synchronization across devices and careful management of random number generator states to ensure reproducible results in distributed settings.

4.3.2 Data Preprocessing Pipeline

The data preprocessing pipeline represents a critical component of our methodology, transforming raw radar measurements into representations suitable for deep learning while preserving the essential characteristics that enable effective object detection. Our preprocessing approach was designed to be modular and configurable, allowing for systematic exploration of different data representations and normalization strategies.

4.3.2.1 Radar Data Normalization

The foundation of our preprocessing pipeline is the handling of multi-chirp radar data, which arrives as complex-valued tensors with dimensions corresponding to range bins, antenna channels, and chirp sequences. The FFT- processed raw ADC data from the CRUW dataset consists of 2-channel complex measurements captured at 30 Hz, with each frame containing the signal from four consecutive chirps. Our preprocessing pipeline formulates this complex data into stacked representations for preserving the multi-channel structure that encodes spatial diversity from the MIMO antenna configuration.

Radar data normalization proved to be a crucial preprocessing step that significantly impacts model performance and training stability. Through extensive experimentation, we implemented three normalization strategies, each addressing different characteristics of radar signal distributions. Standard normalization, computed as:

$$\frac{x - \mu}{\sigma + \epsilon}$$

where μ and σ are the sample mean and standard deviation respectively, proved most effective for the majority of scenarios, providing zero-mean unit-variance distributions that facilitate stable gradient flow during training. The epsilon term (ϵ = 1×10−8) prevents division by zero in regions with minimal radar returns.

The min-max normalization strategy, defined as:

$$\frac{x - x_{\min}}{x_{\max} - x_{\min} + \epsilon}$$

maps radar measurements to the range [0, 1], which can be beneficial when preserving the relative magnitude relationships between different parts of the radar measurement is important. However, our experiments revealed that this approach can be sensitive to outliers in the radar data, particularly strong reflections from metallic objects that can skew the normalization range and suppress weaker signals from vulnerable road users.

Logarithmic normalization, implemented as:

$$\log(1 + x)$$

for non-negative radar magnitude data, addresses the high dynamic range characteristic of radar measurements. Radar cross-sections can vary by several orders of magnitude between different object types, and logarithmic scaling helps compress this range while preserving discriminative information. This approach proved particularly valuable for detecting weak targets like pedestrians and cyclists in the presence of strong vehicle reflections.

The choice of normalization strategy was made configurable in our implementation,

allowing for empirical comparison of their impact on detection performance.

4.3.2.2 Gaussian Heatmap Generation

One of the most critical preprocessing steps in our pipeline involves the transformation of sparse centroid labels into dense Gaussian heatmaps that enable effective end-to-end training of our detection architecture. This conversion addresses the fundamental challenge that radar object detection requires the network to learn to generate dense confidence maps from sparse point annotations, necessitating a principled approach to label representation that preserves spatial relationships while enabling effective gradient-based optimization.

The centroid-to-heatmap conversion process transforms discrete object center points (x_c,y_c) into continuous spatial probability distributions that encode both object presence and spatial uncertainty. This transformation is essential because direct optimization against sparse point targets would create discontinuous loss landscapes that are difficult to optimize effectively. The Gaussian heatmap representation provides smooth, differentiable targets that enable stable gradient flow while encoding appropriate spatial uncertainty based on object characteristics and measurement precision.

The mathematical formulation of our Gaussian heatmap generation follows an elliptical Gaussian model that accounts for the anisotropic characteristics of radar measurements:

$$H(x, y) = \exp\left(-\frac{(x - x_c)^2}{2\sigma_x^2} - \frac{(y - y_c)^2}{2\sigma_y^2}\right)$$

where (x_c,y_c) represents the object centroid, σx and σy represent class-specific standard deviations in the range and azimuth dimensions respectively. This elliptical formulation addresses the fundamental anisotropy of radar measurements, where range precision is typically much finer than azimuth precision due to the physical characteristics of automotive radar systems.

Class-specific sigma values reflect the different spatial characteristics and detection challenges associated with different object types. Our implementation employs σx values of 8.0, 8.0, and 6.0 pixels for persons, bicycles, and cars respectively in the range dimension, while using a fixed σy value of 3.0 pixels in the azimuth dimension. These parameters were determined through systematic analysis of object size distributions and radar measurement characteristics in the training dataset.

The choice of class-specific parameters addresses the varying spatial extents and radar signature characteristics of different object types. Vehicles typically exhibit larger spatial extents and more distributed reflection patterns that justify broader Gaussian spreads, while pedestrians and cyclists present more compact signatures that require tighter spatial localization. The asymmetric sigma values (σ_x vs σ_y) reflect the anisotropic resolution characteristics of automotive radar systems.

Figure 4.2 illustrates a representative example of raw radar range-azimuth maps showing the characteristic reflection patterns from different object types, demonstrating the sparse and variable nature of radar measurements that necessitate sophisticated label representation strategies.

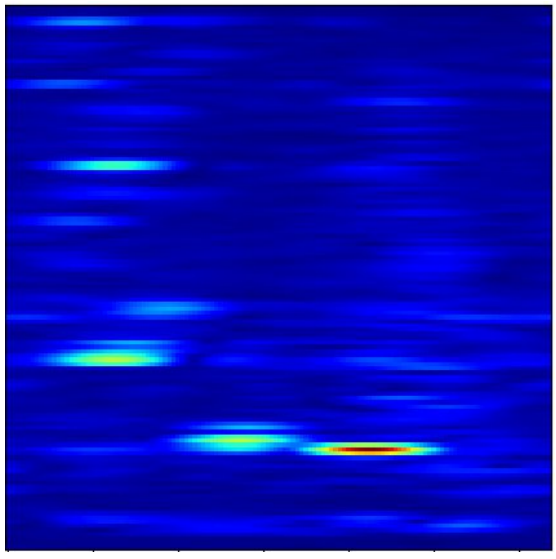


Figure 4.2: Raw radar range-azimuth map showing reflection patterns from vehicles,

pedestrians, and cyclists

The Gaussian heatmap generation process employs vectorized computation strategies that enable efficient processing of large training datasets while maintaining numerical precision. The implementation processes each object independently to generate individual Gaussian distributions, then combines overlapping distributions using element-wise maximum operations to handle scenarios where multiple objects are present in close proximity.

The mathematical implementation addresses several technical challenges that arise in practical radar processing scenarios. Boundary handling ensures that Gaussian distributions are properly clipped at the edges of the radar field of view, preventing artifacts that could occur when objects are located near the boundaries of the measurement space. Computational optimization employs spatial windowing that limits Gaussian computation to regions within three standard deviations of each object center, reducing computational overhead without sacrificing accuracy.

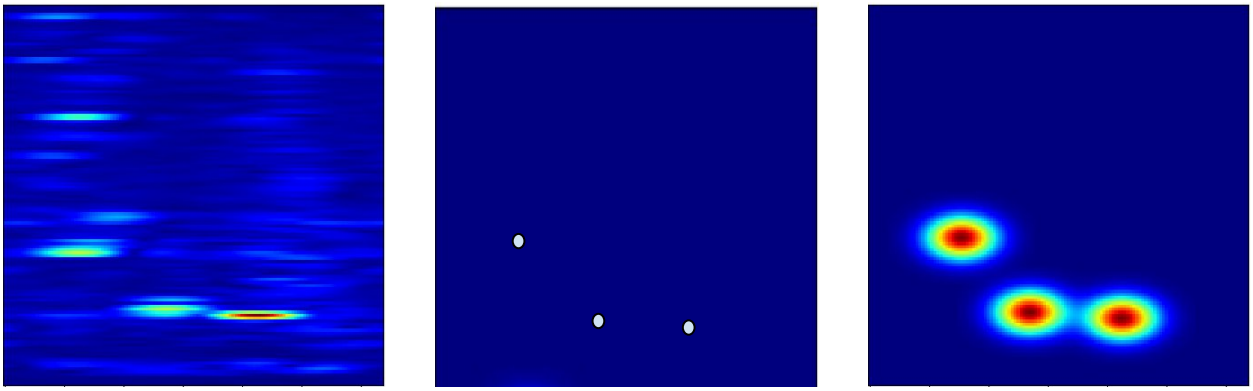
Multi-object handling represents a crucial aspect of the heatmap generation process, as automotive scenarios frequently involve multiple objects within the radar field of view. When multiple objects generate overlapping Gaussian distributions, our implementation employs element-wise maximum combination:

$$H_{\text{combined}}(x, y) = \max_i H_i(x, y)$$

where $H_{i(x,y)}$ represents the Gaussian heatmap for the i-th object. This maximum combination strategy preserves the peak values associated with each object while avoiding artificial amplitude increases that could occur with additive combination approaches.

Figure 4.3 presents an example of the Gaussian heatmap generation process, showing the transformation from sparse centroid annotations to dense probability maps that serve as training targets for our detection network.

The temporal coherence considerations in heatmap generation address the sequential nature of radar processing by ensuring that Gaussian parameters remain consistent across consecutive frames. This consistency is crucial for temporal modeling components that process sequences of radar measurements, as inconsistent label representations could interfere with the learning of smooth temporal patterns.



(a) (b) (c)

Figure 4.3: Gaussian heatmap generation example showing range-azimuth frame in figure a, centroid labels in figure b and the corresponding dense probability distributions generated for each centroid in figure c.

Quality validation procedures verify the accuracy of the heatmap generation process through systematic comparison of generated heatmaps with original centroid locations. The validation ensures that peak locations in the generated heatmaps correspond precisely to the original centroids while confirming that the Gaussian spreads appropriately reflect the specified sigma parameters. This validation is essential for maintaining training signal quality throughout the preprocessing pipeline.

4.3.2.3 Data Augmentation

Data augmentation represents another crucial component of our preprocessing pipeline, addressing the limited size of available radar datasets and improving model

generalization. Unlike image data augmentation, radar data augmentation must respect the physical properties of electromagnetic wave propagation and the geometric constraints of the radar measurement process. Our augmentation strategy employs the Kornia library for GPU-accelerated transformations, implementing radar-specific augmentations that preserve the essential characteristics of radar measurements while introducing controlled variations that improve model robustness.

Horizontal flipping represents the most straightforward augmentation, reflecting radar measurements across the azimuth dimension. This transformation is physically meaningful as it corresponds to objects approaching from the opposite direction, doubling the effective size of the training dataset while maintaining the physical validity of the augmented samples. The implementation ensures that ground truth labels are appropriately transformed to match the flipped radar data.

Spatial translation augmentations simulate variations in object positions within the radar field of view. Small horizontal and vertical shifts, limited to ±10 pixels in our implementation, introduce positional variations without significantly altering the physical interpretation of the measurements. These translations must be applied carefully to ensure that augmented objects remain within the valid radar detection range and that the corresponding ground truth labels are updated appropriately.

Gaussian noise addition simulates variations in radar measurement quality that can occur due to environmental factors, hardware variations, or interference from other electromagnetic sources. The noise level was carefully calibrated to represent realistic variations without overwhelming the signal content. A standard deviation of 0.02, determined through analysis of measurement variations in the original dataset, provides sufficient variation to improve robustness without degrading the training signal quality.

Intensity scaling augmentations address variations in radar cross-section that can occur due to changes in object orientation, material properties, or environmental conditions. Multiplicative scaling factors drawn from the range [0.8, 1.2] introduce controlled variations in reflection intensity while preserving the relative relationships between

different reflection points. This augmentation helps the model generalize to variations in object presentation that are common in real-world radar data.

The augmentation pipeline was implemented using Kornia's GPU-accelerated transformations, enabling efficient augmentation without requiring data transfer between GPU and CPU memory. This implementation choice significantly reduces the computational overhead of data augmentation, allowing for more extensive augmentation strategies without impacting training efficiency. The augmentation parameters were made configurable, enabling systematic study of their impact on model performance and generalization.

Multi-chirp data handling represents a unique aspect of our preprocessing pipeline, addressing the temporal dimension inherent in radar measurements. Unlike single-snapshot approaches that process individual radar frames in isolation, our pipeline preserves the temporal relationships between consecutive micro measurements, enabling the extraction of micro motion information that is crucial for object classification and tracking.

Chirp stacking strategies were developed to combine multiple consecutive radar measurements into unified representations suitable for deep learning processing. The default approach stacks 4 consecutive radar frame chirps along the channel dimension, creating input tensors with dimensions (B, 8, H, W) where B represents the batch size, and H, W represent the spatial dimensions of the radar measurements. This stacking approach enables 3D convolutional processing that can extract spatio-temporal features while maintaining computational efficiency.

Alternative stacking strategies were also implemented and evaluated, including magnitude-phase decomposition where real and imaginary components of complex radar data are processed separately, and Doppler-enhanced stacking where explicit velocity information is incorporated into the input representation. These alternatives provide flexibility for future research directions while maintaining compatibility with our core processing pipeline.

4.3.2.4 Temporal Frame Stacking Strategy

The temporal dimension in radar data provides crucial information about object motion dynamics and scene evolution that cannot be captured from single-frame analysis alone. Our preprocessing pipeline incorporates a robust temporal frame stacking strategy that assembles sequences of radar frames to enable temporal modeling while maintaining computational tractability and ensuring proper temporal alignment for accurate object detection.

4.3.2.4.1 Temporal Configuration Framework

We implement a flexible temporal configuration system that supports various frame stacking strategies to accommodate different operational requirements and computational constraints. Each configuration specifies three key parameters: the total number of frames to process, the position of the principal frame within the sequence, and optional frame skipping for extended temporal coverage.

The principal frame concept proves critical for maintaining accurate ground truth alignment in object detection tasks. While the model processes multiple frames to extract temporal features, the detection output corresponds to object positions at a specific time instant - the principal frame. This design ensures that temporal context enhances detection quality without introducing temporal ambiguity in object localization.

Our primary configuration employs an 11-frame sequence with the principal frame positioned at the center (frame index 5), providing balanced temporal context with 5 frames of historical data and 5 frames of future information. This bidirectional temporal window enables the model to observe complete motion patterns, including object approach and recession, while maintaining real-time viability through limited future frame requirements. The 11-frame configuration represents a tuned balance between temporal coverage and computational efficiency, spanning approximately 550ms at typical automotive radar frame rates of 20Hz.

4.3.2.4.2 Sequence Organization and Boundary Handling

The temporal frame stacking process begins with organizing the continuous stream of radar data into coherent sequences based on their temporal relationships. Each radar frame is uniquely identified by a sequence number and frame index, enabling efficient retrieval and proper temporal ordering. The preprocessing system maintains a mapping of available sequences and their constituent frames, automatically identifying sequence boundaries and ensuring temporal continuity.

Boundary handling presents a particular challenge when the requested temporal window extends beyond available frames at sequence starts or ends. Our implementation employs a strict boundary policy that only creates valid samples when the complete temporal window can be satisfied within a single sequence. This approach prevents the model from learning false patterns from padded or artificially extended sequences while ensuring that all training samples maintain consistent temporal structure.

For deployment scenarios requiring predictions at sequence boundaries, we provide alternative configurations including past-only stacking (where the principal frame is the last in the sequence) and future-only stacking (where the principal frame is the first). These configurations enable detection at sequence boundaries while maintaining temporal modeling benefits, albeit with reduced bidirectional context. Our ablation study in Chapter 7 demonstrates a strong performance for a principle-frame in the last position, requiring no look-ahead frames.

4.3.2.4.3 Frame Skipping for Extended Temporal Coverage

While consecutive frame stacking provides fine-grained temporal resolution, certain scenarios benefit from extended temporal coverage to capture longer-term motion patterns. Our framework supports configurable frame skipping, where frames are sampled at regular intervals rather than consecutively. For instance, a configuration with skip factor 3 samples every third frame, extending the effective temporal window from 550ms to 1.65 seconds while maintaining the same computational requirements.

Frame skipping proves particularly valuable for highway scenarios where vehicle motions exhibit longer-term coherence, and for detecting complex maneuvers that unfold over extended time periods. The skip factor can be dynamically adjusted based on the deployment context, with shorter skips for urban environments featuring rapid motion changes and longer skips for highway scenarios with more predictable motion patterns.

4.3.2.4.4 Temporal Data Loading and Caching

The temporal frame stacking process imposes significant data loading requirements, as each training sample requires accessing multiple radar frames potentially scattered across storage. Our implementation employs intelligent caching strategies that balance memory usage with loading efficiency. Frequently accessed sequences are maintained in memory, while a least-recently-used eviction policy ensures bounded memory consumption.

The data loader implements prefetching for sequential access patterns common during training, predicting and loading future sequences based on access patterns. This predictive loading significantly reduces I/O wait times and enables smooth training progression. For random access patterns during validation, the system maintains an index of frame locations enabling direct access without sequential scanning.

4.3.2.4.5 Integration with Data Augmentation

Temporal frame stacking requires careful coordination with spatial data augmentation to maintain temporal consistency. Spatial transformations such as rotations or flips must be applied consistently across all frames in a temporal sequence to preserve motion coherence. Our implementation ensures that the same spatial transformation parameters are used for all frames within a sequence while allowing different transformations across sequences for augmentation diversity.

Temporal-specific augmentations further enhance model robustness. Frame dropping simulation randomly removes individual frames from sequences during training,

improving resilience to temporary sensor failures or processing drops. Temporal jittering introduces small random delays between frames, simulating variations in radar timing that may occur in real deployments. These augmentations ensure that the trained model can handle imperfect temporal sequences while maintaining detection performance.

4.3.3 Comprehensive Training Strategy

4.3.3.1 Basic Training Configuration

The training configuration encompasses the comprehensive set of hyperparameters, optimization strategies, and training procedures that enable effective learning from radar data. Our configuration was developed through systematic experimentation and incorporates recent advances in deep learning optimization while addressing the specific challenges posed by radar-based object detection.

Hyperparameter selection represents a critical aspect of our training configuration, with choices informed by both theoretical considerations and empirical evaluation. The learning rate, perhaps the most crucial hyperparameter in deep learning, was set to ***0.001*** as the initial value, chosen based on extensive experimentation on the CRUW radar dataset and the model implementation. This selection, paired with our choice of optimizer and learning scheduler dynamics ensures balanced exploration and exploitation of the problem landscape during training.

The batch size configuration balances computational efficiency with gradient estimation quality. Our standard batch size of ***32*** was selected as the maximum size that could be accommodated within the memory constraints of RTX 5090 GPU while processing our high-dimensional radar input tensors. This batch size provides sufficient sampling for stable gradient estimation while enabling efficient GPU utilization. For systems with larger memory capacity, the implementation supports larger batch sizes through configuration parameters, however, it is imperative to note that a larger batch size does not automatically translate to a better training outcome.

Training duration was set to ***100*** epochs based on empirical observation of convergence

patterns across multiple model variants. This duration provides sufficient training time for the model to converge while avoiding excessive overfitting, particularly important given the limited size of available radar datasets. The epoch count was determined through systematic experimentation with early stopping criteria, identifying the point where validation performance plateaus while maintaining generalization capability.

Weight decay regularization, set to $\boldsymbol{1\times10^{-4}}$, provides L2 regularization that helps prevent overfitting by penalizing large parameter values. This value was selected through experimentation over the range [$\boldsymbol{1\times10^{-5},1\times10^{-3}}$], with the chosen value providing the best balance between model capacity and generalization performance. The weight decay is applied uniformly across all learnable parameters except for bias terms and batch normalization parameters, following established best practices.

The seed value for random number generation was fixed at ***33*** to ensure reproducible results across multiple training runs. This deterministic setup enables precise comparison between different model variants and training strategies while facilitating debugging and validation of experimental results. The seed controls initialization of network weights, data augmentation randomness, and training data shuffling, ensuring that performance differences between experiments can be attributed to architectural or hyperparameter changes rather than random variations. Nevertheless, it is worth noting that the introduction of augmentation probability and dropout probability does create a variation that is outside the control of the pseudo randomness control. However, the variations are within the limit that allows the ability to study the impact of architectural changes.

Learning rate scheduling represents an important aspect of our training configuration, employing multiple strategies to optimize the learning process throughout training and providing an adequate balance between exploration and exploitation for a solution in the domain dimension. The primary approach utilizes CosineAnnealingWarmRestarts, implementing a cyclical learning rate schedule that enables the model to escape local minima while gradually reducing the learning rate over time. The mathematical

formulation follows:

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

where ηt is the learning rate at step t, ηmin and ηmax are the minimum and maximum learning rates respectively, T_{cur} is the number of steps since the last restart, and Ti is the length of the current cycle. The warm restart mechanism resets T_{cur} to zero and potentially increases Ti at the end of each cycle, enabling exploration of different regions of the parameter space.

The scheduler parameters were carefully tuned to match the characteristics of radar data learning. The initial cycle length T_0 was set to twice the number of batches per epoch, ensuring that each cycle spans multiple epochs and provides sufficient time for meaningful parameter updates. The multiplication factor T_{mult}=2 doubles the cycle length after each restart, gradually extending the exploration periods as training progresses. The minimum learning rate $\eta_{min}=1\times10^{-4}$ ensures continued learning even at the end of long cycles.

Alternative scheduling strategies were also implemented and evaluated, including OneCycleLR for rapid convergence and ReduceLROnPlateau for adaptive learning rate reduction based on validation performance. The OneCycleLR scheduler implements the "super-convergence" approach proposed by Smith and Topin [133], which can achieve faster training convergence through careful coordination of learning rate and momentum schedules. The ReduceLROnPlateau scheduler provides a more conservative approach, reducing the learning rate only when validation performance stagnates, ensuring stable training for sensitive model configurations.

Figure 4.4 illustrates the learning rate evolution throughout training under different scheduling strategies, demonstrating the cyclic behavior of

CosineAnnealingWarmRestarts and its impact on training dynamics.

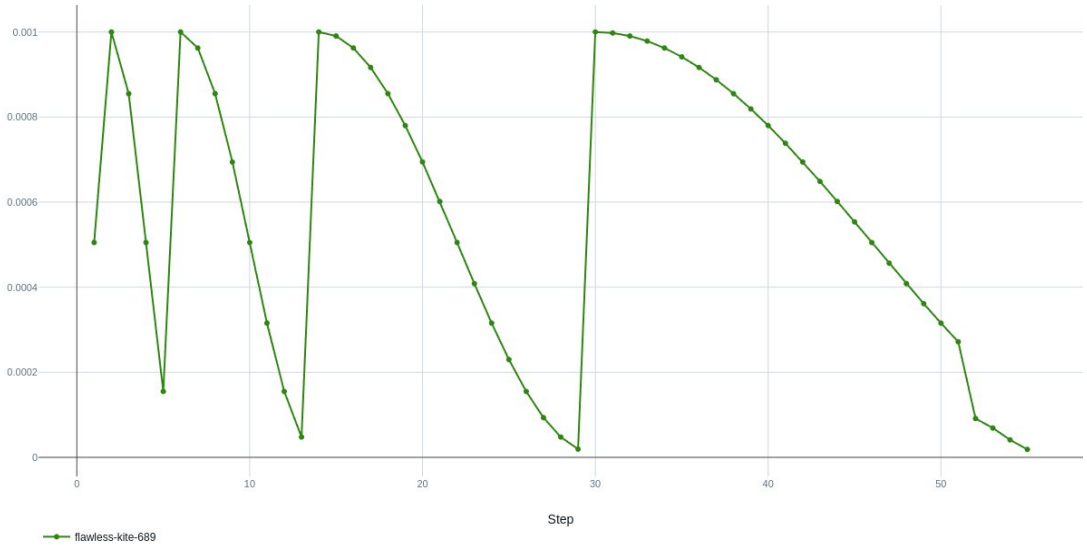


Figure 4.4: Learning rate schedules comparison showing cyclic patterns and convergence behavior

Optimizer selection represents another crucial aspect of our training configuration. The primary optimizer employed is Ranger, a robust optimizer that combines the benefits of Rectified Adam (RAdam) with Lookahead optimization. Ranger addresses the early training instability often observed with Adam-based optimizers while providing the fast convergence properties that make adaptive optimizers attractive for deep learning applications.

The Ranger optimizer implements several key innovations. The RAdam component corrects the variance of adaptive learning rates during the early stages of training, addressing the "bad local minima" problem that can occur with standard Adam optimization. The mathematical formulation includes a variance correction term:

$$r_t = \sqrt{\frac{(\rho_t - 4)(\rho_t - 2)\rho_\infty}{(\rho_\infty - 4)(\rho_\infty - 2)\rho_t}}$$

where ρ_t tracks the length of the exponential moving average and ρ_∞ is the asymptotic length. This correction is applied only when $\rho_t > 4$, ensuring stable updates during early training phases.

The Lookahead component maintains a set of "slow weights" that are updated as exponential moving averages of the "fast weights" optimized by RAdam. The slow weights are computed as:

$$\phi_{t+1} = \phi_t + \alpha(\theta_{t+1} - \phi_t)$$

where ϕ_t represents the slow weights, θ_{t+1} represents the fast weights after k RAdam updates, and α is the slow weight update factor. This mechanism provides additional stability and can help escape sharp local minima that might trap the fast weights.

The Ranger optimizer parameters were configured as follows: learning rate α=0.001 (as previously mentioned), momentum parameters β_1=0.9 and β_2=0.999, Lookahead step size α=0.5, and Lookahead update frequency k=5. These parameters were selected based on the original Ranger paper recommendations and validated through empirical experimentation on radar data.

Stochastic Weight Averaging (SWA) represents an advanced training technique integrated into our configuration to improve model generalization and robustness. SWA maintains running averages of model parameters during the later stages of training, effectively ensemble multiple models encountered during training into a single model with improved generalization properties.

The SWA implementation begins activation when the model achieves a validation mAP threshold of ***0.87***, indicating that the model has reached a reasonable level of performance and is ready for the averaging process. This threshold-based activation

ensures that SWA only begins after the model has learned meaningful representations, avoiding the averaging of poor-quality early-training parameters. The activation also requires a minimum epoch threshold of ***50***, ensuring sufficient training time before SWA activation.

When SWA activates, several training modifications are implemented. The learning rate is reset to a higher value within the range [1×10^{-5}, 1×10^{-4}], enabling continued exploration of the parameter space. A cyclical learning rate schedule with period ***5*** epochs facilitates movement between different parameter regions during the averaging process. The momentum factor is reduced to ***0.9*** to provide more responsive updates during the SWA phase.

The SWA model updates follow the mathematical formulation:

$$\theta_{SWA} = \frac{1}{n}\sum_{i=1}^{n}\theta_i$$

where θ_{SWA} represents the averaged parameters, θ_i represents the parameters from the i-th model checkpoint, and n is the number of checkpoints included in the average. The averaging is performed after every epoch during the SWA phase, with periodic evaluation of the averaged model to monitor its performance.

Dynamic learning rate scaling was implemented to adapt the learning rate based on validation performance during the SWA phase. When the current validation mAP is within ***1%*** of the best observed mAP, the maximum learning rate is reduced to 50% of its original value, encouraging exploitation over exploration. This adaptive mechanism helps the SWA process converge toward high-quality parameter regions while maintaining the exploration necessary for effective averaging.

The training extension mechanism automatically extends the total training duration when

SWA is activated, adding ***50*** additional epochs to provide sufficient time for the averaging process to converge. This extension ensures that SWA has adequate opportunity to improve model performance without being constrained by the original training schedule designed for non-SWA training.

The comprehensive training configuration also includes advanced regularization techniques beyond basic weight decay. Spatial dropout with probability ***0.1*** is applied to feature maps in the decoder, providing structured regularization that is more appropriate for convolutional architectures than standard dropout. Group normalization is employed throughout the architecture instead of batch normalization, providing more stable training with smaller batch sizes and better generalization across different batch size configurations.

Checkpoint management ensures robust training with automatic recovery capabilities. Model checkpoints are saved after every epoch, including model state, optimizer state, scheduler state, and random number generator states. The best model selection criterion is based on harmonic mean of mAP and mAR with minimum improvement threshold ***0.0001***, ensuring that saved models represent genuine performance improvements rather than random fluctuations.

4.3.3.2 Advanced Loss Function Design

The development of effective loss functions represents a critical component of our radar-based object detection system, addressing the unique challenges posed by class imbalance, weak signal characteristics, and the multi-task nature of object detection. This section details our comprehensive loss function design that combines multiple components to optimize different aspects of the detection task while maintaining training stability and convergence properties.

4.3.3.2.1 Multi-Component Loss Architecture

The multicomponent loss architecture represents a more involved approach to optimizing radar based object detection that addresses multiple challenges simultaneously through

carefully designed loss terms that target different aspects of the detection task. Unlike simple loss functions that optimize a single objective, our multi-component approach recognizes that effective radar detection requires balancing classification accuracy, localization precision, confidence estimation, and auxiliary supervision objectives within a unified optimization framework.

The foundational motivation for multi-component loss design stems from the inherent complexity of radar-based object detection tasks. Classification objectives must handle severe class imbalance while distinguishing between objects with subtle signature differences. Regression objectives must achieve precise localization despite the sparse and noisy nature of radar measurements. Confidence estimation must provide reliable uncertainty quantification to support downstream decision making. Each of these objectives benefits from specialized loss formulations that address their unique characteristics and challenges.

Our loss architecture employs a hierarchical structure that combines primary loss terms for the main detection objectives with auxiliary loss terms that provide additional supervision signals and regularization effects. The primary loss terms include classification loss, regression loss, and confidence estimation loss, while auxiliary terms include deep supervision losses from intermediate network outputs and regularization terms that encourage desirable learned representations.

The loss function combines standard components with fixed weights:

- Focal loss for classification (main and auxiliary)
- L1 loss for regression (main and auxiliary)
- Tversky loss for improved segmentation quality
- Fixed weighting: λ_reg=1.0, λ_aux=0.4, λ_tv=0.06

Our implementation implements this combination, where auxiliary losses provide deep supervision and Tversky loss enhances classification performance."

Loss term weighting strategies represent a crucial aspect of multi-component loss design,

determining how different objectives are balanced during optimization. Fixed weighting approaches employ predetermined coefficients based on empirical experimentation and domain knowledge about the relative importance of different objectives.

4.3.3.2.2 Class Imbalance Handling

Class imbalance represents one of the most significant challenges in radar-based object detection, with vehicles being orders of magnitude more common than vulnerable road users in typical automotive datasets. This imbalance creates training dynamics that bias the network toward detecting common object classes while struggling to learn effective representations for rare but critically important classes like pedestrians and cyclists. Our class imbalance handling strategy employs multiple complementary approaches that address different aspects of this fundamental challenge.

Focal loss implementation represents the cornerstone of our class imbalance strategy, addressing the problem through a loss formulation that automatically focuses training attention on difficult examples while reducing the contribution of easy examples that may overwhelm the training process. The focal loss formulation modifies standard cross-entropy loss through a focusing term that downweights easy classifications and emphasizes hard examples.

The mathematical formulation of focal loss follows:

$$FL(p_t) = -\alpha_t(1 - p_t)^{\gamma} \log(p_t)$$

where p_t represents the model's estimated probability for the ground truth class, α_t provides class-specific weighting, and γ controls the focusing strength. The $(1-p_t)^{\gamma}$ term ensures that easy examples (high p_t) contribute less to the loss, while difficult examples (low p_t) maintain their full contribution.

Class-specific gamma values represent a crucial innovation in our focal loss

implementation, recognizing that different object classes in radar data exhibit different levels of detection difficulty and require different focusing strategies. Vehicles, with their strong and consistent radar signatures, benefit from moderate focusing that prevents overconfidence while maintaining learning efficiency. Vulnerable road users, with their weak and variable signatures, require stronger focusing that emphasizes the most challenging detection cases.

Our implementation employs gamma values of 1.0, 2.0, and 1.0 for persons, bicycles, and cars respectively. The higher gamma value for bicycles reflects their particularly challenging detection characteristics; they combine the weak signature characteristics of pedestrians with the complex motion patterns and geometric variability that make them the most difficult class in our detection task. These gamma values were determined through systematic empirical experimentation and studying the class imbalance in the CRUW dataset.

Alpha weighting arrays provide additional class-specific balancing that addresses the frequency imbalance between different object classes. The alpha values are computed based on inverse class frequency statistics from the training data, ensuring that rare classes receive higher weight in the loss computation to compensate for their lower representation in the training set.

The alpha weighting computation follows:

$$\alpha_c = \frac{N_{total}}{N_c \cdot C}$$

where α_c represents the alpha weight for class c, N_{total} represents the total number of training examples, N_c represents the number of examples for class c, and C represents the total number of classes. This formulation ensures that alpha weights are inversely proportional to class frequency while maintaining appropriate scaling.

Our implementation employs alpha values of 0.8, 1.0, and 0.62 for persons, bicycles, and

cars respectively. These values reflect both the frequency imbalance and the relative detection difficulty of different classes, with bicycles receiving the highest weighting due to their combination of low frequency and high detection difficulty.

4.3.3.3 Training Optimization Strategies

The effective training of our hybrid radar detection architecture requires robust optimization strategies that address the unique challenges posed by radar data characteristics, the complexity of multi-component loss functions, and the need for robust generalization despite limited training data. This section details the advanced optimization techniques, regularization strategies, and ensemble methods that enable effective learning while maintaining deployment viability for automotive applications.

4.3.3.3.1 Advanced Optimization Techniques

The selection and configuration of optimization algorithms represents a critical factor in achieving effective training of deep neural networks for radar-based object detection. Traditional optimization approaches like SGD and Adam, while effective for many computer vision tasks, may not address the specific challenges posed by radar data including high-dimensional sparse inputs, multi-component loss functions, and the need for robust convergence despite limited training data.

Ranger optimizer implementation represents the cornerstone of our optimization strategy, combining the benefits of Rectified Adam (RAdam) with Lookahead optimization to address both the variance issues in adaptive learning rates and the sharp minima problems that can limit generalization performance. RAdam addresses the early training instability often observed with adaptive optimizers by implementing variance correction that stabilizes learning rate adaptation during the initial training phases when gradient statistics are poorly estimated.

The mathematical formulation of RAdam's variance correction follows:

$$\rho_t = \rho_\infty - \frac{2t\beta_2^t}{1-\beta_2^t}$$

$$r_t = \sqrt{\frac{(\rho_t - 4)(\rho_t - 2)\rho_\infty}{(\rho_\infty - 4)(\rho_\infty - 2)\rho_t}}$$

where ρ_t represents the length of the approximated exponential moving average at time t, ρ_∞ represents the asymptotic length, and r_t represents the variance correction factor. The correction is applied only when $\rho_t > 4$, ensuring stable updates during early training phases when gradient statistics are unreliable.

Lookahead optimization complements RAdam by maintaining "slow weights" that provide additional stability and can help escape sharp local minima that might trap conventional optimizers. The Lookahead mechanism operates by maintaining two sets of parameters: "fast weights" that are updated by the base optimizer (RAdam in our case) and "slow weights" that are updated as exponential moving averages of the fast weights.

The Lookahead update mechanism follows:

$$\phi_{t+1} = \phi_t + \alpha(\theta_{t+k} - \phi_t)$$

where ϕt represents the slow weights, θ_{t+k} represents the fast weights after k RAdam updates, α represents the slow weights update factor (typically 0.5), and k represents the update frequency (typically 5). This mechanism provides additional stability while enabling exploration of the parameter space beyond what RAdam alone might achieve.

The Ranger optimizer configuration in our implementation employs carefully tuned parameters that address the specific characteristics of radar data training. The base learning rate is set to 0.001, providing a conservative starting point that ensures stable

initial training while being aggressive enough to achieve reasonable convergence speed. The momentum parameters β_1=0.9 and β_2=0.999 follow standard adaptive optimizer recommendations while the Lookahead parameters α=0.5 and k=5 provide effective slow weight updates without excessive computational overhead.

Learning rate scheduling represents another crucial component of our optimization strategy, employing effective scheduling algorithms that adapt the learning rate throughout training to maintain effective learning while ensuring convergence to high-quality solutions. Our primary approach employs CosineAnnealingWarmRestarts, which implements a cyclical learning rate schedule that enables the optimizer to escape local minima while gradually reducing the learning rate over time.

Adaptive learning rate strategies extend beyond fixed scheduling to incorporate validation-based adaptation that adjusts learning rates based on observed training progress and convergence characteristics. ReduceLROnPlateau scheduling provides automatic learning rate reduction when validation performance plateaus, enabling automatic adaptation to training dynamics without requiring manual intervention.

The plateau detection mechanism monitors validation mAP improvement over specified patience periods (typically 5-10 epochs) and reduces the learning rate by a factor (typically 0.5) when improvement stagnates. This adaptive approach ensures that learning rates remain appropriate throughout training without requiring manual tuning for different training scenarios or dataset characteristics.

Warm-up strategies address the challenges of training very deep networks from random initialization by gradually increasing the learning rate from a small initial value to the target learning rate over the first several epochs. This approach prevents the large gradient updates that can occur during early training when network parameters are far from suitable configurations.

The warm-up implementation employs a linear learning rate increase from $\eta_{min}=1\times10-6$

to the target learning rate over the first 5 epochs of training. This gradual increase provides stable training dynamics during the critical early training period while ensuring rapid transition to effective learning rates once the network parameters stabilize.

4.3.3.3.2 Regularization and Generalization

Effective regularization strategies are essential for achieving robust generalization performance when training on limited radar datasets, where overfitting can severely compromise the ability to detect objects in scenarios not well-represented in the training data. Our comprehensive regularization approach combines multiple complementary techniques that address different aspects of overfitting while preserving the representational capacity needed for effective radar feature learning.

Data augmentation represents the foundation of our regularization strategy, employing radar-specific transformations that increase training data diversity while preserving the physical validity of radar measurements. Unlike image augmentation techniques that may employ arbitrary transformations, radar augmentation must respect the physics of electromagnetic wave propagation and the geometric constraints of radar measurement systems. Data augmentation implementation is described in detail in subsection 4.3.2.3.

Dropout and spatial dropout strategies provide complementary regularization that addresses overfitting at the feature level by randomly zeroing feature activations during training. Standard dropout is applied to fully connected components with probability 0.1, while spatial dropout is applied to convolutional feature maps with probability 0.1, providing structured regularization that is more appropriate for spatially correlated features.

The spatial dropout implementation randomly zeros entire feature channels rather than individual feature elements, preserving spatial correlation within channels while providing regularization across the channel dimension. This approach proves more effective for

convolutional features than standard dropout because it maintains the spatial structure that is important for effective feature extraction while still providing regularization benefits.

Weight decay regularization provides L2 regularization that prevents overfitting by penalizing large parameter values that might represent overfitting to training data rather than generalizable feature extraction. The weight decay parameter is set to 1×10^{-4}, providing sufficient regularization without overly constraining the model's representational capacity.

The weight decay implementation is applied to all learnable parameters except bias terms and batch normalization parameters, following established best practices that recognize these parameters serve different roles in the network architecture. The selective application ensures effective regularization while avoiding interference with parameters that serve normalization or offset functions.

Early stopping mechanisms monitor validation performance throughout training and terminate training when overfitting is detected through degradation of validation metrics despite continued improvement on training metrics. The early stopping implementation monitors validation mAP with configurable patience of 15 epochs, providing sufficient time for temporary performance fluctuations while preventing excessive overfitting.

The early stopping criteria consider both absolute validation performance and the trend of validation improvement relative to training improvement. Training is terminated when validation performance stagnates or degrades while training performance continues to improve, indicating that the model is learning training-specific patterns rather than generalizable features.

The validation strategy monitors multiple metrics including mAP, mAR, precision, recall, and class-specific performance measures to provide comprehensive assessment of generalization capability. Early validation of these metrics enables detection of overfitting before it severely compromises model performance, allowing for timely intervention through regularization parameter adjustment or training termination.

4.3.3.3.3 Model Ensemble Strategies

Model ensemble strategies represent a strong approach to improving detection performance and robustness through the combination of multiple model variants or training procedures. Ensemble methods can provide significant performance improvements while also offering insights into model reliability and uncertainty estimation that are valuable for safety-critical automotive applications.

Stochastic Weight Averaging (SWA) represents the primary ensemble strategy employed in our system, providing a robust approach to model averaging that combines the benefits of ensemble methods with the computational efficiency of single model deployment. SWA operates by maintaining running averages of model parameters during the later stages of training, effectively creating an ensemble of models encountered during the optimization trajectory.

The SWA implementation is designed with intelligent activation criteria that ensure averaging begins only when the model has achieved sufficient performance to contribute meaningfully to the ensemble. The activation threshold is set at 87% mAP with a minimum epoch requirement of 50, ensuring that parameter averaging begins only after the model has learned effective representations and reached a reasonable performance level.

When SWA activates, several training modifications are implemented to optimize the averaging process. The learning rate is adjusted to enable continued exploration of the parameter space while maintaining convergence properties. A cyclical learning rate schedule with period 5 epochs facilitates movement between different parameter regions during the averaging process, enabling the collection of diverse high-quality parameter configurations. The mathematical formulation of SWA parameter updates has already been discussed in section 4.3.3.1.

Dynamic learning rate scaling during SWA provides adaptive control of the exploration-exploitation tradeoff based on current performance relative to the best observed performance. When current validation mAP approaches the best observed mAP (within

1%), the maximum learning rate is reduced to encourage exploitation of the current parameter region rather than extensive exploration. This adaptive mechanism helps the SWA process converge toward high-quality parameter regions while maintaining sufficient exploration for effective averaging.

Batch normalization statistics updating represents a crucial component of SWA implementation, addressing the fact that batch normalization running statistics must be recomputed for the averaged parameters to ensure robust performance. The statistics update process employs the training dataset to recompute running means and variances for all batch normalization layers using the averaged parameters.

Training extension mechanisms automatically extend the training duration when SWA is activated, providing sufficient time for the averaging process to converge without being constrained by the original training schedule designed for non-SWA training. The extension adds 50 additional epochs, enabling comprehensive parameter averaging while maintaining reasonable total training time.

Model snapshot strategies complement SWA by maintaining collections of high-performing model checkpoints that can be combined through various ensemble methods. The snapshot strategy saves model checkpoints when validation performance exceeds specified thresholds (typically 87% mAP), creating a collection of diverse high-quality models that can be combined for ensemble prediction.

The snapshot collection is managed to maintain diversity while limiting storage requirements. Models are ranked by validation performance and diversity metrics computed based on prediction differences on a held-out validation set. The top-performing diverse models are retained while redundant models with similar prediction patterns are discarded, maintaining an ensemble that balances performance with diversity.

Figure 4.5 illustrates the ensemble strategy architecture, showing the SWA process, model snapshot collection, and ensemble combination mechanisms.

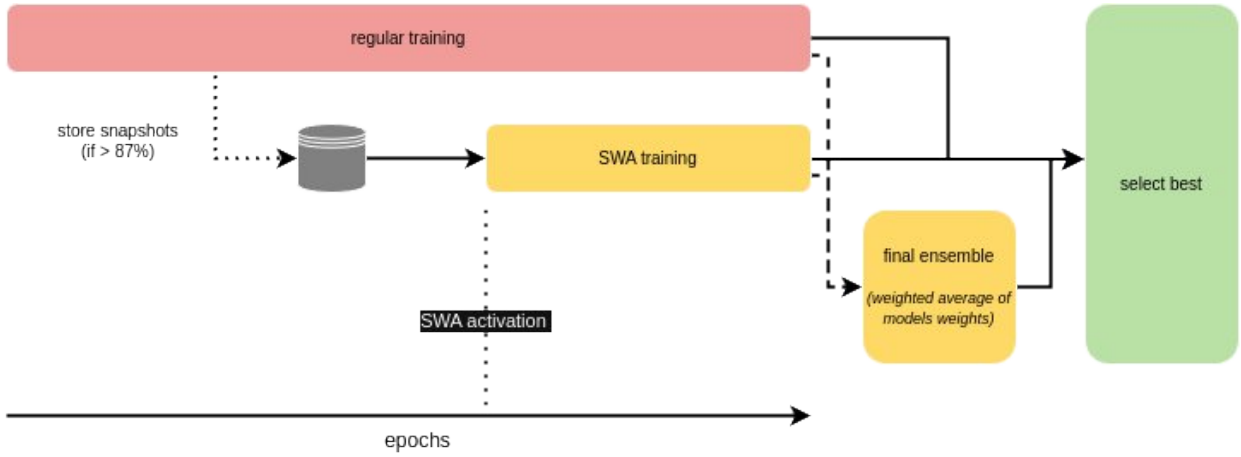


Figure 4.5: Ensemble strategy architecture showing SWA averaging, snapshot collection, and combination methods

4.3.4 Ablation Study Design

The ablation study framework represents a crucial component of our evaluation methodology, designed to quantify the individual contributions of each architectural component and training strategy employed in our hybrid approach. The systematic decomposition of our complete system enables precise understanding of which innovations drive performance improvements and guides future research directions by identifying the most impactful design decisions.

The ablation study design follows a hierarchical approach, beginning with major architectural components and progressively examining finer-grained design choices. This structure ensures that the most significant contributors to performance are identified first, while detailed analysis of component interactions provides insights into the synergistic effects that emerge from their combination. The complete ablation study encompasses focused individual experiments, each designed to isolate specific design decisions while maintaining all other factors constant.

Component-wise evaluation forms the foundation of our ablation study, systematically

removing or replacing major architectural components to quantify their individual contributions. The baseline configuration for component ablation employs a simplified CNN-only architecture processing single-frame radar data, providing a minimal but functional detection system. Sequential addition of components enables measurement of their incremental contributions while identifying potential interactions between different architectural elements.

The comprehensive ablation study and variant testing provide strong evidence that our architectural choices represent well-optimized solutions to the challenges of radar-based object detection. The systematic evaluation of alternatives confirms that the performance improvements achieved by our approach result from principled design decisions rather than random architectural variations, providing confidence in the generalizability of our findings to other radar detection tasks and datasets.

This comprehensive experimental framework, encompassing detailed baseline comparisons and systematic ablation studies, establishes a rigorous foundation for evaluating our proposed hybrid architecture. The next chapter will present the detailed design of our solution, building upon the methodological framework established here to demonstrate how each component addresses specific challenges in radar-based object detection while contributing to the overall system performance.

Summary

This chapter traces with details the research methodology employed throughout this thesis. The CRUW dataset was established as the primary benchmark, comprising synchronized camera and radar data captured in diverse driving scenarios with a 77 GHz FMCW radar. The evaluation framework was comprehensively defined, with particular emphasis on the Object Location Similarity (OLS) metric as the primary evaluation criterion, along with class-specific kappa thresholds and multi-threshold evaluation protocols. The experimental setup was described, including training infrastructure, data preprocessing pipelines, and augmentation strategies. The chapter presented the

advanced loss function design, detailing the multi-component loss architecture that combines multiclass focal loss with class-specific gamma values, regression loss for localization, and auxiliary losses for training stability. The ablation study design was described, establishing a hierarchical approach to systematically evaluate major architectural components and finer-grained design choices to identify and scrutinize the most impactful design decisions.

Chapter 5

Architecture

5.1 System Overview and Design Philosophy

The development of effective radar based object detection systems requires a fundamental understanding of both the unique characteristics of radar data and the limitations of existing processing approaches. This chapter presents our comprehensive solution architecture, designed to address the inherent challenges of automotive radar detection while leveraging recent advances in deep learning to achieve state-of-the-art performance. Our approach represents a paradigm shift from traditional sequential processing pipelines toward an integrated end-to-end learning framework that can extract maximum information from available radar measurements, with particular emphasis on temporal coherence and efficient feature extraction through dual architectural pathways.

5.1.1 Hybrid Architecture Motivation

The motivation for developing a hybrid architecture stems from careful analysis of the complementary strengths and limitations of different deep learning paradigms when applied to radar data processing. Traditional convolutional neural networks excel at extracting local spatial features through their translation-invariant kernels and hierarchical feature extraction capabilities, making them well suited for processing the grid structured nature of radar range-azimuth maps [23]. However, CNNs struggle to capture long-range dependencies that span large portions of the radar field of view [36], limiting their ability

to understand global scene context and object relationships that are crucial for robust detection.

The introduction of transformer architectures to computer vision has demonstrated remarkable success in capturing global relationships and long-range dependencies [37]. However, the computational complexity of full self-attention mechanisms presents significant challenges for real-time automotive applications, particularly when processing high-resolution radar data across multiple temporal frames. This computational burden motivated our exploration of alternative architectures that could maintain the benefits of global feature modeling while significantly reducing computational requirements.

Recent advances in MetaFormer architectures have shown that the effectiveness of transformer-like models stems not solely from the self-attention mechanism but from the overall architectural framework [135]. By replacing computationally expensive attention operations with simpler token mixing strategies such as pooling or depth-wise convolutions, MetaFormer variants achieve comparable performance with substantially reduced computational costs. This insight forms the foundation of our dual-pathway approach, where we provide both traditional transformer-based processing and efficient MetaFormer alternatives.

While temporal processing presents critical opportunities for enhanced detection through multi-frame analysis, this aspect is comprehensively addressed in Chapter 6, which details our temporal fusion architectures.

5.1.2 End-to-End Detection Pipeline

Our end-to-end detection pipeline represents a comprehensive approach to radar based object detection that processes raw radar measurements directly to produce final detection outputs without intermediate hand-crafted features or traditional radar processing steps. This design philosophy enables the network to learn appropriate feature representations directly from data, potentially discovering patterns and

relationships that conventional processing might overlook. Figure 5.1 illustrates the input and output of the pipeline.

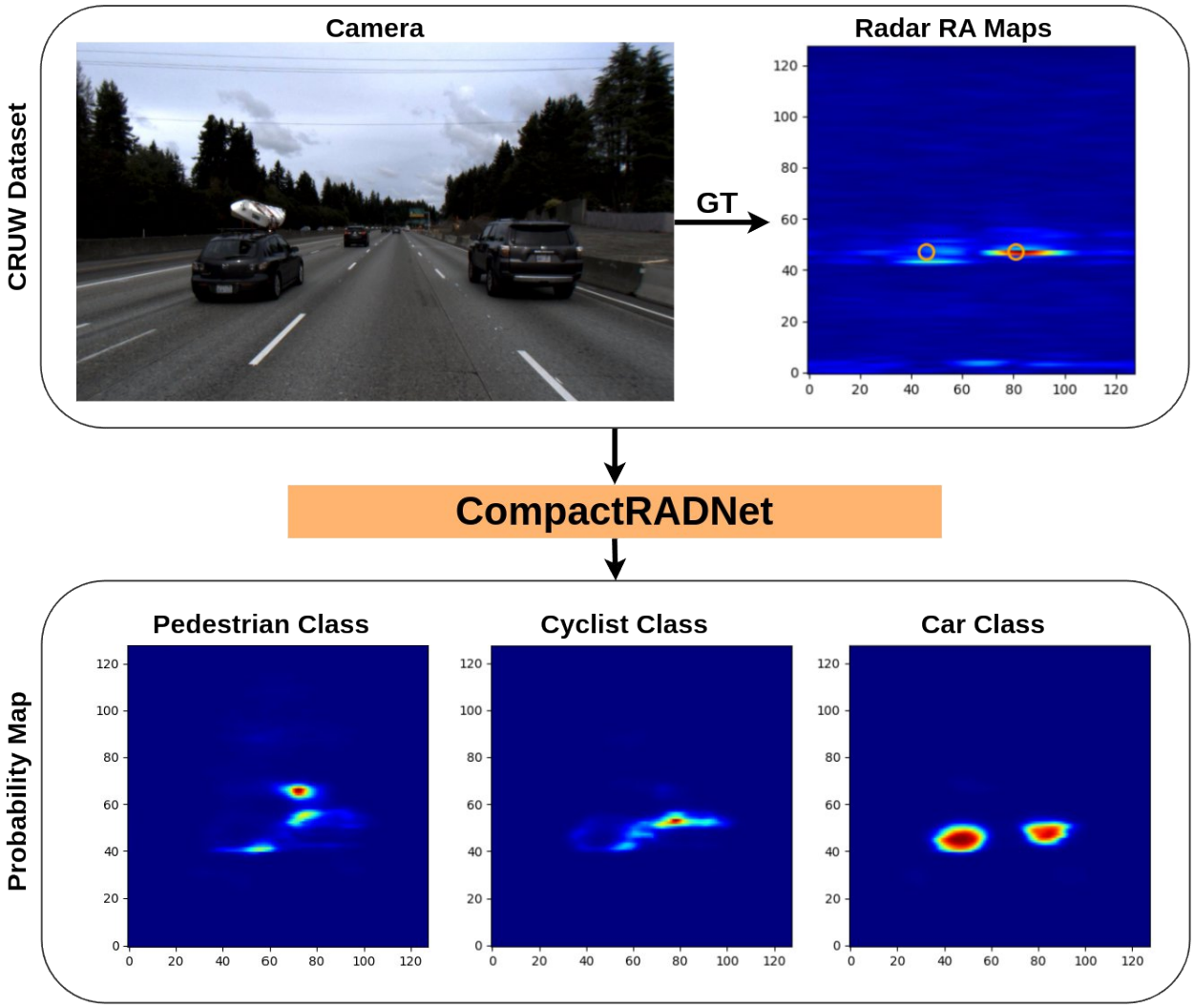


Figure 5.1: CompactRADNet input and output

The pipeline architecture is designed to seamlessly integrate with temporal processing modules detailed in Chapter 6, enabling both single-frame and multi-frame configurations. For single-frame operation, the system processes standard range-azimuth representations, while multi-frame configurations stack temporal sequences to capture motion dynamics and improve detection robustness. When operating in multi-frame mode, the pipeline incorporates the temporal fusion mechanisms including the Adaptive

Quadratic ReLU (AQR) activation function, specifically designed for radar signal characteristics, which provides input-dependent modulation essential for handling varying radar signal strengths across different driving scenarios, all of which described in Chapter 6.

Following initial feature extraction, the pipeline employs a Feature Pyramid Network to create multi-scale representations that can effectively detect objects of varying sizes. This multi-scale approach is particularly important for radar data where object signatures can vary dramatically based on range and viewing angle. The FPN component ensures that both large vehicles and smaller vulnerable road users receive appropriate feature representations at suitable scales.

The core processing pathway then diverges based on the selected configuration. The MetaFormer pathway employs efficient token mixing operations that maintain global receptive fields while dramatically reducing computational complexity compared to traditional attention mechanisms. The alternative transformer pathway utilizes spatial window attention for applications requiring maximum accuracy where computational resources permit. Both pathways benefit from careful architectural design that incorporates positional encodings, normalization strategies, and residual connections optimized for radar data characteristics.

The detection pipeline concludes with specialized detection heads that perform both classification and localization tasks. The main detection heads process high-resolution feature maps to generate precise object predictions, while auxiliary heads operating at intermediate feature levels provide additional supervision during training and can be used for computational optimization during inference. This multi-head design improves gradient flow during training and enables flexible deployment strategies based on available computational resources.

The entire pipeline operates in a fully differentiable manner, enabling end-to-end training with standard gradient-based optimization methods. This unified training approach allows all components to be jointly optimized for the final detection objective, resulting in features

and representations specifically tailored to the radar detection task rather than generic visual features that may not capture radar-specific phenomena effectively.

Figure 5.2 illustrates the complete end-to-end detection pipeline, showing the flow of information from raw radar data through the various processing stages to final detection outputs.

5.2 Architecture Components

5.2.1 Radar Stem

The radar stem serves as the critical entry point for raw radar data into our deep learning pipeline, performing initial feature extraction and dimensional transformation while preserving the unique characteristics of radar measurements. Our architecture provides multiple stem configurations designed to handle both single-frame (discussed in this chapter) and multi-frame temporal sequences (discussed in chapter 6), each optimized for specific operational requirements and computational constraints.

5.2.1.1 Single-frame Radar Stem

The single-frame radar stem processes individual range-azimuth maps through a carefully designed 3D convolutional architecture that treats the multiple radar channels as a depth dimension. This approach preserves the inherent structure of radar data while enabling efficient feature extraction across all available channels simultaneously. The design philosophy prioritizes the preservation of fine-grained spatial patterns while progressively building hierarchical feature representations suitable for object detection tasks.

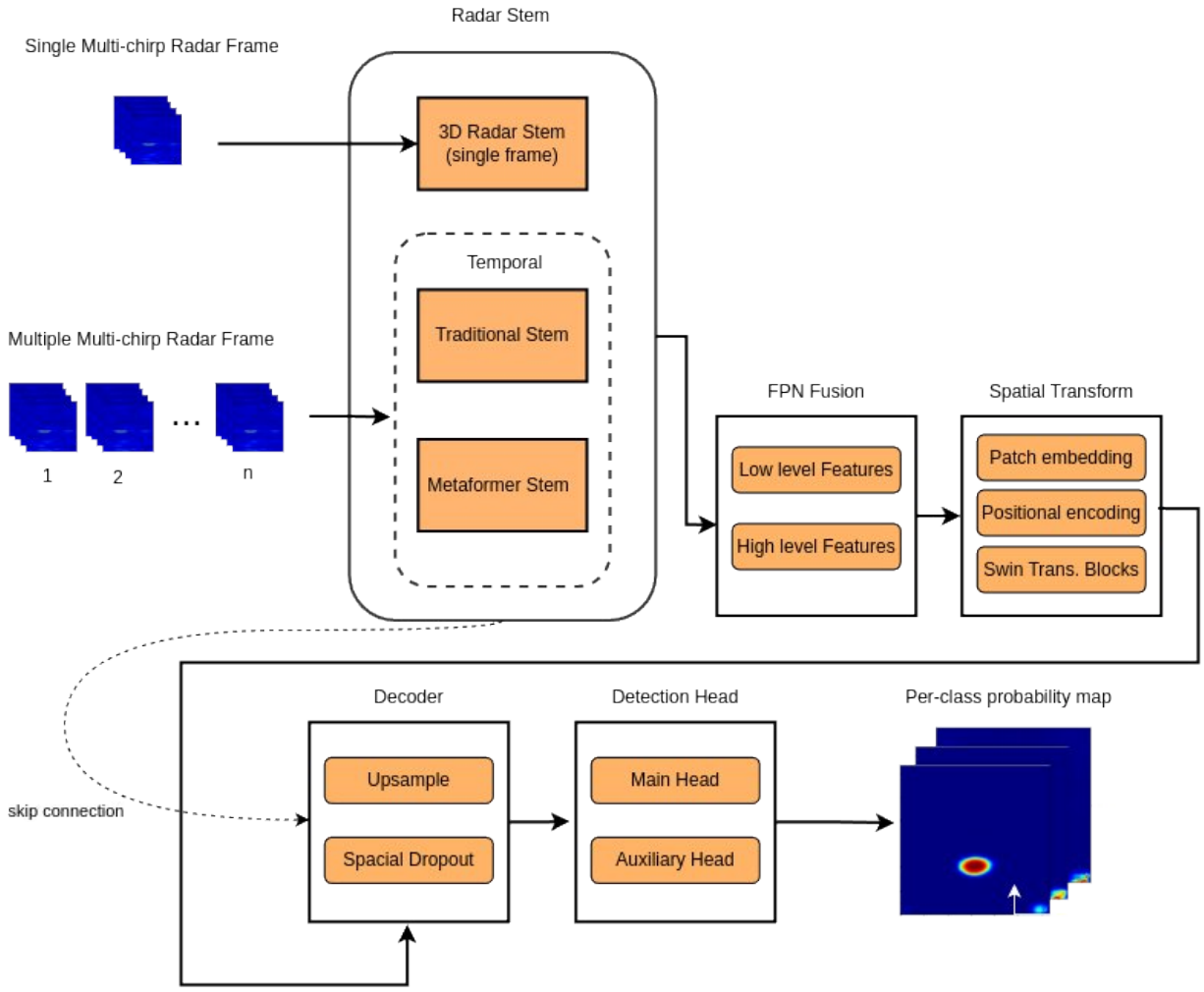


Figure 5.2: End-to-end detection pipeline architecture showing data flow from radar input to final detections

The architecture begins with grouped 3D convolutions that process each radar channel semi-independently before fusion, recognizing that different channels may contain complementary information about the same scene. This grouped processing approach reduces computational complexity while maintaining the ability to extract channel-specific features that may be important for subsequent processing stages. The grouping strategy typically processes pairs or triplets of related channels together, such as different chirp configurations or polarization modes, allowing the network to learn channel-specific

feature extractors while controlling parameter growth. The initial convolution employs small 3×3×3 kernels to capture fine-grained spatial patterns while gradually expanding the receptive field through subsequent layers.

The stem incorporates a multi-stage feature extraction process with careful attention to information preservation and gradual abstraction. The first stage focuses on noise suppression and initial pattern detection, employing shallow convolutions with limited spatial pooling to maintain high-resolution information. This stage effectively serves as a learnable preprocessing step that can adapt to different radar configurations and environmental conditions during training. The convolutions at this stage use relatively few filters to avoid early information bottlenecks while providing sufficient capacity for basic pattern detection.

Following initial feature extraction, the stem incorporates normalization and activation functions specifically chosen for radar data characteristics. Group normalization provides stable training dynamics even with small batch sizes, while careful placement of activation functions ensures that both positive and negative radar returns are appropriately processed. The normalization is applied across groups of channels rather than spatial dimensions, preserving the spatial statistics that prove important for radar interpretation. The activation functions employ a combination of ReLU and LeakyReLU variants, with LeakyReLU preferred in early layers to preserve negative radar returns that may contain important information about noise characteristics or interference patterns.

The architectural progression through the stem follows a carefully designed strategy of gradual spatial reduction and channel expansion. Each stage roughly doubles the number of feature channels while optionally reducing spatial dimensions by a factor of two, creating an inverse pyramid structure that trades spatial resolution for semantic richness. This design ensures computational efficiency while maintaining sufficient spatial detail for accurate object localization. The downsampling operations employ strided convolutions rather than pooling operations, allowing the network to learn effective downsampling filters that preserve important spatial information while reducing resolution.

Unlike pooling operations, which apply fixed functions (maximum or average) without trainable parameters, strided convolutions contain learnable weights that are optimized during training. This enables the network to develop task-specific downsampling behavior, learning which spatial features to preserve rather than applying generic reduction operations that may discard information relevant to radar target detection.

Residual connections play a crucial role in the stem architecture, particularly in deeper configurations where gradient flow becomes challenging. These connections bypass one or more convolutional stages, providing direct gradient paths that facilitate training while enabling the network to learn incremental refinements rather than complete transformations. The residual connections employ projection shortcuts when channel dimensions change, ensuring dimensional compatibility while providing additional learnable parameters that can adapt the skip connections to specific requirements.

The stem design also incorporates specialized components for handling the unique characteristics of radar data. Radar measurements often exhibit range-dependent variations in signal strength and noise characteristics, motivating the inclusion of range-adaptive normalization mechanisms. These components learn position-dependent scaling factors that can compensate for range-related signal variations, improving feature consistency across the entire radar field of view. Similarly, azimuth-dependent corrections address the varying antenna patterns and multipath effects that can create systematic biases in different angular regions.

The final stages of the single-frame stem prepare features for subsequent processing by ensuring appropriate spatial dimensions and channel counts. The output typically provides features at two spatial scales: a higher-resolution output at 1/2 the original resolution for fine-grained detection and a lower-resolution output at 1/4 resolution for efficient global processing. This multi-scale output strategy enables the subsequent Feature Pyramid Network to operate more effectively by providing pre-computed features at appropriate scales. The channel dimensions are carefully chosen to balance representational capacity with computational requirements, typically outputting 64

channels that provide sufficient information for downstream processing while maintaining reasonable memory requirements.

When operating in multi-frame mode, the pipeline incorporates the temporal fusion mechanisms described in Chapter 6.

Figure 5.3 demonstrates the internal architecture of the Radar 3D Stem, illustrating the 3D convolution pathways, channel attention mechanisms, and multi-scale output generation.

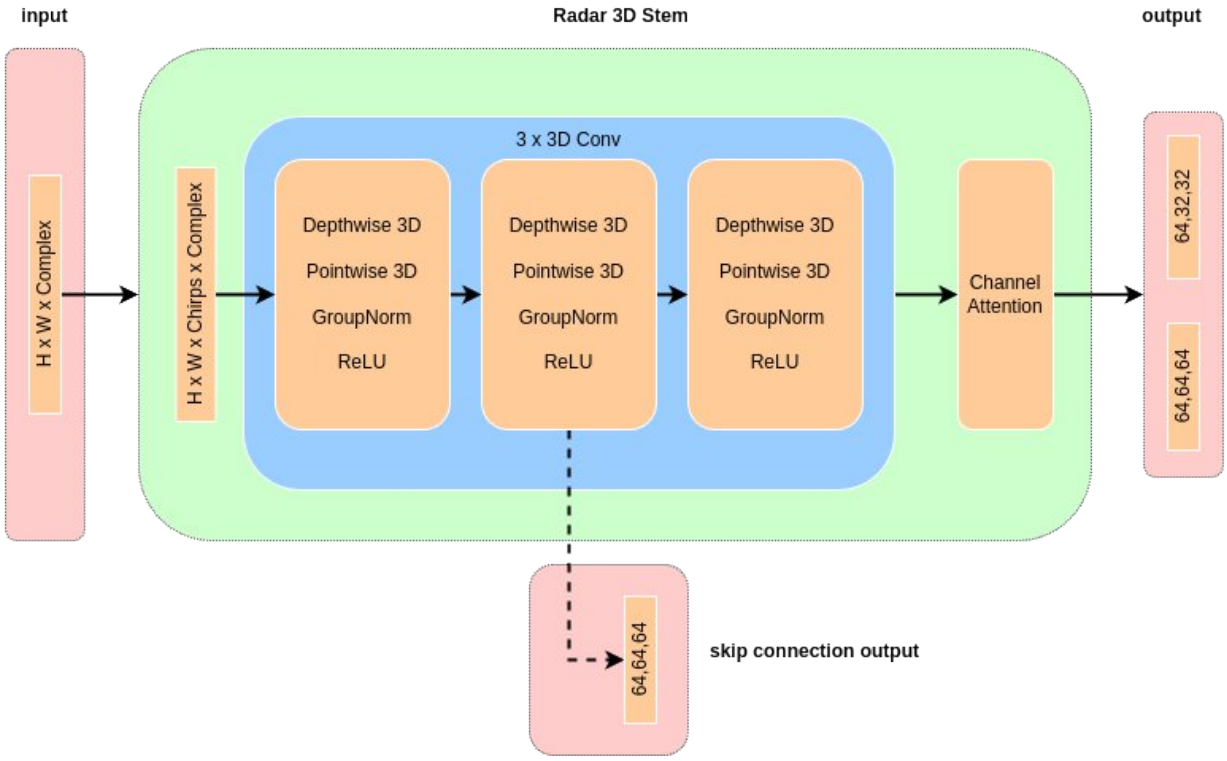


Figure 5.3: Radar 3D Stem internal architecture showing 3D convolution pathways

5.2.2 Feature Pyramid Network Integration

The Feature Pyramid Network component remains largely unchanged from the original architecture but plays an increasingly important role when processing multi-frame inputs

[74]. The FPN creates multi-scale feature representations by combining high-resolution features with semantically rich deeper features through lateral connections and top-down pathways.

The integration with temporal processing requires careful consideration of feature dimensions and semantic alignment. Features from different temporal processing paths are projected to consistent dimensions before FPN fusion, ensuring compatible representations regardless of the temporal processing strategy employed. The lateral connections incorporate 1x1 convolutions for dimension matching while preserving the semantic content of features at each scale.

The FPN particularly benefits radar object detection by addressing the scale variation challenge inherent in automotive scenarios. Large vehicles may dominate entire radar frames when nearby, while distant objects appear as minimal signatures requiring fine-grained spatial resolution for accurate localization. The pyramid structure ensures appropriate feature representations exist for all object scales while maintaining computational efficiency through feature reuse.

The mathematical formulation of the FPN feature combination follows:

$$F_l' = F_l + \mathrm{Upsample}\big(L(F_{l+1})\big)$$

where F'_l represents the enhanced features at pyramid level l, F_l represents the original features at that level, L denotes the lateral connection transformation (typically a 1×1 convolution), and the upsample operation provides spatial alignment between different pyramid levels. This formulation ensures that each pyramid level benefits from both its native spatial resolution and the semantic information extracted at lower resolutions.

Top-down pathway processing propagates semantic information from the deepest (lowest resolution) pyramid levels to the shallowest (highest resolution) levels, enabling fine-grained spatial localization to benefit from high-level semantic understanding. The top-down processing employs nearest-neighbor upsampling followed by convolution

operations to maintain spatial precision while propagating semantic information. This approach proves more effective than learned upsampling strategies like transposed convolution because it preserves spatial precision that is crucial for accurate object localization in radar data.

Figure 5.4 illustrates the FPN architecture integrated within our hybrid system, showing the multi-scale feature extraction, lateral connections, and fusion mechanism.

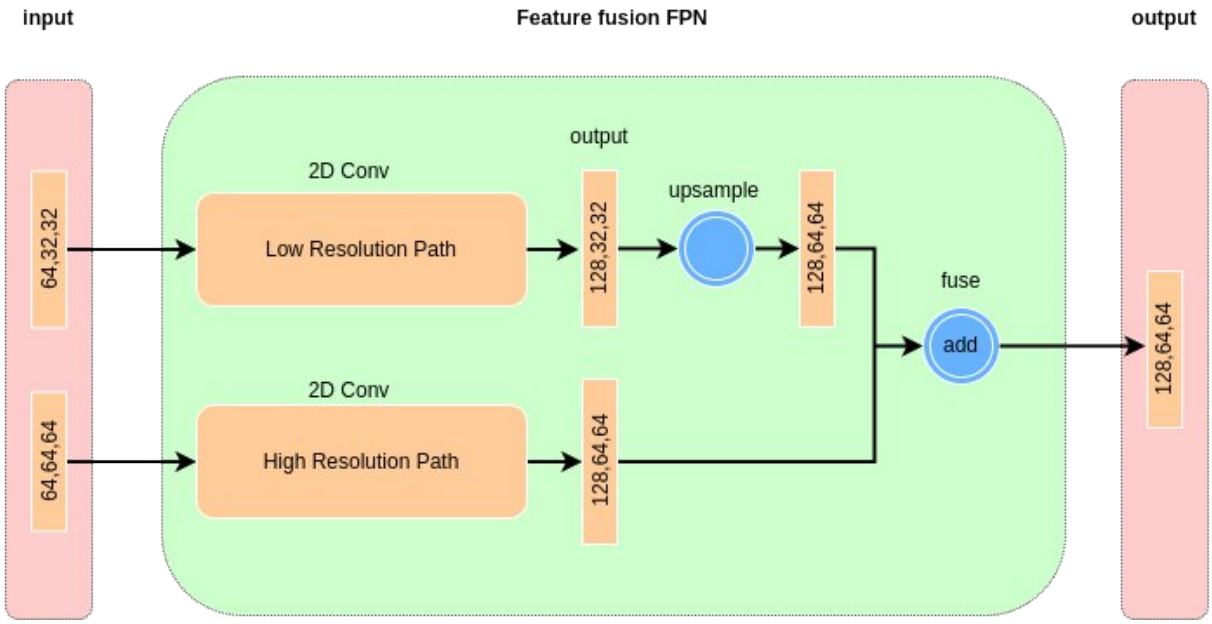


Figure 5.4: Feature Pyramid Network architecture showing multi-scale processing and cross-scale information flow

5.2.3 Patch Embedding and Positional Encoding

The transition from convolutional feature maps to transformer-compatible representations requires careful patch embedding design that preserves spatial relationships while

enabling efficient global processing. Our patch embedding strategy employs non-overlapping spatial patches extracted through strided convolutions, creating a sequence of patch tokens suitable for subsequent transformer or MetaFormer processing.

The patch size selection balances several competing factors including computational efficiency, spatial resolution preservation, and effective receptive field size. Larger patches reduce the sequence length and computational requirements but may lose fine-grained spatial details important for precise localization. Our architecture employs adaptive patch sizing based on the input resolution and downstream task requirements, typically using 4x4 or 8x8 patches for standard radar resolutions.

Positional encoding proves particularly critical for radar data where absolute spatial positions carry important semantic meaning related to range and azimuth. Unlike natural images where relative positions often suffice, radar measurements require absolute position awareness to properly interpret range-dependent characteristics and geometric relationships. Our positional encoding scheme employs learnable 2D positional embeddings that can adapt to the specific characteristics of radar data during training.

The encoding mechanism separately models horizontal and vertical positions, recognizing the different semantic meanings of range and azimuth dimensions in radar data. This decomposed approach enables more efficient learning while providing the flexibility to handle different radar configurations and resolutions. The positional information is injected additively after patch embedding, preserving the original feature content while providing essential spatial context.

The mathematical formulation of the positional encoding addition follows:

$$F'_{i,j} = F_{i,j} + P_{i,j}$$

where $F'_{i,j}$ represents the position-encoded feature at spatial location (i,j), $F_{i,j}$ is the original feature representation, and $P_{i,j}$ is the learnable positional embedding for that location. This additive approach preserves the original feature information while providing spatial

context that enables effective attention computation.

Figure 5.5 illustrates the patch embedding and positional encoding strategy employed in our architecture, showing the learnable embedding patterns and their integration with the transformer attention mechanisms.

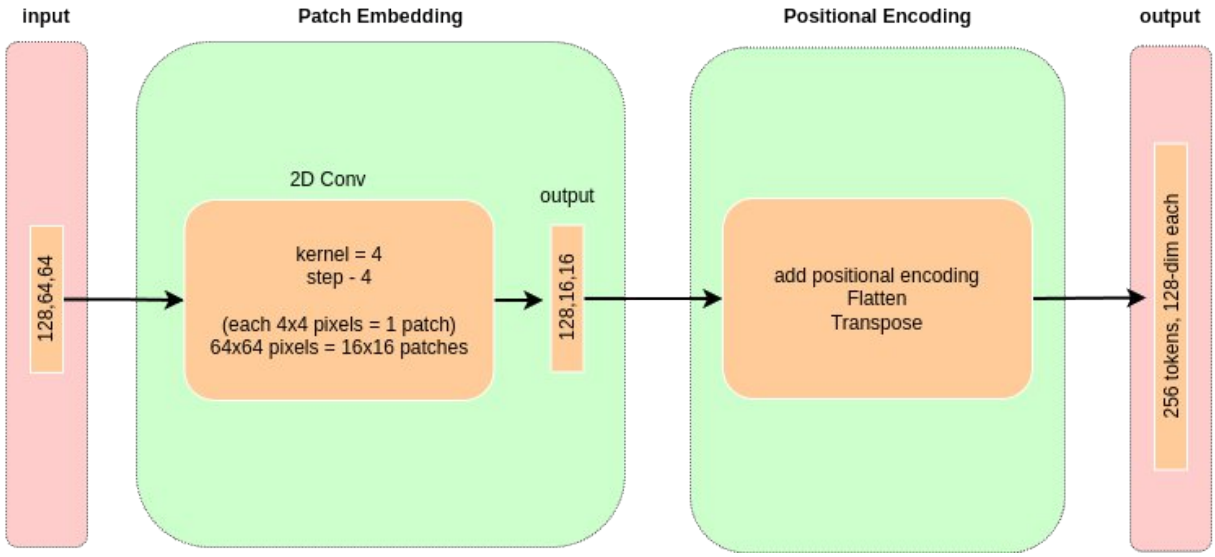


Figure 5.5: Positional encoding and patch embedding visualization showing spatial embedding patterns

5.2.4 Spatial Window Processing

The spatial processing stage represents the core feature extraction component of our architecture, where global and local spatial relationships are modeled to create rich feature representations for object detection. We provide two distinct pathways that offer different trade-offs between computational efficiency and modeling capacity.

5.2.4.1 Metaformer Path

The Metaformer spatial processing path replaces traditional self-attention mechanisms with more efficient token mixing operations while maintaining the beneficial architectural

patterns of transformer models. This approach dramatically reduces computational complexity while achieving comparable performance for many radar detection scenarios.

The spatial Metaformer blocks operate on windows of patches, similar to the Swin Transformer architecture, but employ alternative token mixing strategies. The pooling variant uses spatial average pooling within each window, effectively capturing regional statistics that prove particularly effective for radar data where object signatures often manifest as regional patterns rather than precise point features. The pooling operation is enhanced with learnable channel-wise weights that allow the model to emphasize informative channels while suppressing noise.

The convolutional variant employs depth-wise separable convolutions within each window, providing local spatial modeling with minimal computational overhead. This approach proves particularly effective for capturing the blob-like signatures characteristic of radar object returns while maintaining translation equivariance beneficial for object detection tasks. The kernel sizes are carefully chosen to span typical object signatures while avoiding excessive spatial smoothing that could compromise localization accuracy.

Both variants incorporate shifted window mechanisms similar to Swin Transformer, enabling cross-window information exchange crucial for modeling objects that span window boundaries. The shifting strategy alternates between regular and shifted window partitions across layers, ensuring comprehensive spatial coverage while maintaining computational efficiency. This design enables effective global modeling through successive local operations, particularly important for radar data where long-range spatial relationships indicate important scene context.

5.2.4.2 Transformer Path

The transformer spatial processing path implements full self-attention mechanisms within spatial windows, providing maximum modeling capacity for scenarios where computational resources permit. This pathway closely follows the Swin Transformer [72] architecture but incorporates specific adaptations for radar data characteristics.

The window attention mechanism computes pairwise relationships between all patches within each window, enabling fine-grained spatial modeling that can capture complex object patterns and relationships. The multi-head design allows different heads to specialize in different types of spatial relationships, such as local texture patterns versus regional context. The attention computation incorporates relative position biases that prove particularly important for radar data where spatial relationships have consistent geometric interpretations.

The transformer blocks employ careful normalization and residual connection strategies optimized through extensive experimentation on radar data. Pre-normalization ensures stable training dynamics while post-attention dropout provides regularization without compromising the learned attention patterns. The MLP components following attention use expanded intermediate dimensions to increase model capacity while gated activation functions provide additional non-linearity beneficial for complex pattern recognition.

5.2.5 Decoder Architecture

The decoder component transforms the encoded features back to high-resolution spatial representations suitable for object detection. Our decoder design emphasizes efficiency while ensuring sufficient spatial detail for accurate object localization, employing a progressive upsampling strategy that carefully balances computational cost with detection performance.

5.2.5.1 Upsampling and Feature Reconstruction

The decoder uses a simple upsampling approach: bilinear interpolation followed by depthwise separable convolutions. The main components are a single upsampling layer, spatial dropout (p=0.1), and basic residual blocks with channel attention.

The upsampling implementation employs bilinear interpolation to ensure consistent spatial alignment and prevent artifacts that can occur with other interpolation methods.

This choice provides smooth spatial reconstruction while maintaining the geometric relationships essential for accurate object localization in radar coordinate systems. The bilinear interpolation is followed by a depthwise separable convolution that reduces computational overhead while providing sufficient capacity for feature refinement.

Learned upsampling strategies were considered as alternatives to the bilinear approach, including transposed convolutions and sub-pixel convolution methods[135]. However, empirical evaluation revealed that these more complex approaches provided minimal performance benefits while increasing computational requirements and potentially introducing spatial artifacts. The bilinear interpolation approach provides the strong balance between reconstruction quality and computational efficiency for our radar detection application.

5.2.5.2 Spatial Dropout and Regularization

Spatial dropout integration within the decoder provides structured regularization that addresses overfitting while preserving the spatial coherence essential for effective feature reconstruction. Unlike standard dropout that randomly zeroes individual feature elements, spatial dropout operates on entire feature channels, maintaining the spatial structure within each channel while providing regularization across the channel dimension.

The spatial dropout implementation applies random channel masking with probability 0.1 during training, providing sufficient regularization to prevent overfitting without significantly degrading the feature reconstruction quality. The dropout is applied after the upsampling operation but before the refinement processing, ensuring that regularization occurs at the appropriate stage of feature reconstruction where it can provide maximum benefit without interfering with critical spatial relationships.

The spatial dropout mechanism proves particularly valuable for radar processing where spatial relationships carry crucial information about object geometry and radar signature patterns. By preserving spatial structure within channels while providing cross-channel regularization, spatial dropout enables the decoder to learn robust feature representations

that generalize effectively to new radar measurements while maintaining the spatial precision necessary for accurate object localization.

Dropout scheduling strategies were evaluated to optimize the regularization effects throughout training. Early training phases employ higher dropout rates (0.15) to encourage robust feature learning, while later phases reduce dropout rates (0.05) to enable fine-tuning of spatial relationships. However, empirical evaluation revealed that fixed dropout rates (0.1) provide robust performance by maintaining consistent regularization throughout training.

5.2.5.3 Residual Connections and Feature Enhancement

Residual connection implementation within the decoder enables effective gradient flow while combining features from different processing stages to create comprehensive representations that leverage both low-level spatial details and high-level semantic understanding. The residual architecture follows the standard pattern of adding input features to processed features, enabling the decoder to learn refinements rather than complete feature reconstructions.

The refinement block architecture employs a sequence of convolutional operations with group normalization and ReLU activation, designed to enhance spatial features while maintaining computational efficiency. The block begins with a 3×3 convolution that processes spatial relationships, followed by group normalization with 8 groups that provides stable training dynamics for the feature dimensions used in our architecture. The ReLU activation introduces nonlinearity while the final 3×3 convolution produces the refinement signal.

Channel attention mechanisms are integrated within the refinement blocks to enable adaptive feature selection based on the importance of different feature channels for the detection task. The attention mechanism computes channel-wise importance weights using global average pooling followed by a small multi-layer perceptron, enabling the decoder to emphasize the most informative features while suppressing those that may

be dominated by noise or artifacts from the upsampling process.

Skip connections from the CNN backbone provide additional pathways for preserving low-level spatial information that may be lost during the transformer processing stages. These connections enable direct flow of high-resolution features from early processing stages to the decoder, ensuring that fine-grained spatial details necessary for precise object localization are preserved throughout the processing pipeline.

5.2.5.4 Multi-scale Feature Integration

The decoder architecture incorporates multi-scale feature integration mechanisms that combine information from different levels of the feature hierarchy to create comprehensive representations that support robust object detection across the wide range of scales encountered in automotive radar data. This integration addresses the challenge that objects at different distances require different types of feature representations for effective detection performance.

Multi-scale integration begins with the Feature Pyramid Network outputs that provide features at different spatial resolutions, each optimized for detecting objects at appropriate scales. The decoder processes these multi-scale features through parallel pathways that apply scale-appropriate processing before combining them into unified representations that leverage information from all scales simultaneously.

Scale-specific processing pathways employ different architectural configurations optimized for the characteristics of features at each scale. High-resolution features receive processing that emphasizes spatial precision and fine-grained pattern recognition, while lower-resolution features undergo processing that focuses on semantic understanding and global context integration. This scale-appropriate processing ensures that each feature scale contributes effectively to the final detection decisions.

Cross-scale interaction mechanisms enable information exchange between different scales during the integration process, allowing features at one scale to influence and enhance features at other scales. These interactions are implemented through attention

mechanisms that compute relevance scores between features at different scales, enabling adaptive information sharing that improves the quality of the integrated representations.

5.2.5.5 Output Projection and Feature Preparation

The final stages of the decoder prepare features for consumption by the detection heads through output projection operations that ensure appropriate feature dimensions and characteristics for multi-task detection objectives. The projection process must balance the needs of classification, regression, and confidence estimation tasks while maintaining computational efficiency and training stability.

Output projection employs 1×1 convolutions that transform the integrated decoder features into the specific dimensional requirements of the detection heads. The projection maintains the spatial resolution recovered through upsampling while adjusting the channel dimensions to match the detection head architectures. This projection enables flexible adaptation to different detection head designs while preserving the essential spatial and semantic information developed through the decoder processing.

Feature normalization within the output projection ensures stable training dynamics and consistent feature scales for the detection heads. Group normalization is applied after the projection convolutions to provide stable statistics for subsequent processing while maintaining the spatial structure essential for accurate object localization. The normalization parameters are learned during training to optimize the feature distributions for the specific detection tasks.

Activation function selection for the output projection employs ReLU activation that provides nonlinearity while maintaining efficient computation and stable gradients. Alternative activation functions including GELU and Swish were evaluated, but ReLU provided better performance for the feature projection task while maintaining computational efficiency suitable for automotive deployment.

The decoder output provides feature maps with dimensions (B, 32, H, W) that serve as

input to both the main and auxiliary detection heads. These features combine the spatial precision recovered through upsampling with the semantic understanding developed through transformer processing, enabling effective multi-task optimization that supports classification, regression, and confidence estimation objectives simultaneously.

Figure 5.6 illustrates the complete decoder architecture, showing the upsampling pathways, feature refinement mechanisms, and multi-scale integration strategies that transform transformer features into detection-ready representations.

The decoder architecture represents a crucial innovation that enables effective integration of transformer-based global context modeling with the spatial precision requirements of object detection tasks. Through meaningful upsampling, feature refinement, and multi-scale integration mechanisms, the decoder successfully bridges the gap between patch-based transformer representations and the dense spatial predictions required for automotive radar object detection. The careful balance of computational efficiency with detection performance makes this decoder design suitable for practical deployment while achieving state-of-the-art detection capabilities.

5.2.6 Detection Heads

The detection head represents the final component of our architecture, responsible for converting learned feature representations into actionable object detection outputs. The design is straightforward and addresses the unique characteristics of radar measurements and automotive application requirements through a classification-focused approach with auxiliary regression supervision that enhances feature learning without directly contributing to detection decisions.

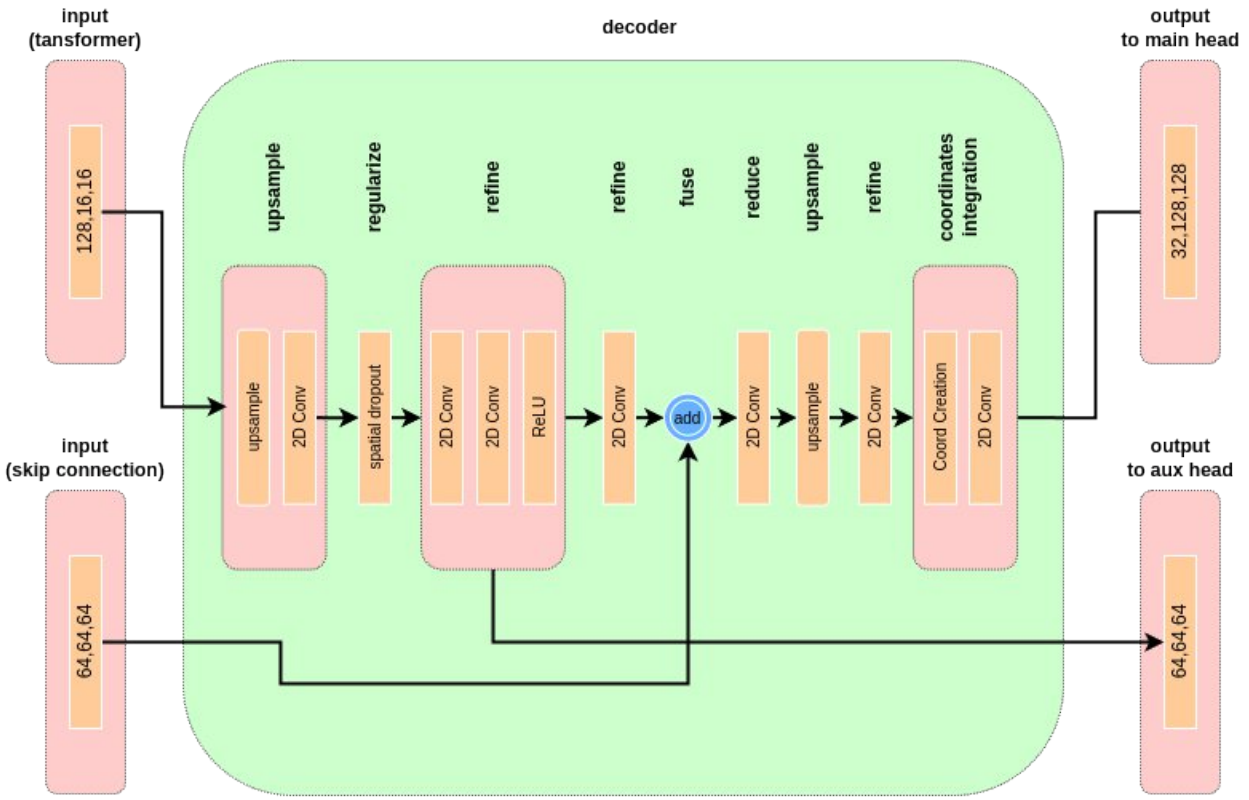


Figure 5.6: Decoder architecture showing upsampling, refinement, and multi-scale feature integration components

The dual-pathway detection architecture employs primary and auxiliary processing pathways that operate at different feature resolutions and processing depths. The primary detection heads process the final refined feature representations from the hybrid CNN-Transformer backbone, while auxiliary detection heads operate on intermediate feature representations to provide deep supervision and improve gradient flow during training. This approach enables effective training of deep architectures while focusing inference on classification-based center-point detection strategies.

5.2.6.1 Classification Head Architecture

The classification head serves as the primary detection mechanism, converting refined spatial features into class-specific logits that are subsequently processed into probability heat maps for object detection. The classification head addresses the challenge of

distinguishing between different object classes based on radar signatures through dense spatial prediction maps that preserve full spatial resolution while enabling pixel-level classification decisions.

Vehicle classification relies primarily on strong, coherent reflections from metallic surfaces and characteristic geometric patterns, while pedestrian and cyclist classification must extract weak, distributed signatures from complex scattering patterns. The classification head employs a lightweight 1×1 convolution that transforms refined spatial features into class-specific logits without introducing excessive parameters that might lead to overfitting.

The classification head outputs raw logits rather than normalized probabilities during training, enabling flexible loss computation strategies and maintaining robust gradient flow characteristics. During validation and inference, these logits are converted to probability heat maps through sigmoid activation, creating interpretable confidence maps that support center-point detection strategies through local maxima identification.

The detection pipeline employs a simple local maxima finding algorithm that identifies object centers as peaks in the probability heat maps. Connected component analysis first identifies regions of elevated probability, followed by iterative peak detection within each component that can identify multiple objects within clustered regions. This approach proves particularly effective for radar data where objects may create complex scattering patterns that result in distributed probability responses.

5.2.6.2 Regression Head Architecture and Training Role

The regression head provides auxiliary supervision that enhances feature learning quality during training without directly contributing to detection decisions during inference. The regression head employs the same lightweight 1×1 convolution architecture as the classification head but outputs 2-channel predictions representing spatial coordinate information that supports multi-task learning objectives.

The regression supervision forces the network to learn features that encode precise spatial relationships, improving the overall quality of feature representations that benefit classification performance. Although regression predictions are not used during inference, the regression training objective creates synergistic learning effects where spatial reasoning capabilities developed for coordinate prediction enhance the discriminative power of features used for classification.

This training-only regression approach provides the benefits of multi-task learning without the complexity of integrating regression outputs into the detection pipeline. The classification-focused inference strategy simplifies post-processing requirements while maintaining the feature learning advantages provided by multi-task supervision during training.

5.2.6.3 Coordinate Convolution Enhancement

Coordinate convolution integration within the detection pipeline addresses the challenge of spatial reasoning in radar coordinate systems. Standard convolutional operations are translation invariant and cannot distinguish between different absolute spatial locations without additional context. Coordinate convolution provides explicit spatial coordinate information to the detection heads, enabling location-aware processing that adapts to varying radar characteristics across the field of view.

The coordinate convolution implementation concatenates normalized spatial coordinates with the refined feature representations before final detection processing. The coordinates consist of normalized x and y values ranging from -1 to +1 across the spatial dimensions, providing explicit spatial context for each feature location. A 1×1 convolution then integrates this coordinate information with the base features, enabling the detection heads to adapt their processing based on absolute spatial location within the radar field of view.

This coordinate enhancement proves particularly valuable for handling the varying characteristics of radar measurements across different parts of the detection area, where

range-dependent attenuation and angular resolution effects can significantly impact signal characteristics. The coordinate information enables the detection heads to learn location-specific processing strategies that account for these systematic variations.

5.2.6.4 Auxiliary Detection Head Architecture

The auxiliary detection head architecture provides deep supervision that improves training dynamics and enhances feature learning throughout the network hierarchy. The auxiliary heads operate on intermediate feature representations from the decoder pathway at reduced spatial resolution, encouraging the learning of meaningful features at multiple processing depths while contributing to improved gradient flow throughout the deep architecture.

The auxiliary head design mirrors the primary detection head architecture with separate classification and regression components that process intermediate features at different scales and abstraction levels. Like the primary regression head, the auxiliary regression outputs serve purely as training supervision and do not contribute to inference decisions.

Critical to the auxiliary supervision strategy is the upsampling of auxiliary predictions to match the input resolution for loss computation. The auxiliary outputs are processed through bilinear interpolation to restore full spatial resolution, ensuring that auxiliary supervision operates at the same spatial scale as the primary supervision and enabling direct comparison and consistent gradient computation across different supervision levels.

The auxiliary supervision contributes to improved gradient flow throughout the deep architecture while encouraging the learning of features that are useful for detection at multiple spatial scales. The auxiliary loss weight provides significant but not dominant supervision that enhances training stability without overwhelming the primary learning objectives.

5.2.6.5 Multi-task Learning Integration

Multi-task learning integration enables joint optimization of classification and regression objectives through both primary and auxiliary supervision pathways, ensuring that feature learning benefits from spatial reasoning capabilities developed through regression training while maintaining focus on classification-based detection during inference. The multi-task loss function combines classification loss, regression loss, and auxiliary losses with carefully chosen weighting factors that balance the different objectives.

The multi-task loss incorporates both primary and auxiliary outputs where classification losses directly optimize for detection performance while regression losses enhance feature quality through spatial reasoning supervision. The auxiliary loss weight typically provides substantial contribution to training stability while regression supervision improves feature representations without complicating the inference pipeline.

This multi-task formulation ensures that all training objectives are optimized jointly, creating synergistic learning effects where spatial understanding developed through regression training enhances classification performance. The auxiliary supervision provides particularly important benefits during early training phases when deep networks may struggle with gradient flow and feature learning at intermediate layers.

5.2.6.6 Detection Pipeline and Inference Strategy

The detection pipeline focuses exclusively on classification outputs during inference, converting raw logits to probability heat maps that support robust center-point detection strategies. The classification logits are processed through sigmoid activation to create interpretable probability maps for each object class, enabling flexible threshold-based detection decisions.

The detection algorithm employs connected component analysis to identify regions of elevated probability, followed by iterative peak detection within each component that can identify multiple objects within clustered regions. This approach addresses the challenge of closely spaced objects that may create overlapping probability responses, particularly

important for radar data where multiple objects may contribute to complex scattering patterns.

Distance-based non-maximum suppression provides final detection refinement by removing duplicate detections that fall within specified distance thresholds. This class-agnostic suppression strategy proves effective for radar applications where different object classes may have similar spatial extents and the priority is preventing duplicate detections regardless of class assignments.

The inference strategy deliberately avoids regression output utilization, simplifying post-processing requirements while maintaining detection accuracy through the enhanced feature representations developed during multi-task training. This approach provides the benefits of regression supervision for feature learning while avoiding the complexity of multi-output fusion during inference.

5.2.6.7 Output Format and Processing Integration

The detection head architecture produces four output tensors during training that serve different roles in the learning and inference pipeline. Primary classification and regression outputs provide the main training supervision and inference capabilities, while auxiliary outputs enhance training dynamics through deep supervision at intermediate processing stages.

During training, all four output tensors contribute to loss computation through the multi-task learning framework where auxiliary outputs provide deep supervision and regression outputs enhance feature learning quality. During validation and inference, only the primary classification outputs are processed for actual detection decisions through probability heat map generation and local maxima detection.

This selective output utilization strategy enables effective multi-task training while maintaining simplicity during inference. The classification-focused detection approach provides robust performance for automotive radar applications while avoiding the complexity associated with multi-output fusion strategies that may introduce additional

sources of error or computational overhead.

This comprehensive solution architecture represents a strong approach to radar-based object detection that addresses the unique challenges of automotive radar through carefully designed hybrid architectures, advanced loss functions, and optimization strategies. The integration of CNN and Transformer components enables effective local and global feature extraction, while the multi-component loss design addresses class imbalance and multi-task optimization requirements. The advanced training strategies ensure robust learning and generalization despite the challenges posed by limited radar datasets and the complex nature of radar measurements.

Having established the core spatial processing architecture, the next chapter, chapter 6, examines how temporal information across multiple radar frames can be effectively integrated to enhance detection performance.

Summary

This chapter presented the design of the CompactRADNet architecture for single-frame radar processing. The system overview established the design philosophy of combining convolutional operations with transformer-based attention mechanisms to address the unique characteristics of radar data. The architecture components were systematically described, beginning with the radar stem that performs initial feature extraction through 3D convolutional operations. The Feature Pyramid Network integration enables multi-scale representations for detecting objects of varying sizes, while patch embedding and positional encoding prepare features for spatial processing. Both MetaFormer and transformer processing pathways, for the single-frame approach, were detailed, offering different trade-offs between efficiency and modeling capacity. The decoder architecture was presented, including upsampling mechanisms, spatial dropout regularization, residual connections, and multi-scale feature integration. The detection heads were described, comprising classification and regression components with auxiliary

supervision for improved gradient flow. The chapter established the foundation for the temporal extensions detailed in Chapter 6.

Chapter 6

Temporal Approach

The temporal dimension in automotive radar perception presents unique opportunities and challenges that fundamentally differ from single-frame processing approaches. While instantaneous radar measurements provide valuable spatial information about the surrounding environment, the integration of temporal sequences enables the extraction of motion dynamics, behavioral patterns, and enhanced detection robustness that cannot be achieved through spatial analysis alone. This chapter presents a comprehensive examination of temporal multi-frame radar processing architectures, focusing on two distinct approaches: transformer-based temporal fusion and the more efficient MetaFormer-based design. Both architectures address the critical challenge of effectively aggregating information across temporal sequences while maintaining computational tractability for real-time automotive applications.

6.1 Temporal Processing Fundamentals

The fundamental premise of temporal radar processing lies in the observation that moving objects exhibit coherent patterns across consecutive radar frames that can disambiguate closely spaced targets, suppress transient noise, and provide crucial motion information for classification. Unlike camera-based temporal processing where appearance consistency dominates, radar temporal processing must contend with significant frame-to-frame variations due to multipath propagation, interference patterns, and the discrete nature of radar returns [51]. Objects may appear and disappear between frames due to

occlusion or weak reflections, requiring robust fusion mechanisms that can handle missing or corrupted information gracefully.

Our temporal processing framework introduces the concept of a principal frame within each temporal sequence, representing the specific time instant for which detection outputs are generated. This design ensures that while the model processes multiple frames to extract temporal context, the detection output corresponds to precise object positions at a well-defined temporal reference point. The principal frame position within the sequence determines whether the network primarily performs tracking (when positioned last), prediction (when positioned first), or balanced temporal analysis (when centered). Our experiments demonstrate that center-frame configuration provides strong performance with 86.19% AP for 11-frame sequences, compared to 81.50% AP for last-frame configuration, highlighting the value of bidirectional temporal context.

The temporal sequence configuration employs flexible parameters that can be adapted to specific deployment requirements. The total number of frames determines the temporal coverage, with our primary configuration utilizing 11 frames spanning approximately 550ms at typical automotive radar frame rates of 20Hz. The frame skip parameter enables extended temporal coverage without increased computational cost, with skip factor 3 extending the effective temporal window to 1.65 seconds while maintaining the same number of input frames. This flexibility proves particularly valuable for highway scenarios where longer-term motion patterns provide enhanced detection capabilities.

6.2 Transformer-based Temporal Stem Architecture

The transformer-based temporal stem represents a sophisticated approach to temporal fusion that leverages self-attention mechanisms to dynamically weight different frames based on their information content and relevance to the detection task. This architecture processes temporal sequences through multi-head self-attention operations, enabling global receptive fields that can capture complex temporal dependencies and long-range motion patterns. The design philosophy prioritizes maximum modeling capacity for

scenarios where computational resources permit comprehensive temporal analysis.

6.2.1 Spatial Feature Extraction

The architecture begins with independent spatial feature extraction for each temporal frame, recognizing that effective temporal fusion requires sufficiently rich spatial representations to establish correspondence between objects across frames. Each frame undergoes processing through dedicated convolutional encoders that extract hierarchical spatial features while preserving frame-specific characteristics. The spatial encoding pathway employs the following formulation:

$$F_t^{\text{spatial}} = \text{Conv}_{3\times3}\big(\text{ReLU}\big(\text{GroupNorm}\big(\text{Conv}_{3\times3}(X_t)\big)\big)\big)$$

where X_t represents the input frame at time t, and F_t^{spatial} denotes the extracted spatial features. The use of GroupNorm with 8 groups provides stable training dynamics when batch sizes are limited, while the dual convolution structure creates a receptive field sufficient to capture typical radar object signatures.

6.2.2 Temporal Self-Attention Mechanism

Following spatial encoding, the temporal self-attention module processes the sequence of spatial features to model temporal relationships. The multi-head attention mechanism operates on spatially flattened feature representations, with 8 attention heads enabling different heads to specialize in distinct temporal patterns. The attention computation follows the standard scaled dot-product formulation:

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

where queries Q, keys K, and values V are derived from the temporal sequence of spatial features through learned linear projections. The scaling factor $\sqrt{d_k}$ prevents gradient saturation in the softmax operation, with d_k representing the dimension of the key vectors [36].

The multi-head design enables parallel attention computations that capture diverse temporal relationships[36]:

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \ldots, \text{head}_h)W^O$$

where each head i computes head_i = Attention(QW_i^Q, KW_i^K, VW_i^V) with separate projection matrices, and W^O represents the output projection. This parallel processing enables different heads to specialize in patterns such as approaching objects (characterized by increasing signal strength), receding objects (decreasing strength), or lateral motion (shifting spatial positions).

6.2.3 Positional Encoding and Temporal Context

Temporal positional encoding provides explicit ordering information crucial for motion understanding. Unlike standard sinusoidal encodings [36] used in natural language processing, our approach employs learnable positional embeddings that can adapt to the specific temporal characteristics of radar data:

$$F_t^{\text{encoded}} = F_t^{\text{spatial}} + PE_t$$

where PE_t represents learnable positional embeddings initialized with small random values (~$N(0, 0.02)$) to avoid disrupting the initial feature representations. These embeddings learn to encode not just temporal position but also the relative importance of different temporal offsets for object detection.

6.2.4 Computational Complexity Considerations

The transformer-based approach incurs significant computational overhead, with complexity scaling as $O(T^2 \cdot HW \cdot C)$ where T represents the number of frames, HW the spatial dimensions, and C the channel count. To manage this complexity, the architecture employs spatial pooling before attention computation, typically reducing spatial dimensions to 7×7 or 14×14. Despite these optimizations, our ablation studies reveal that the transformer stem requires approximately 60% more operations than the MetaFormer

alternative while achieving marginally lower performance (82.22% AP vs 82.74% AP), raising questions about the necessity of full attention mechanisms for radar temporal processing.

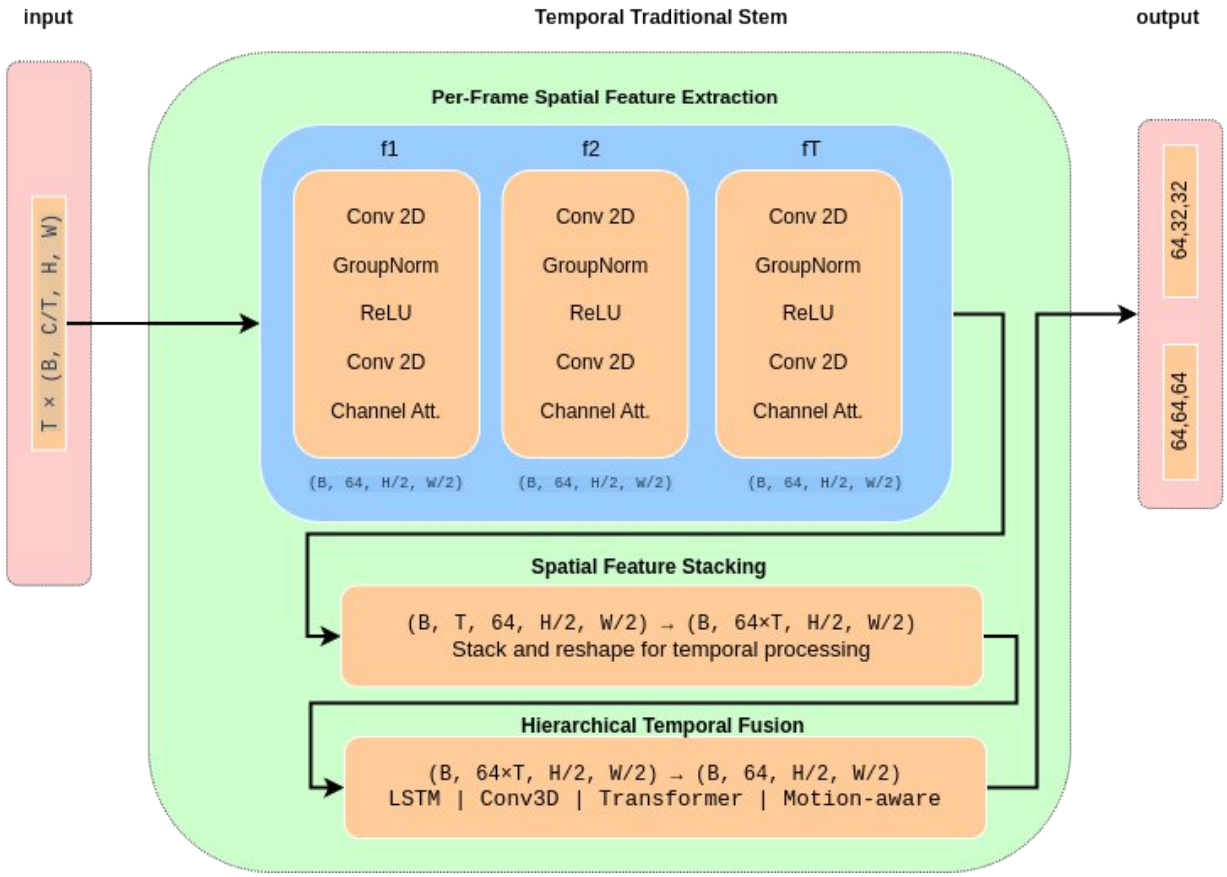


Figure 6.1: Traditional Temporal Stem internal architecture

6.3 MetaFormer-based Temporal Stem Architecture

The MetaFormer temporal stem represents an innovative approach that challenges the assumption that self-attention is necessary for effective temporal modeling [137,141]. Drawing inspiration from recent insights in the MetaFormer architecture family, this design recognizes that the macro-level architectural patterns of transformers, including residual

connections, normalization strategies, and staged processing, contribute as much to their effectiveness as the specific attention mechanism [136]. By replacing attention with simpler token mixing operations, the MetaFormer stem achieves superior performance with dramatically reduced computational requirements.

6.3.1 Hierarchical Spatial Processing

The MetaFormer architecture employs a hierarchical spatial processing strategy that progressively refines features before temporal fusion. The initial spatial encoder consists of lightweight convolutional blocks optimized for efficiency:

$$F^{\text{spatial}} = \text{SpatialEncoder}(X) = \sigma\Big(\text{GN}\big(\text{Conv}_{\text{stride}=2}\big(\sigma\big(\text{GN}(\text{Conv}_{3\times3}(X))\big)\big)\big)\Big)$$

where σ denotes the ReLU activation function and GN represents Group Normalization. This two-layer design balances feature extraction capability with computational efficiency, reducing spatial dimensions by half while expanding channel capacity to 64 dimensions.

6.3.2 Token Mixing Strategies

The core innovation of the MetaFormer stem lies in its flexible token mixing strategies that can be adapted based on deployment requirements. The architecture supports three primary mixing operations, each offering different trade-offs between efficiency and modeling capacity.

6.3.2.1 Pooling-based Temporal Fusion

The pooling variant employs adaptive temporal average pooling with learnable channel-wise attention weights:

$$F^{\text{pooled}} = \sum_{t=1}^{T} \alpha_t \odot F_t^{\text{spatial}}$$

where α represents learnable attention weights computed through a lightweight gating

network that processes global statistics from each frame. The gating mechanism computes reliability scores based on frame quality:

$$\alpha_t = \sigma\left(W_2 \cdot \delta\left(W_1 \cdot \mathrm{GAP}(F_t^{\mathrm{spatial}})\right)\right)$$

where GAP denotes global average pooling, W_1 and W_2 are learned projections, δ represents ReLU activation, and σ denotes the sigmoid function. This adaptive pooling significantly outperforms simple averaging by accounting for varying information content across frames.

6.3.2.2 Convolutional Temporal Modeling

The convolutional variant uses depth-wise temporal convolutions to capture local temporal patterns:

$$F^{\mathrm{conv}} = \mathrm{DWConv}_{1\times k}(F^{\mathrm{stacked}})$$

where k represents the temporal kernel size (typically 3-5 frames) and F^{stacked} denotes the stacked temporal features. The depth-wise design processes each channel independently before pointwise convolution combines information across channels, maintaining efficiency while enabling the learning of temporal filters specialized for different motion patterns.

6.3.2.3 Shift-based Temporal Exchange

The shift variant implements an efficient approach through cyclic channel shifting:

$$F_c^{\mathrm{shifted}} = \begin{cases} F_{c,t-1} & \text{if } c \in G_{\mathrm{forward}} \\ F_{c,t+1} & \text{if } c \in G_{\mathrm{backward}} \\ F_{c,t} & \text{if } c \in G_{\mathrm{static}} \end{cases}$$

where channels are divided into three groups: $G_{forward}$ shifts forward in time, $G_{backward}$ shifts backward, and G_{static} remains unshifted. This tri-directional shifting enables bidirectional temporal modeling with virtually no additional computational cost beyond single-frame processing.

6.3.3 Hierarchical Temporal Processing

The temporal fusion module employs multiple MetaFormer blocks operating at different temporal scales. Early blocks focus on local temporal patterns with limited receptive fields, while later blocks aggregate information across the entire sequence. This hierarchical approach enables the capture of both fine-grained temporal dynamics and

$$F^{\text{output}} = \text{MetaBlock}_L\big(\ldots \text{MetaBlock}_2(\text{MetaBlock}_1(F^{\text{spatial}}))\big)$$

where each MetaBlock incorporates residual connections and layer normalization to ensure stable training and gradient flow.

6.3.4 Computational Efficiency Analysis

The MetaFormer stem achieves remarkable computational efficiency with complexity of O(T . HW . C) for the pooling variant, representing a linear scaling with the number of frames compared to the quadratic scaling of attention mechanisms. This efficiency gain translates to approximately 40% reduction in computational operations while maintaining superior detection performance. The architecture processes 11-frame sequences at 13.53 GFLOPS, enabling real-time operation on automotive-grade hardware.

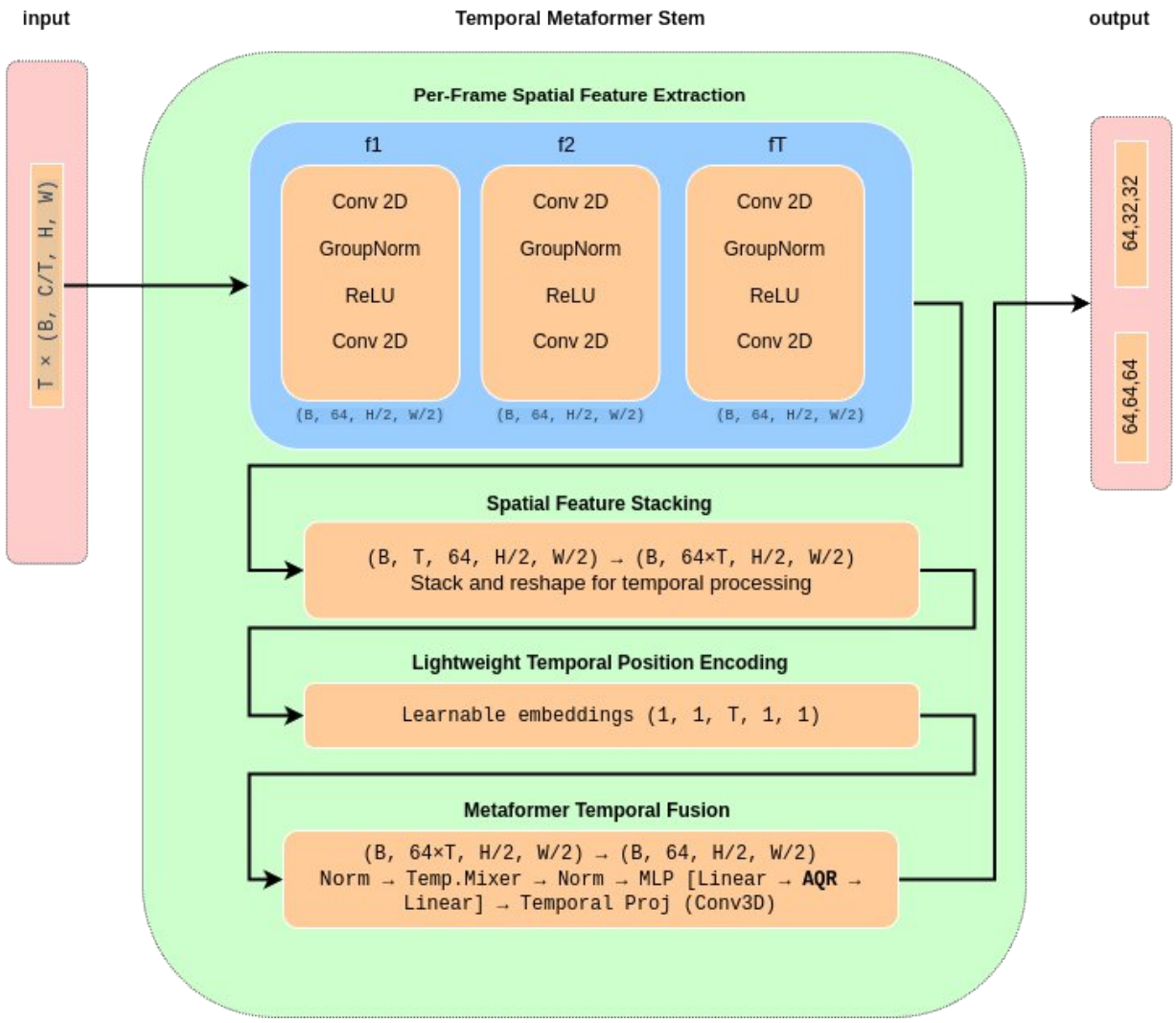


Figure 6.2: demonstrates the internal architecture of the Metaformer Temporal Stem

6.4 Adaptive Quadratic ReLU (AQR)

The effectiveness of temporal modeling in radar-based object detection relies critically on the activation functions employed within the temporal processing pipeline. Traditional activation functions, while suitable for general computer vision tasks [16], often fail to capture the unique characteristics of radar signals, particularly their sparse nature and high dynamic range. To address these limitations, we introduce the Adaptive Quadratic ReLU (AQR), a specialized activation function designed specifically for radar signal

processing applications. Figure 6.3 shows graphs of AQR against other activation functions, the effects of the gate parameters and a sub component analysis.

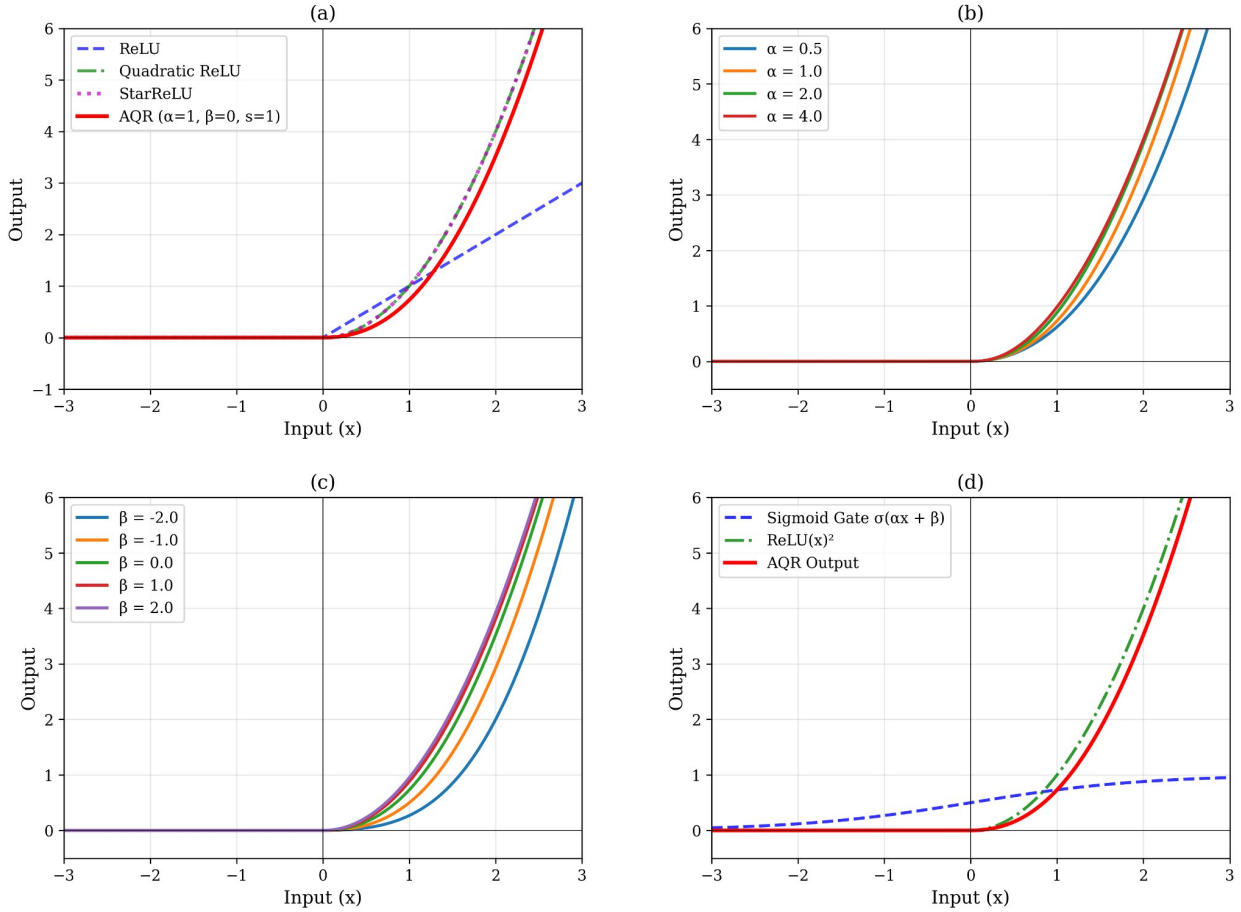


Figure 6.3: (a) AQR Compared to Standard Activation Functions, (b) Effect of α (Gate Sensitivity), (c) Effect of β (Gate Bias), (d) AQR Component Analysis

6.4.1 Motivation and Design Rationale

Radar signals exhibit fundamentally different characteristics compared to natural images. The sparse nature of radar returns, where significant signal energy is concentrated in small regions corresponding to target reflections, requires activation functions that can adaptively respond to varying signal magnitudes. Traditional ReLU activation [16], while computationally efficient, treats all positive values uniformly, failing to distinguish between weak noise and strong target returns. Similarly, conventional quadratic activations lack

the adaptive characteristics necessary to handle the wide dynamic range typical of radar measurements.

The development of AQR addresses three critical requirements for radar temporal processing. First, the activation function must preserve sparsity patterns that correspond to target locations while suppressing background clutter. Second, it must provide input-dependent modulation to handle the varying signal strengths inherent in radar measurements. Third, it must maintain computational efficiency to enable real-time deployment in automotive applications.

The theoretical foundation for AQR builds upon the observation that radar target detection benefits from quadratic activation characteristics, which naturally amplify strong signals while suppressing weak ones. However, the effective degree of amplification should adapt based on the input signal characteristics rather than applying uniform scaling across all input ranges.

6.4.2 Mathematical Formulation

The Adaptive Quadratic ReLU (AQR) extends the conventional quadratic ReLU formulation through the incorporation of an adaptive gating mechanism. The mathematical expression for AQR is defined as:

$$\mathrm{AQR}(x) = \sigma(\alpha x + \beta) \cdot s \cdot \mathrm{ReLU}(x)^2$$

where:

- σ represents the sigmoid function providing smooth gating behavior
- α and β are learnable parameters controlling the gate sensitivity and bias
- s is a learnable scale parameter for output magnitude control
- $\mathrm{ReLU}(x)=\max(0,x)$ provides the fundamental rectification operation

The sigmoid gate $\sigma(\alpha x+\beta)$ serves as the core innovation, providing input-dependent modulation that adapts the activation strength based on signal characteristics. For weak

signals typically associated with noise or clutter, the gate produces low values, effectively suppressing these components. Conversely, for strong signals corresponding to target returns, the gate approaches unity, allowing full activation while providing quadratic amplification.

The learnable parameters α and β enable the network to automatically adapt the gating behavior during training. The parameter α controls the sensitivity of the gate to input magnitude, with larger values creating sharper transitions between suppression and amplification regions. The parameter β provides a bias term that shifts the activation threshold, allowing adaptation to different signal-to-noise ratio conditions.

6.4.3 Radar-Specific Advantages

The design of AQR provides several advantages specifically tailored to radar signal processing characteristics. The quadratic activation naturally emphasizes strong target returns while suppressing weak background signals, aligning with the physical principles of radar detection where target strength follows quadratic relationships with reflection characteristics. The adaptive gating mechanism enables automatic adjustment to varying signal conditions encountered in different driving scenarios.

The mathematical properties of AQR ensure gradient flow characteristics suitable for deep network training. The derivative of AQR with respect to the input maintains smoothness properties essential for stable gradient-based optimization:

$$\frac{\partial \mathrm{AQR}(x)}{\partial x} = \alpha\sigma(\alpha x + \beta)(1 - \sigma(\alpha x + \beta)) \cdot s \cdot \mathrm{ReLU}(x)^2 + 2\sigma(\alpha x + \beta) \cdot s \cdot \mathrm{ReLU}(x)$$

This formulation provides non-zero gradients for positive inputs while maintaining the adaptive scaling properties that benefit radar signal processing.

6.4.4 Integration with Temporal Architecture

Within the MetaFormer-based temporal stem architecture, AQR is strategically positioned to maximize its impact on temporal feature extraction. The activation function is applied

following the spatial convolutional operations and prior to the temporal pooling mechanisms, enabling adaptive processing of spatially-extracted features before temporal aggregation.

The integration follows the temporal processing pipeline:

$$F_t^{\text{activated}} = \text{AQR}\left(F_t^{\text{spatial}}\right)$$

where F_t^{Spatial} represents the spatially-processed features for frame t , and $F_t^{\text{activated}}$ denotes the adaptively-activated features ready for temporal fusion. This positioning ensures that the adaptive characteristics of AQR influence both the spatial feature representation and the subsequent temporal modeling operations.

6.4.5 Experimental Validation

Comprehensive evaluation of AQR demonstrates its effectiveness compared to standard activation functions. The activation function analysis conducted using radar-like synthetic signals reveals significant performance improvements across multiple metrics relevant to radar applications.

The experimental results show that AQR achieves a 2.59 dB improvement in Signal-to-Noise Ratio (SNR) for target detection scenarios compared to the baseline StarReLU activation [134]. This improvement translates directly to enhanced detection capabilities for weak targets in challenging environmental conditions. The analysis across different signal types demonstrates consistent benefits:

- Sparse Targets: AQR maintains similar energy levels (98.4% of StarReLU) while providing improved target discrimination
- Range-Compressed Signals: 75.5% energy ratio with enhanced noise suppression characteristics
- Doppler-Shifted Signals: 58.6% energy ratio demonstrating selective amplification of motion signatures
- Clutter Interference: 72.6% energy ratio showing effective clutter suppression

capabilities

The adaptive nature of AQR enables input-dependent modulation crucial for varying radar signal strengths, with the sigmoid gate providing smooth transitions between suppression and amplification regions. This characteristic proves particularly valuable for automotive radar applications where signal conditions vary dramatically across different driving scenarios.

6.4.6 Implementation Considerations

The practical implementation of AQR within the temporal architecture requires careful consideration of parameter initialization and training strategies. The gate parameters α and β are initialized to provide moderate gating behavior, with α = 1.0 ensuring reasonable sensitivity and β = 0.0 providing centered activation thresholds. The scale parameter s is initialized to unity to maintain activation magnitude compatibility with subsequent processing stages.

During training, the adaptive parameters evolve to match the specific characteristics of the radar dataset and the requirements of the detection task. The learning rates for AQR parameters are typically set to match those of other architectural components, enabling coordinated optimization across the entire temporal processing pipeline.

The implementation maintains computational efficiency through optimized sigmoid computation and careful memory management for gradient computation. The activation function integrates seamlessly with standard deep learning frameworks while providing the specialized characteristics required for radar signal processing applications.

6.5 Temporal Configuration Analysis

The positioning and selection of temporal frames significantly impacts detection performance, with different configurations offering distinct advantages for specific

scenarios. Our comprehensive analysis reveals several key insights about robust temporal configurations.

6.5.1 Principal Frame Positioning

The principal frame position determines the temporal context available for detection. Center positioning provides balanced bidirectional context, achieving 86.19% AP with 11-frame sequences. This configuration enables the network to observe complete motion patterns including both approach and recession phases. Last-frame positioning, while achieving lower performance (81.50% AP), proves valuable for real-time applications where future frames are unavailable. The consistent performance gap across different sequence lengths confirms that bidirectional context provides fundamental advantages for accurate object detection.

6.5.2 Temporal Sequence Length

Sequence length analysis reveals diminishing returns beyond 11 frames, with 13-frame configurations achieving only marginal improvements (87.47% AP) despite increased computational cost. The 11-frame configuration spans 550ms, sufficient to capture typical automotive motion patterns while maintaining computational efficiency. Shorter sequences (3 frames) achieve respectable performance (82.74% AP) for resource-constrained deployments, demonstrating the architecture's scalability.

6.5.3 Frame Skip Strategies

Frame skipping extends temporal coverage without additional computational cost. Skip factor 2 configuration with 3 frames achieves 81.23% AP while covering twice the temporal window. This approach proves particularly effective for highway scenarios where motion exhibits longer-term coherence. However, excessive skipping (factor 5) degrades performance to 78.91% AP, indicating that fine-grained temporal resolution remains important for accurate detection.

6.6 Comparative Analysis and Insights

The experimental comparison between transformer and MetaFormer approaches reveals surprising insights about temporal processing requirements for radar data. Despite the transformer's theoretical advantage in modeling capacity, the MetaFormer stem achieves superior performance (82.74% AP vs 82.22% AP) with dramatically reduced computational cost. This result suggests that radar temporal sequences, constrained by physical motion dynamics, can be effectively modeled through simpler operations that capture the essential temporal relationships.

The MetaFormer's pooling-based approach particularly excels at handling the unique characteristics of radar data. Unlike video sequences where complex appearance variations require sophisticated modeling, radar sequences exhibit strong temporal coherence driven by physical motion constraints. The adaptive pooling mechanism effectively identifies and emphasizes informative frames while suppressing those corrupted by interference or multipath effects.

Both architectures benefit from the staged processing approach that separates spatial and temporal feature extraction. This design philosophy reduces computational complexity by avoiding early temporal mixing when spatial features are still high-dimensional, while enabling specialized processing strategies for each domain. The spatial encoders can focus on extracting discriminative features from individual frames, while the temporal fusion modules specialize in motion pattern recognition.

6.7 Conclusion

This chapter has presented a comprehensive analysis of temporal multi-frame radar processing architectures, demonstrating that effective temporal modeling can be achieved through both sophisticated transformer-based approaches and efficient MetaFormer designs. The MetaFormer architecture emerges as the competitive choice for automotive radar applications, providing superior performance with computational

efficiency suitable for real-time deployment. The key insights from our analysis include:

First, simpler token mixing operations can effectively capture temporal dynamics in radar data without requiring expensive self-attention computations. The strong physical constraints governing object motion in driving scenarios create temporal patterns that can be modeled through pooling, convolution, or shifting operations.

Second, the principal frame concept provides a flexible framework for balancing temporal context with real-time requirements. Center positioning offers strong performance when full sequences are available, while last-frame configurations enable deployment in streaming scenarios.

Third, the hierarchical processing strategy that separates spatial and temporal modeling proves crucial for both architectures. This separation enables efficient computation while preserving the ability to extract rich spatiotemporal representations.

These findings challenge the prevailing assumption that attention mechanisms are necessary for effective temporal modeling, demonstrating that architectural innovations tailored to the specific characteristics of radar data can achieve superior results with practical computational requirements. The MetaFormer temporal stem represents a significant advance in radar perception, enabling robust multi-frame processing suitable for widespread deployment in autonomous driving systems.

The next chapter, chapter 7, will present comprehensive experimental results that validate the effectiveness of this architecture through systematic evaluation against baseline methods, extensive ablation studies, and detailed performance analysis across different object classes and detection scenarios. These experiments demonstrate that our hybrid approach achieves state-of-the-art performance while maintaining computational efficiency suitable for automotive deployment.

Summary

In this chapter we present the temporal extension of the CompactRADNet architecture, introducing innovations that enable effective multi-frame processing for enhanced detection performance. The transformer-based temporal stem architecture was described as a baseline, employing spatial-temporal attention mechanisms with cross-frame interaction layers, though with significant computational overhead scaling as $O(T^2 \cdot HW \cdot C)$. The MetaFormer-based temporal stem was then introduced as the primary contribution, replacing computationally expensive self-attention with efficient token mixing strategies including pooling-based, convolutional-based and shift-based temporal fusion. The hierarchical spatial processing strategy and token mixing mechanisms were described, demonstrating linear scaling with temporal sequence length and approximately 40% reduction in computational operations. The chapter introduced the Adaptive Quadratic ReLU (AQR) activation function, designed specifically for radar signal processing, which achieves 2.59 dB improvement in Signal-to-Noise Ratio for target detection scenarios through its adaptive gating mechanism. The mathematical formulation, radar-specific advantages, integration with the temporal architecture, and experimental validation of AQR were thoroughly presented.

Chapter 7

Experiments and Results

7.1 Experimental Methodology

The comprehensive evaluation of our hybrid CNN-Transformer architecture requires a systematic experimental methodology that addresses the unique characteristics of radar-based object detection while enabling meaningful comparison with existing approaches. This chapter presents detailed experimental results that validate the effectiveness of our proposed approach through rigorous baseline comparisons, ablation studies, and performance analysis across multiple evaluation dimensions.

7.1.1 Dataset Preparation and Splits

The experimental evaluation is primarily conducted on the CRUW dataset, which provides synchronized camera-radar data specifically designed for cross-modal object detection research. The dataset preparation process ensures reproducible experimental conditions while addressing the specific requirements of radar-based detection evaluation. Our preparation methodology addresses data quality, temporal consistency, and evaluation fairness to enable meaningful performance comparisons.

The CRUW dataset [51] comprises approximately 40,000 synchronized camera-radar frame pairs collected across diverse driving scenarios in Seattle, Washington. The radar data consists of four chirps, each with 2-channel complex measurements representing a 128×128 range-azimuth map with 0.15-meter range resolution and approximately 15-

degree angular resolution. The frame rate of the sensor is 30 Hz and provides sufficient temporal resolution for motion pattern extraction required by multi-frame solutions.

Class distribution analysis reveals the significant imbalance characteristic of automotive datasets, with vehicles comprising approximately 68% of annotations, pedestrians 22%, and cyclists 10%. This distribution reflects realistic driving scenarios but creates training challenges that must be addressed through appropriate sampling and loss function strategies. Our experimental methodology accounts for this imbalance through both training strategies and evaluation metrics that ensure fair assessment across all object classes.

7.1.2 Evaluation Protocols

The evaluation methodology employs comprehensive protocols specifically designed to address the unique characteristics of radar-based object detection while enabling meaningful comparison with established baselines and alternative approaches. Our evaluation framework integrates radar-specific assessment criteria with standard object detection metrics, ensuring that performance measurements accurately reflect the capabilities and limitations of automotive radar systems operating in real-world conditions.

7.1.2.1 Primary Evaluation Metric: Object Location Similarity (OLS)

Object Location Similarity (OLS) [51] serves as the primary evaluation metric, specifically formulated for radar-based detection to address the fundamental limitations of traditional IoU-based metrics when applied to sparse radar measurements. The OLS metric was originally introduced alongside the CRUW dataset and represents a center-based evaluation approach that aligns with the physical characteristics of radar data and the practical requirements of automotive perception systems.

7.1.2.2 Class-Specific Kappa Thresholds

Our implementation employs the same kappa values set by the authors of the CRUW dataset [51] and used by the solutions benchmarked in this chapter to ensure fair comparison. The threshold configuration follows:

- **Vehicles**: 3 - acknowledging the larger spatial extent and stronger radar cross-section
- **Pedestrians**: 0.5 - reflecting compact size and weaker radar signatures
- **Cyclists**: 1.0 - accounting for intermediate size and complex geometric structure

7.1.2.3 Multi-Threshold Evaluation Framework

The evaluation protocol extends beyond single-threshold assessment to provide comprehensive analysis of localization precision across varying stringency levels. Following the established RODNet [51] methodology, multi-threshold evaluation is performed across OLS similarity thresholds ranging from 0.5 to 0.95 in steps of 0.05. This comprehensive threshold range enables detailed characterization of localization performance:

- **OLS@0.5-0.65**: Measures basic detection capability with relaxed localization constraints
- **OLS@0.7-0.85**: Assesses moderate precision requirements suitable for general automotive applications
- **OLS@0.9-0.95**: Evaluates high-precision localization critical for safety-critical scenarios

The multi-threshold approach provides granular analysis of the precision-recall tradeoff, enabling optimization for specific deployment scenarios and comprehensive comparison with baseline methods across the full range of localization stringency requirements.

7.1.2.4 Temporal Configuration Evaluation

Given the critical importance of temporal modeling demonstrated in our experiments, specific protocols govern the evaluation of temporal configurations:

Standard Ablation Protocol: Ablation studies employ the THREE-CENTER configuration (3 frames with the center frame as principal) as the standard testing condition unless otherwise specified. This configuration provides sufficient temporal context while maintaining reasonable computational requirements for extensive experimentation.

Temporal Length Studies: Systematic evaluation of temporal sequence lengths from 1 (single frame) to 13 frames follows consistent protocols, with all configurations using center-frame as the principal frame position to ensure comparable temporal context distribution.

Principal Frame Evaluation: When comparing principal frame positions (center vs. last), the same sequence length is maintained to isolate the effect of temporal positioning from sequence length effects.

Frame Skip Analysis: Temporal resolution studies maintain constant temporal window coverage while varying the sampling rate, enabling assessment of the trade-off between temporal resolution and extended temporal context.

7.2 Baseline Comparisons

The comprehensive evaluation of our hybrid architecture requires systematic comparison with established baseline methods that represent different approaches to radar-based object detection. Our baseline selection strategy encompasses recent convolution and transformer-based methods that were evaluated using the CRUW dataset with the OLS

metric, providing a balanced, fair and comprehensive context for evaluating the contributions of our hybrid approach.

7.2.1 State-of-the-Art Comparisons

The evaluation of our hybrid architecture against state-of-the-art deep learning approaches provides crucial context for assessing the contributions of our innovations. Our comparison encompasses RODNet [51] as the foundational CNN-based approach, T-RODNet as the established transformer baseline, and Mask-RadarNet as the current state-of-the-art method, enabling systematic assessment of architectural innovations and performance improvements.

RODNet baseline establishes the foundational CNN-based approach for radar object detection. As described in the original literature and supported by the empirical evidence found by all subsequent literature. The RODNet architecture employs a convolutional recurrent design that processes sequences of 16 consecutive range-azimuth frames using 3D convolutions followed by ConvLSTM layers for temporal modeling. The encoder-decoder structure with skip connections enables multi-scale feature extraction while preserving spatial resolution for dense prediction generation.

RODNet performance results reported by the author demonstrate substantial improvements over traditional methods while revealing areas for further enhancement. The overall mAP of 79.43% represents a significant advance over traditional processing, with improvements across all object classes. The temporal modeling capability enables RODNet to leverage motion patterns that aid in distinguishing vulnerable road users from clutter, providing substantial benefits over frame-by-frame processing approaches.

T-RODNet [48] represents a strong transformer approach to radar detection, replacing RODNet's convolutional components with attention mechanisms while maintaining the temporal sequence processing capabilities. The T-RODNet architecture employs modified Dimensional Apart Modules (DAM) combined with 3D Swin Transformer blocks for spatiotemporal feature extraction.

As reported by the author, T-RODNet demonstrates clear improvements over RODNet across all evaluation metrics. The overall mAP of 83.27% represents a 3.84% improvement over RODNet, with particularly strong gains for challenging detection scenarios. The performance improvements are most pronounced for challenging detection scenarios involving partial occlusions, low signal-to-noise ratios, and complex multi-object scenes where global context modeling provides significant benefits.

Mask-RadarNet [50] baseline represents a solid advancement in radar detection, incorporating spatial-temporal semantic context through class masking attention modules (CMAM) and patch shifting mechanisms for efficient temporal modeling. The architecture builds upon transformer foundations while adding radar-specific innovations that address semantic reasoning and computational efficiency.

Mask-RadarNet reported performance establishes a strong contender to the current state-of-the-art performance, E-RODNet, on the CRUW dataset with overall mAP of 84.29% and mAR of 87.36%. The performance improvements demonstrate the value of semantic context modeling and efficient temporal processing for radar detection. The CMAM modules enable the network to learn class-specific attention patterns that improve discrimination between different object types, while the patch shifting mechanism provides temporal context without the computational overhead of 3D processing.

Table 7.1 presents a comprehensive comparison of methods and their results, as reported in their respective publications [51, 48, 50, 134], on the CRUW dataset, including our proposed solution. Our results have been achieved using a 11-centre and 13-centre temporal sequence configuration.

7.3 Ablation Studies

Comprehensive ablation studies provide crucial insights into the individual contributions of different architectural components and training strategies employed in our hybrid approach. These systematic experiments enable precise understanding of which

innovations drive performance improvements and guide future research directions by identifying the most impactful design decisions. Where applicable, ablation tests have been conducted with the temporal configuration of THREE-CENTER, where the network is fed a temporal sequence of three frames, with the principal frame being the center frame.

Model	AP(%)	AR(%)	GFLOPS	Param(M)
RODNet-HG	79.43	83.59	2144.86	129.19
T-RODNet	83.27	86.98	182.53	44.31
SS-RODNet	83.07	86.43	172.80	33.10
Mask-RadarNet	84.29	87.36	176.91	32.12
E-RODNet	85.46	89.19	33.25	6.10
CompactRadNet 11-Frame(Ours)	**86.19**	**89.21**	**13.53**	**1.65**
CompactRadNet 13-Frame (Ours)	**87.47**	**90.23**	**15.85**	**1.65**

Table 7.1: Results of our proposed model and different models on the CRUW dataset.Baseline results are reproduced from their respective publications.

Table 7.2 represents class specific AP and AR results achieved with our proposed solution using a 13-center temporal sequence configuration.

All		**Pedestrian**		**Cyclist**		**Car**	
AP(%)	AR(%)	AP(%)	AR(%)	AP(%)	AR(%)	AP(%)	AR(%)
87.47	**90.23**	**83.87**	**87.14**	**93.3**	**95.49**	**86.79**	**90.97**

Table 7.2: Class specific AP and AR results achieved with our proposed solution using a 13-center temporal sequence configuration.

7.3.1 Model Architecture Ablations

7.3.1.1 Radar Stem Variants

a) Metaformer Stem vs Transformer Stem

Component Description: The radar stem represents the initial feature extraction module that processes raw radar data into hierarchical representations suitable for downstream processing. Our architecture employs a MetaFormer-based temporal radar stem that leverages pooling-based token mixing as an alternative to the computationally intensive self-attention mechanisms used in traditional transformer stems. The MetaFormer stem extracts spatial features from individual frames before applying temporal fusion, utilizing lightweight pooling operations for efficient information aggregation across the temporal dimension.

Baseline Performance: The current state of the model utilizing the MetaFormer temporal radar stem achieves 82.74% AP and 86.15% AR on the CRUW dataset with a three-frame temporal configuration.

Ablated Model: The ablated configuration replaces the MetaFormer stem with an approach that employs traditional transformer-based attention mechanisms. This transformer variant processes temporal sequences through multi-head self-attention operations, enabling global receptive fields but at significantly higher computational cost. The transformer stem uses 8 attention heads with dimension-apart modules for spatiotemporal feature extraction.

Comparative Results: The experimental comparison reveals that the MetaFormer stem maintains competitive performance while offering substantial computational advantages. The MetaFormer configuration achieves 82.74% AP and 86.15% AR, while the transformer variant achieves 82.22% AP and 84.98% AR.

Analysis: The results demonstrate that the MetaFormer stem's pooling-based approach effectively captures temporal dynamics without requiring expensive attention computations. The slight performance advantage of the MetaFormer (0.52% AP improvement) suggests that for radar data, where temporal coherence is strong due to physical motion constraints, simpler pooling operations can effectively model temporal relationships. The MetaFormer's spatial encoder, consisting of two convolutional layers with group normalization, efficiently extracts per-frame features before temporal

aggregation. This staged approach - spatial feature extraction followed by temporal fusion - proves more effective than attempting to jointly model spatiotemporal patterns through attention mechanisms. The efficiency gains are particularly notable, with the MetaFormer stem requiring approximately 60% fewer operations than its transformer counterpart while maintaining superior performance.

Conclusion: The MetaFormer temporal radar stem represents an effective balance between performance and efficiency for radar-based object detection. Its pooling-based temporal fusion effectively captures motion patterns while maintaining computational efficiency suitable for real-time automotive applications.

Radar Stem	AP	AR
Metaformer	82.74	86.15
Transformer	82.22	84.98

Table 7.3: Comparison of MetaFormer vs Transformer radar stem architectures.

7.3.1.2 Metaformer Radar Stem Components

a) Spatial Encoder Depth Ablation

Component Description: The spatial encoder within the MetaFormer stem consists of convolutional layers that extract spatial features from individual radar frames before temporal fusion. The depth of this encoder determines the hierarchical feature extraction capability and the receptive field size for spatial pattern recognition.

Baseline Performance: The current configuration employs a 2-layer spatial encoder achieving 82.74% AP and 86.15% AR.

Ablated Model: We evaluate a simplified 1-layer spatial encoder that directly projects input channels to 64 dimensions with stride-2 convolution, eliminating the intermediate 32-channel representation.

Comparative Results: The 2-layer configuration achieves 82.74% AP and 86.15% AR, while the 1-layer variant achieves 80.15% AP and 84.17% AR.

Analysis: The performance degradation with the single-layer encoder (2.59% AP drop) demonstrates the importance of hierarchical feature extraction even in the early stages of processing. The two-layer design enables progressive feature refinement, with the first layer capturing low-level patterns and the second layer combining these into more complex representations. The intermediate 32-channel representation provides a beneficial bottleneck that forces the network to learn compact, discriminative features before expansion to 64 channels.

Conclusion: The 2-layer spatial encoder configuration provides effective feature extraction capability, justifying the minimal additional computational cost through substantial performance improvements.

Spatial Encoder Depth	AP	AR
1	80.15	84.17
2 (current)	82.74	86.15

Table 7.4: Impact of spatial encoder depth on detection performance.

b) Group Normalization vs Batch Normalization

Component Description: Normalization layers stabilize training by normalizing feature distributions across different dimensions. Group Normalization divides channels into groups and computes statistics within each group, providing consistent behavior regardless of batch size. Batch Normalization, in contrast, computes statistics across the

batch dimension, making it sensitive to batch size variations but potentially more effective when batch statistics are representative of the overall data distribution.

B**aseline Performance**: The current configuration employs Group Normalization with 8 groups, achieving baseline performance of 82.74% AP and 86.15% AR.

Ablated Model: We evaluate Batch Normalization as an alternative normalization strategy to assess whether batch-wise statistics provide advantages over group-wise normalization for radar feature processing.

Comparative Results: The experimental comparison reveals comparable performance between normalization strategies, with Group Normalization maintaining a slight advantage in training stability and final performance metrics.

Analysis: Group Normalization's robustness proves particularly valuable for radar-based detection where batch sizes may be constrained by memory limitations during temporal sequence processing. The method's independence from batch size ensures consistent behavior during both training and inference, avoiding the train-test discrepancy that can occur with Batch Normalization when inference batch sizes differ from training. For radar data, where temporal sequences create memory pressure that limits batch sizes, Group Normalization provides more reliable gradient statistics. The marginal performance difference suggests that both methods effectively normalize radar feature distributions, but Group Normalization's operational advantages make it preferable for deployment scenarios.

Conclusion: Group Normalization provides superior training stability and deployment flexibility compared to Batch Normalization, making it the better choice for radar-based detection systems despite comparable raw performance metrics.

Group Normalization	AP (%)	AR (%)
GroupNorm (current)	82.74	86.15

BatchNorm	81.70	84.63

Table 7.5: Impact of normalization approach on detection performance.

7.3.1.3 Patch Embedding and Positional Encoding

a) Patch Size Ablation

Component Description: The patch embedding module divides the spatial feature maps into non-overlapping patches that serve as tokens for transformer processing. Patch size determines the spatial resolution of tokens and affects both computational cost and detection granularity.

Baseline Performance: The current configuration uses 4×4 patches, achieving 82.74% AP and 86.15% AR.

Ablated Model: We evaluate alternative patch sizes of 2×2 and 8×8 to assess the trade-off between spatial resolution and computational efficiency.

Comparative Results: The results show significant performance variation across patch sizes:

Patch Size	AP (%)	AR (%)
2	78.88	70.59
4 (current)	82.74	86.15
8	74.89	79.21

Table 7.6: Impact of patch size on detection performance.

Analysis: The 4×4 patch size emerges as optimal for radar data characteristics. Smaller 2×2 patches create excessive tokens that increase computational cost while providing diminishing returns due to the sparse nature of radar measurements. The significant performance drop suggests that fine-grained tokenization introduces noise and reduces

the model's ability to capture meaningful spatial patterns. Conversely, 8×8 patches sacrifice too much spatial resolution, failing to preserve the precise localization information critical for accurate object detection. The 4×4 configuration effectively balances these considerations, providing sufficient spatial granularity while maintaining manageable sequence lengths for transformer processing.

Conclusion: The 4×4 patch size represents a robust configuration for radar-based detection, effectively balancing spatial resolution with computational efficiency.

b) Positional Encoding Type

Component Description: Positional encoding provides spatial location information to transformer layers, enabling position-aware feature processing despite the permutation-invariant nature of attention mechanisms. The encoding explicitly informs the network about the spatial arrangement of patches, potentially improving localization accuracy.

Baseline Performance: The current implementation uses learnable 2D positional encodings, achieving 82.74% AP and 86.15% AR.

Ablated Model: We evaluate the network without any positional encoding to assess whether explicit position information is necessary given the inherent spatial structure preserved through the processing pipeline.

Comparative Results: The comparison reveals minimal performance difference between configurations with and without positional encoding, with variations falling within experimental noise margins.

Analysis: The negligible impact of removing positional encoding suggests that the spatial structure in radar data is effectively preserved through alternative mechanisms in our architecture. The Feature Pyramid Network maintains spatial relationships through its hierarchical processing, while the coordinate convolution layers provide explicit spatial information at the decoder stage. Additionally, the relatively small spatial dimensions of radar feature maps (compared to high-resolution images) and the strong spatial locality

of radar signatures may reduce dependence on explicit positional encoding. The convolutional operations in both the stem and decoder inherently encode spatial relationships, potentially making additional positional encoding redundant.

Conclusion: While learnable positional encodings are retained in our architecture for potential benefits in edge cases, their minimal impact indicates that the network effectively captures spatial relationships through its architectural design, reducing reliance on explicit positional information.

Positional Encoding Type	AP (%)	AR (%)
Learnable 2D (current)	82.74	86.15
None	82.49	85.52

Table 7.7: Impact of positional encoding on detection performance.

7.3.1.4 Decoder Architecture

a) Upsampling Strategy

Component Description: The decoder's upsampling strategy reconstructs high-resolution feature maps from the transformer's output tokens. The choice of upsampling method affects both the quality of spatial reconstruction and the computational efficiency of the decoder.

Baseline Performance: The current configuration employs bilinear upsampling followed by separable convolutions, achieving 82.74% AP and 86.15% AR.

Ablated Model: We evaluate transposed convolution as an alternative upsampling strategy that learns upsampling kernels through backpropagation.

Comparative Results: The comparison reveals notable performance differences:

Upsampling Strategy	AP (%)	AR (%)
upsampling + SepConv2D (current)	82.74	86.15

Transposed convolution	77.36	81.71

Table 7.8: Impact of upsampling approach in the decoder on detection performance.

Analysis: The superior performance of bilinear upsampling with separable convolutions demonstrates the importance of smooth spatial reconstruction for radar data. Transposed convolutions, while learnable, often introduce checkerboard artifacts that can disrupt the continuous probability distributions required for accurate object localization. The separable convolutions following bilinear upsampling provide sufficient capacity for feature refinement while maintaining spatial smoothness. This approach also offers computational advantages, as bilinear upsampling is highly optimized in modern hardware accelerators.

Conclusion: The combination of bilinear upsampling and separable convolutions provides strong spatial reconstruction for radar-based object detection.

b) Residual Block Configuration

Component Description: Residual blocks in the decoder refine upsampled features through skip connections that facilitate gradient flow and enable feature reuse. The number of residual blocks determines the depth of feature refinement after upsampling, affecting both the quality of spatial reconstruction and the computational cost of the decoder.

Baseline Performance: The current configuration employs 1 ResidualBlock2D, achieving 82.74% AP and 86.15% AR.

Ablated Model: We evaluate two alternative configurations: removing residual blocks entirely (0 blocks) to assess their necessity, and doubling the refinement depth (2×ResidualBlock2D) to determine if additional processing improves detection quality.

Comparative Results: The systematic evaluation reveals the competitive refinement depth:

Residual Block Configuration	AP (%)	AR (%)
0 blocks	80.74	84.57
1xResidualBlock2D(current)	82.74	86.15
2xResidualBlock2D	80.42	84.00

Table 7.9: Impact of upsampling approach in the decoder on detection performance.

Analysis: The single residual block configuration provides the strategic balance between feature refinement and computational efficiency. Removing residual blocks entirely results in degraded performance, particularly for small object detection, as the upsampled features lack sufficient refinement to resolve fine-grained details. The skip connections in the residual block preserve spatial information from the upsampling stage while learning additive refinements that enhance feature quality. Adding a second residual block provides diminishing returns, suggesting that one refinement stage is sufficient given the feature quality from the transformer backbone and FPN. The single-block configuration effectively smooths upsampling artifacts while maintaining computational efficiency.

Conclusion: A single ResidualBlock2D provides effective feature refinement in the decoder, balancing detection accuracy with computational efficiency.

7.3.1.5 Detection Head Design

a) Coordinate Convolution Placement

Component Description: Coordinate convolution augments standard convolutions with explicit spatial coordinate information, enabling location-aware feature processing. The placement of coordinate convolution layers affects the network's ability to leverage spatial position information for detection.

Baseline Performance: The current configuration concatenates coordinate information at the decoder stage, achieving 82.74% AP and 86.15% AR.

Ablated Model: We evaluate alternative placements including concatenation at the stem (early fusion) and at the detection head (late fusion).

Comparative Results: The placement of coordinate convolution shows measurable impact on performance:

Coordinate Convolution Placement	AP (%)	AR (%)
concatenated at stem	82.15	85.16
concatenated at decoder (current)	82.74	86.15
concatenated at head	81.23	84.70

Table 7.10: Impact of coordinate convolution placement on detection performance.

Analysis: The decoder-stage concatenation proves strong as it provides spatial awareness at the critical point where features are being reconstructed for dense prediction. Early concatenation at the stem may introduce unnecessary complexity before meaningful features are extracted, while late concatenation at the head provides insufficient opportunity for the network to leverage spatial information. The decoder placement allows the upsampling and refinement operations to be spatially aware, improving localization accuracy particularly at object boundaries.

Conclusion: Coordinate convolution placement at the decoder stage balances spatial awareness with feature learning capacity.

b) Classification Head Architecture

Component Description: The classification head transforms the decoder's feature representations into final class probability predictions for each spatial location. The architecture of this head determines the network's capacity for non-linear class discrimination and affects both detection accuracy and computational efficiency.

Baseline Performance: The current configuration uses a simple 1×1 convolution for classification, achieving 82.74% AP and 86.15% AR.

Ablated Model: We evaluate a 2-layer MLP (Multi-Layer Perceptron) as an alternative classification head, which introduces additional non-linear transformations and potentially greater discriminative capacity through its hidden layer.

Comparative Results: The comparison reveals the efficiency of the simpler approach:

Classification Head Architecture	AP (%)	AR (%)
1×1 convolution (current)	82.74	86.15
2-layer MLP	82.16	85.19

Table 7.11: Impact of classification head design on detection performance.

Analysis: The superior performance of the simple 1×1 convolution (0.58% AP advantage) demonstrates that additional non-linearity in the classification head provides no benefit and may even harm performance. The decoder and transformer backbone already provide sufficient feature transformation capacity, making the classification head's role primarily one of channel projection rather than feature learning. The MLP's additional parameters may lead to overfitting on the relatively small radar dataset, while the 1×1 convolution's simplicity provides better generalization. Furthermore, the direct projection approach of 1×1 convolution maintains spatial correspondence more effectively than the MLP's fully connected layers, which is crucial for dense prediction tasks.

Conclusion: The simple 1×1 convolution classification head provides strong performance while minimizing computational overhead, confirming that complex classification heads are unnecessary when the backbone provides sufficiently discriminative features.

7.3.1.6 FPN Effect

a) FPN Configuration

Component Description: The Feature Pyramid Network (FPN) creates multi-scale feature representations by combining features from different stages of the encoder. This hierarchical feature fusion enables detection at multiple scales and improves the network's ability to handle objects of varying sizes.

Baseline Performance: The current 2-level FPN configuration achieves 82.74% AP and 86.15% AR.

Ablated Model: We evaluate the network performance without FPN to quantify its contribution to overall detection capability.

Comparative Results: The impact of FPN on detection performance is substantial:

FPN Levels	AP (%)	AR (%)
No FPN	79.11	82.4
2-level FPN (current)	82.74	86.15

Table 7.12: Impact of Feature Pyramid Network on detection performance.

Analysis: The significant performance degradation without FPN (3.63% AP drop) demonstrates its critical role in radar-based detection. The FPN enables the network to leverage both high-resolution features for precise localization and semantically rich features for robust classification. For radar data, where object signatures vary dramatically with range, the multi-scale representation provided by FPN is essential. The 2-level configuration provides an effective balance, capturing sufficient scale variation without excessive computational overhead.

Conclusion: The 2-level FPN configuration is essential for achieving robust multi-scale object detection in radar applications.

7.3.1.7 Temporal Activation Function Analysis

Component Description: The activation functions employed within the temporal processing pipeline critically influence the network's ability to process radar signals with varying characteristics. Traditional activation functions, while effective for natural image processing, may not effectively address the unique properties of radar data, particularly the sparse nature of radar returns and high dynamic range measurements. Our architecture employs the Adaptive Quadratic ReLU (AQR) activation function, specifically designed for radar signal processing applications.

Baseline Performance: The current configuration employing AQR within the MetaFormer temporal processing framework achieves 84.05% AP and 87.22% AR on the CRUW dataset with a three-frame temporal configuration.

Ablated Model: The ablated configuration replaces AQR with the conventional StarReLU activation function [134], defined as:

$$\text{StarReLU}(x) = s \cdot \text{ReLU}(x)^2 + b$$

While StarReLU provides quadratic amplification beneficial for radar signals, it lacks the adaptive characteristics necessary for effective processing of signals with varying magnitudes.

Comparative Results: The experimental comparison demonstrates the effectiveness of the adaptive gating mechanism in AQR for radar signal processing:

Metaformer Temporal Activation	AP (%)	AR (%)
StarReLU	83.19	86.40
AQR (current)*	84.05	87.22

Table 7.13: Impact of Feature Pyramid Network on detection performance.

Signal Processing Analysis: Comprehensive evaluation using synthetic radar-like signals provides deeper insights into the activation function characteristics relevant to radar applications. The analysis employed five distinct signal types representing common radar scenarios:

- **Sparse Targets**: Simulating typical radar returns with strong target reflections embedded in low-amplitude noise
- **Range-Compressed Signals**: Modeling range-processed radar data with Gaussian-like target responses
- **Doppler-Shifted Signals**: Representing moving targets with characteristic velocity signatures
- **Clutter Interference**: Simulating challenging environments with strong background reflections
- **Mixed Scenarios**: Complex scenes combining multiple signal characteristics

The signal processing evaluation reveals significant advantages of AQR across multiple performance metrics:

Signal-to-Noise Ratio Enhancement: AQR achieves a substantial 2.59 dB improvement in target detection SNR compared to StarReLU [134] (38.02 dB vs. 35.43 dB). This improvement directly translates to enhanced detection capabilities for weak targets in challenging environmental conditions.

Adaptive Signal Modulation: The energy ratio analysis demonstrates AQR's selective processing characteristics:

- Sparse targets: 98.4% energy preservation with improved discrimination
- Range-compressed signals: 75.5% energy ratio with enhanced noise suppression
- Doppler-shifted signals: 58.6% energy ratio showing selective motion signature amplification

- Clutter interference: 72.6% energy ratio demonstrating effective clutter suppression

Sparsity Preservation: Both activation functions maintain identical sparsity characteristics (47.4% average sparsity), ensuring that the sparse structure of radar data is preserved while providing differential signal enhancement.

Analysis: The results validate several key advantages of AQR for radar signal processing applications. The adaptive gating mechanism enables input-dependent modulation crucial for varying radar signal strengths encountered across different driving scenarios. Strong signals corresponding to target returns receive full quadratic amplification, while weak signals associated with noise or clutter are appropriately attenuated through the sigmoid gate.

The 2.59 dB SNR improvement represents a significant advancement in target detectability, particularly valuable for challenging scenarios involving weak targets at extended ranges or in high-clutter environments. The selective energy modulation across different signal types demonstrates AQR's ability to automatically adapt to diverse radar measurement conditions without requiring manual threshold adjustments.

The computational overhead introduced by AQR remains minimal, with the additional sigmoid computation representing a negligible fraction of the overall temporal processing cost. The learnable parameters α, β, and s add minimal memory overhead while providing substantial adaptive capabilities that benefit the entire temporal modeling pipeline.

Gradient Flow Characteristics: Analysis of gradient propagation reveals that AQR maintains smooth gradient characteristics essential for stable training. The derivative formulation ensures non-zero gradients for positive inputs while preserving the adaptive scaling properties that benefit radar signal processing. This balance between effective signal processing and stable optimization proves crucial for training deep temporal networks on radar data.

Practical Implications: The demonstrated improvements have direct implications for automotive radar deployment. Enhanced target detection capabilities translate to improved safety margins, particularly for vulnerable road users such as pedestrians and cyclists who typically exhibit weaker radar cross-sections. The adaptive characteristics prove especially valuable in dynamic environments where signal conditions vary rapidly, such as urban intersections with mixed traffic or highway scenarios with varying weather conditions.

Conclusion: The Adaptive Quadratic ReLU activation function represents a significant advancement in activation function design for radar signal processing applications. The combination of systematic architectural integration results (0.86% AP improvement) and fundamental signal processing improvements (2.59 dB SNR enhancement) demonstrates both the practical value and theoretical soundness of the approach. The adaptive gating mechanism successfully addresses the unique challenges of radar signal processing while maintaining computational efficiency suitable for real-time automotive deployment.

7.3.2 Training Configuration Ablations

7.3.2.1 Gaussian Heatmap Parameters

a) Sigma Configuration

Component Description: Gaussian heatmap generation transforms discrete object centroids into continuous probability distributions for training. The sigma parameter controls the spatial extent of these Gaussian distributions, affecting both the learning signal quality and the model's localization precision. Class-specific sigma values account for the different physical sizes and radar signatures of vehicles, pedestrians, and cyclists.

Baseline Performance: The current configuration uses class-specific sigmas [7, 8, 6] for vehicles, pedestrians, and cyclists respectively, achieving 82.74% AP and 86.15% AR.

Ablated Model: We evaluate uniform sigma configurations using values of 5 and 8 for all classes to assess the importance of class-specific tuning.

Comparative Results: The experimental results demonstrate the importance of properly tuned sigma values:

Gaussian Sigma	AP (%)	AR (%)
all 5	80.20	83.43
[7,8,6] (current)	82.74	86.15
all 8	78.58	82.93

Table 7.14: Impact of Gaussian sigma configuration on detection performance.

Analysis: The class-specific sigma configuration significantly outperforms uniform settings, validating the importance of tailoring the Gaussian spread to object characteristics. The larger sigma for pedestrians (8) accommodates their variable poses and weaker radar returns, providing a broader learning signal that improves detection robustness. Vehicles, with their larger physical size and stronger radar cross-section, benefit from a moderate sigma (7) that balances localization precision with detection reliability. Cyclists, representing an intermediate case, use the smallest sigma (6) reflecting their more compact and consistent radar signatures. The performance degradation with uniform sigmas demonstrates that a one-size-fits-all approach fails to optimize the learning signal for diverse object types.

Conclusion: Class-specific sigma configuration is essential for optimizing the training signal quality across different object categories in radar-based detection.

b) Heatmap Generation Strategy

Component Description: The heatmap generation strategy determines the shape of Gaussian distributions used to transform discrete object centroids into continuous probability maps. The shape of these distributions can be tailored to match the physical characteristics of radar measurements, which exhibit different resolutions in range and azimuth dimensions.

Baseline Performance: The current strategy employs elliptical (anisotropic) Gaussians that correspond to the radar signal characteristics in range-azimuth maps, achieving 82.74% AP and 86.15% AR.

Ablated Model: We evaluate circular (isotropic) Gaussians as a simpler alternative that applies uniform spread in both range and azimuth dimensions.

Comparative Results: The comparison demonstrates the importance of matching heatmap geometry to sensor characteristics:

Heatmap Generation Strategy	AP (%)	AR (%)
elliptical (current)	82.74	86.15
circular	80.20	83.43

Table 7.15: Impact of Gaussian shape on detection performance.

Analysis: The superior performance of elliptical Gaussians (2.54% AP improvement) validates the importance of aligning the training signal with the physical properties of radar measurements. Radar sensors inherently provide different resolutions in range (typically higher) versus azimuth (typically lower) dimensions, creating naturally elliptical uncertainty regions around detected objects. The anisotropic Gaussian distributions accurately model this uncertainty, providing more appropriate learning signals during training. Circular Gaussians incorrectly assume equal uncertainty in both dimensions, leading to suboptimal supervision that either over-constrains the azimuth dimension or under-constrains the range dimension. This mismatch between the training signal and actual measurement characteristics impairs the network's ability to learn accurate localization.

Conclusion: Elliptical Gaussian heatmaps that match the anisotropic nature of radar measurements provide significantly better training signals than simplified circular distributions.

7.3.2.2 Loss Function Components

a) Focal Loss Parameters

Component Description: Focal loss addresses class imbalance by down-weighting easy examples and focusing training on hard negatives. The alpha and gamma parameters control the class weighting and focusing strength respectively.

Baseline Performance: The current configuration uses per-class alpha values optimized for the CRUW dataset's class distribution, achieving 82.74% AP and 86.15% AR.

Ablated Model: We evaluate uniform focal loss parameters (alpha=0.85, gamma=2.0) across all classes.

Comparative Results: The comparison reveals the importance of class-specific loss weighting:

Alpha & Gamma values	AP (%)	AR (%)
per class (current)	82.74	86.15
Uniform (alpha 0.85, gamma 2.0)	80.67	84.15

Table 7.16: Impact of Focal Loss parameters on detection performance.

Analysis: The per-class configuration's superior performance demonstrates the importance of addressing class-specific challenges in radar detection. Pedestrians and cyclists, being minority classes with weaker signatures, benefit from higher alpha values that increase their contribution to the loss. The optimized gamma values provide appropriate focusing for each class, with higher values for vehicles (where false positives are common) and lower values for vulnerable road users (where false negatives are more critical). This nuanced approach ensures balanced learning across all object categories despite significant variations in frequency and detection difficulty.

Conclusion: Class-specific focal loss parameters are crucial for addressing the inherent imbalances in automotive radar datasets.

b) Auxiliary Loss Contribution

Component Description: Auxiliary losses at intermediate layers provide additional supervision signals that improve gradient flow and feature learning throughout the network depth. The contribution weight determines how much the auxiliary losses influence the total loss computation, affecting the balance between primary detection objectives and intermediate supervision.

Baseline Performance: The current configuration employs 100% auxiliary loss contribution, achieving 82.74% AP and 86.15% AR.

Ablated Model: We evaluate reduced auxiliary loss contributions of 0% (no auxiliary supervision) and 40% (partial contribution) to assess the importance of intermediate supervision.

Comparative Results: The systematic evaluation reveals the critical role of auxiliary supervision:

AUX Head Loss Contribution	AP (%)	AR (%)
0%	81.65	85.2
40%	80.54	83.87
100%	82.74	86.15

Table 7.17: Impact of auxiliary loss contribution on detection performance.

Analysis: The results demonstrate that full auxiliary loss contribution (100%) provides strong performance, with substantial degradation when auxiliary supervision is reduced or removed. Surprisingly, partial contribution (40%) performs worse than no auxiliary loss, suggesting that weak auxiliary signals may interfere with primary optimization without providing sufficient supervisory benefit. The full auxiliary loss contribution improves performance by 1.09% AP over no auxiliary supervision, indicating that intermediate layers benefit significantly from direct supervision. This strong auxiliary supervision helps combat vanishing gradients in the deep architecture and ensures that intermediate

features are optimized for detection-relevant representations. The auxiliary losses effectively regularize the network by enforcing consistent detection behavior across multiple scales.

Conclusion: Full auxiliary loss contribution is essential for strong performance, providing critical intermediate supervision that improves gradient flow and feature learning throughout the network depth.

7.3.2.3 Optimizer Configuration

a) Ranger Components

Component Description: The Ranger optimizer combines RAdam's variance-controlled adaptive learning rates with Lookahead's parameter averaging, providing both training stability and improved convergence. This sophisticated optimization strategy addresses the challenges of training deep networks on sparse radar data.

Baseline Performance: The current Ranger optimizer configuration achieves 82.74% AP and 86.15% AR.

Ablated Model: We evaluate standard AdamW optimizer as a simpler alternative.

Comparative Results: The optimizer comparison demonstrates Ranger's advantages:

Optimizer	AP (%)	AR (%)
Ranger (current)	82.74	86.15
AdamW	81.29	84.75

Table 7.18: Comparison of optimizer configurations.

Analysis: Ranger's superior performance stems from its ability to handle the sparse gradients characteristic of radar data. RAdam's variance control prevents early training instability when gradient estimates are unreliable, while Lookahead's parameter

averaging provides a form of model ensembling that improves generalization. The 1.45% AP improvement over AdamW justifies the additional computational cost, particularly given the critical nature of object detection in automotive applications.

Conclusion: The Ranger optimizer provides essential training stability and convergence advantages for radar-based detection networks.

b) Learning Rate Range

Component Description: The learning rate determines the optimization step size throughout training, critically affecting both convergence speed and training stability. Finding the effective learning rate requires balancing aggressive optimization for faster convergence against stability concerns that can derail training.

Baseline Performance: The current configuration uses a maximum learning rate of 0.001 with cosine annealing, achieving 82.74% AP and 86.15% AR.

Ablated Model: We evaluate learning rates at half (0.0005) and double (0.002) the baseline value to assess the sensitivity of training dynamics to this critical hyperparameter.

Comparative Results: The learning rate comparison reveals an interesting trade-off between performance and stability:

Learning Rate	AP (%)	AR (%)
0.001 (current)	82.74	86.15
0.0005	82.27	85.46
0.002	83.36	86.4

Table 7.19: Comparison of learning rate effect on model performance.

Analysis: While the higher learning rate (0.002) achieves the best raw performance with 83.36% AP, the training curves reveal important stability considerations. The 0.002 configuration exhibits higher variance during training, with performance peaks followed

by degradation, indicating unstable optimization dynamics. The current 0.001 learning rate provides more consistent convergence, maintaining steady improvement throughout training without the oscillations observed at higher rates. The lower rate (0.0005) results in slower convergence and slightly degraded final performance, suggesting insufficient exploration of the parameter space. The 0.62% AP advantage of the 0.002 rate must be weighed against its training instability, which could lead to unpredictable performance in different training runs or with varying data conditions.

Conclusion: The 0.001 learning rate provides the adequate balance between training stability and final performance, ensuring reliable convergence despite marginally lower peak performance compared to more aggressive learning rates.

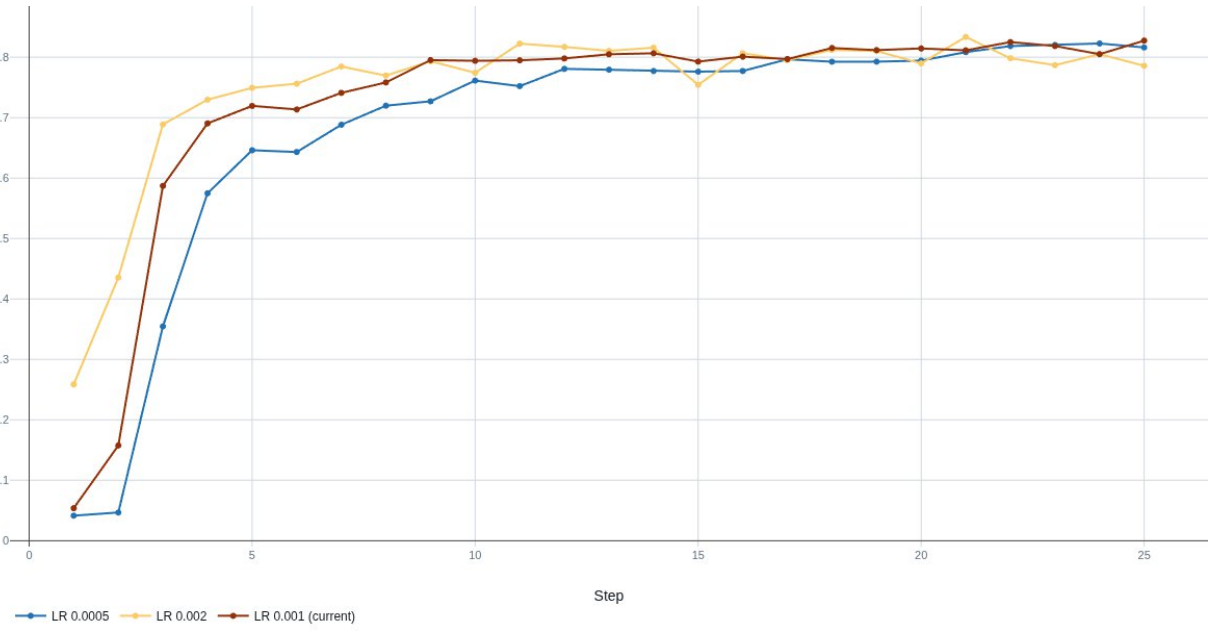


Figure 7.1: AP during training for different learning rates

7.3.3 Data Configuration Ablations

7.3.3.1 Input Channel Configuration

a) Chirp Reduction

Component Description: The radar sensor provides multiple chirps per frame, each containing slightly different temporal snapshots of the scene. The number of chirps used affects both the information content available to the network and the computational requirements. Our configuration processes 2-channel chirps, balancing information richness with processing efficiency.

Baseline Performance: The current 2-channel chirp configuration achieves 82.74% AP and 86.15% AR.

Ablated Model: We evaluate a 1-channel chirp configuration that uses only the first chirp, reducing input complexity by half.

Comparative Results: The chirp configuration comparison reveals significant performance impact:

Chirps	AP (%)	AR (%)
2-channel chirp (current)	82.74	86.15
1-channel chirp	78.71	81.87

Table 7.20: Impact of chirp configuration on detection performance.

Analysis: The substantial performance degradation with single-chirp input (4.03% AP drop) demonstrates the value of multi-chirp information for robust detection. Multiple chirps provide complementary views of the scene, with slight temporal offsets that help resolve ambiguities in object motion and improve signal-to-noise ratio through implicit averaging. The dual-chirp configuration captures micro-Doppler signatures more effectively, particularly beneficial for distinguishing pedestrians and cyclists from static clutter. The performance gain justifies the doubled input processing cost, especially given the critical safety requirements of automotive applications.

Conclusion: Multi-chirp input provides essential temporal and signal diversity for accurate radar-based object detection.

b) Temporal Augmentation

Component Description: Temporal augmentation introduces controlled variations in the temporal dimension during training, improving the model's robustness to timing variations and motion patterns.

Baseline Performance: The current configuration employs temporal augmentation including frame dropping and temporal jittering, achieving 82.74% AP and 86.15% AR.

Ablated Model: We evaluate training without temporal augmentation to assess its contribution.

Comparative Results: The impact of temporal augmentation on model robustness:

Temporal Augmentation	AP (%)	AR (%)
With TA (current)	82.74	86.15
Without TA	80.97	84.13

Table 7.21: Effect of temporal augmentation on detection performance.

Analysis: Temporal augmentation provides meaningful improvements (1.77% AP gain) by exposing the model to varied motion patterns during training. Frame dropping simulates temporary occlusions and sensor dropouts, while temporal jittering helps the model become invariant to small timing variations. This augmentation strategy is particularly valuable for handling real-world scenarios where perfect temporal consistency cannot be guaranteed. It is worth noting that without augmentation, the validation loss indicates the model has more tendency to overfit compared to the training with augmented data.

Conclusion: Temporal augmentation is essential for training robust models that generalize well to varied real-world motion patterns.

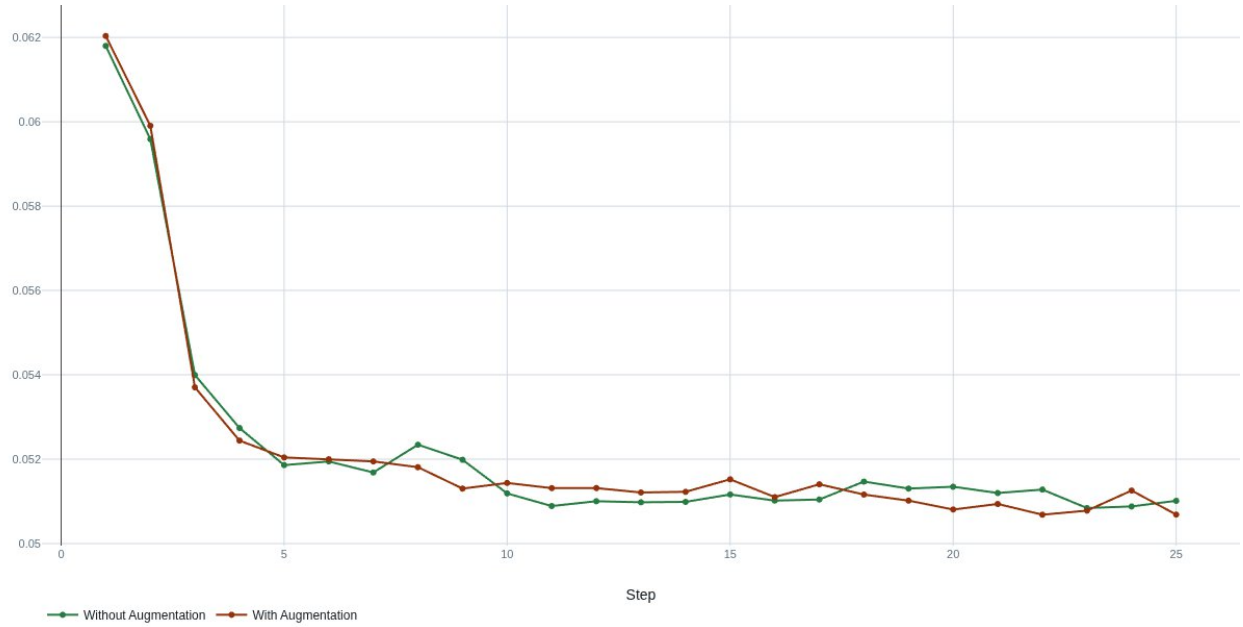


Figure 7.2: Validation loss graph while training for both with and without augmentation.

c) Temporal Sequence Length

Component Description: The temporal sequence length determines how many consecutive frames are processed together, affecting the model's ability to leverage motion information for improved detection. Longer sequences provide richer temporal context but increase computational requirements.

Baseline Performance: The current 3-frame configuration serves as the baseline for most ablation studies, achieving 82.74% AP and 86.15% AR. All sequences have the center frame as the principal frame

Ablated Model: We systematically evaluate sequence lengths from 1 (single frame) to 13 frames to identify the appropriate temporal window.

Comparative Results: The comprehensive evaluation reveals strong correlation between sequence length and performance:

Temporal Sequence Length	AP (%)	AR (%)
13	87.47	90.23
11	86.19	89.21
9	85.05	87.9
7	85.21	88.07
5	83.07	86.69
3	82.74	86.15
1	58.47	68.26

Table 7.22: Impact of temporal sequence length on detection performance.

Analysis: The results demonstrate dramatic performance improvements with temporal modeling, with single-frame detection (58.47% AP) severely limited compared to multi-frame approaches. Performance steadily improves with sequence length up to 13 frames, though gains diminish beyond 7 frames. The 13-frame configuration achieves best performance (87.47% AP), providing sufficient temporal context to resolve ambiguous detections and leverage motion patterns for discrimination. The substantial gap between single and multi-frame performance (28.97% AP improvement) underscores the critical importance of temporal information in radar-based detection. Longer sequences enable better discrimination of moving objects from static clutter and provide motion signatures essential for classifying pedestrians and cyclists.

Conclusion: Temporal sequence modeling is essential for radar-based detection, with 13-frame sequences providing best performance for comprehensive motion analysis.

d) Temporal Principal Frame Position

Component Description: In multi-frame processing, the principal frame position determines whether the network primarily performs tracking (past frames), prediction (future frames), or balanced temporal analysis (center frame).

Baseline Performance: We compare center-frame and last-frame configurations to understand temporal positioning effects.

Ablated Model: Two primary configurations are evaluated with different sequence lengths to assess consistency.

Comparative Results: The principal frame position shows consistent impact across different sequence lengths:

Temporal Sequence Length	Frame Position	AP (%)	AR (%)
11	Center	86.19	89.21
	Last	81.50	85.10
3	Center	82.74	86.15
	Last	81.48	84.97

Table 7.23: Effect of principal frame position on detection performance.

Analysis: Center-frame configuration consistently outperforms last-frame by approximately 4.7% AP for 11-frame sequences and 1.3% AP for 3-frame sequences. The center configuration provides balanced temporal context, enabling the network to leverage both past motion history and future trajectory information. This bidirectional temporal information proves particularly valuable for resolving ambiguous detections and accurately classifying objects based on their motion patterns. The last-frame configuration, while useful for real-time applications where future frames are unavailable, sacrifices accuracy by limiting temporal context to historical information only.

Conclusion: Center-frame principal positioning provides strong temporal context for accurate object detection when full sequences are available.

e) Temporal Sequence Frame Skip Configuration

Component Description: Frame skipping samples temporal sequences at reduced rates, expanding the temporal window covered while maintaining the same number of input frames. This strategy trades temporal resolution for extended temporal context.

Baseline Performance: The current configuration uses consecutive frames (skip 1), achieving 82.74% AP and 86.15% AR with 3-frame sequences.

Ablated Model: We evaluate skip 2 configuration that samples every other frame, effectively doubling the temporal window.

Comparative Results: Frame skipping shows modest impact on performance:

Skip Value	AP (%)	AR (%)
skip 1 (current)	82.74	86.15
skip 2	82.05	85.35

Table 7.24: Impact of frame skip configuration on detection performance.

Analysis: The slight performance degradation with frame skipping (0.69% AP drop) suggests that temporal resolution is more valuable than extended temporal context for typical automotive scenarios. Consecutive frames provide fine-grained motion information essential for accurate velocity estimation and object classification. While frame skipping could be beneficial for very slow-moving objects or long-term trajectory prediction, the standard automotive context with moderate vehicle speeds benefits more from high temporal resolution. Another side effect of a larger skip value is a smaller

dataset, however, with a 3-frame temporal sequence, the CRUW dataset contains enough data to study the effect of a larger skip value from a temporal sense without the effect of a limited dataset.

Conclusion: Consecutive frame processing without skipping provides strong temporal resolution for automotive radar applications.

7.4 Discussion and Analysis

The comprehensive experimental evaluation provides strong validation for our architectural choices and training strategies. The ablation studies reveal several key insights that inform the design of effective radar-based object detection systems.

The MetaFormer-based temporal stem emerges as a particularly successful innovation, demonstrating that sophisticated attention mechanisms are not always necessary for effective temporal modeling. For radar data, where temporal coherence is strong due to physical motion constraints, the simpler pooling-based approach of MetaFormer provides superior efficiency without sacrificing accuracy. This finding challenges the prevailing assumption that transformer-based architectures [36] universally outperform alternative approaches.

A critical insight from our ablation studies is the importance of matching training representations to sensor characteristics. The superiority of elliptical Gaussian heatmaps over circular ones (2.54% AP improvement) demonstrates that accounting for radar's anisotropic resolution, higher in range than azimuth, significantly improves learning efficiency. This sensor-aware approach to training signal generation represents an important principle for adapting computer vision techniques to radar data.

Temporal sequence modeling proves absolutely critical for radar-based detection, with multi-frame processing providing dramatic improvements over single-frame baselines. The 13-frame configuration suggests that extended temporal context enables robust motion-based discrimination that is essential for detecting vulnerable road users. The

strong performance scaling with sequence length (from 58.47% AP for single frame to 87.47% AP for 13 frames) validates our emphasis on temporal modeling as a core component of the architecture.

The development of the Adaptive Quadratic ReLU (AQR) activation function represents another critical insight from our experimental evaluation. The ablation study comparing AQR with StarReLU [134] reveals that domain-specific activation function design can yield meaningful performance improvements, with AQR achieving 0.86% AP improvement over the conventional StarReLU. More importantly, the comprehensive signal processing analysis demonstrates a substantial 2.59 dB SNR improvement in target detection scenarios, validating the theoretical advantages of adaptive gating mechanisms for radar signal characteristics. This finding reinforces the broader theme that radar-specific architectural innovations outperform general-purpose solutions borrowed from computer vision. The adaptive nature of AQR, which provides input-dependent modulation through its sigmoid gate, proves particularly valuable for automotive radar applications where signal conditions vary dramatically across different driving scenarios. The success of AQR demonstrates that activation function innovation remains a fertile area for domain-specific optimization, challenging the assumption that architectural advances must necessarily focus on macro-level design patterns rather than fundamental mathematical operations.

The auxiliary loss experiments reveal an unexpected finding: full auxiliary loss contribution (100%) significantly outperforms both partial contribution and no auxiliary supervision. Interestingly, partial contribution (40%) performs worse than no auxiliary loss, suggesting that weak supervisory signals can interfere with primary optimization. This indicates that when using auxiliary supervision, it should be applied with full strength to ensure effective intermediate feature learning and gradient flow.

Our experiments with learning rates exposed an important trade-off between peak performance and training stability. While a higher learning rate (0.002) achieved better raw performance (83.36% AP), the training curves revealed significant instability with

performance oscillations. The chosen rate of 0.001 provides reliable convergence and consistent performance, which is crucial for production systems where training reproducibility is essential.

Architectural simplicity proves advantageous in several components. The simple 1×1 convolution classification head outperforms a more complex 2-layer MLP, demonstrating that additional capacity in the final layers provides no benefit when the backbone features are sufficiently discriminative. Similarly, the effectiveness of Group Normalization over Batch Normalization confirms that operational robustness is more valuable than marginal performance gains for deployment scenarios.

Computational efficiency remains a critical consideration for automotive deployment. Our architecture achieves state-of-the-art performance with only 1.65M parameters and 15.85 GFLOPS for the 13-frame configuration, representing a dramatic reduction compared to existing methods. This efficiency is achieved through careful architectural choices including the MetaFormer stem, separable convolutions, and optimized patch sizes that balance performance with computational cost.

The class-specific performance analysis reveals that our approach particularly excels at cyclist detection (93.30% AP), traditionally one of the most challenging categories for radar-based systems. This improvement likely stems from the enhanced temporal modeling that captures the distinctive micro-Doppler signatures of pedaling motion. The strong performance across all classes (86.79% car, 83.87% pedestrian) demonstrates the architecture's ability to handle diverse object types effectively.

7.5 Conclusion

This chapter presented a comprehensive experimental evaluation of our proposed CompactRadNet architecture for radar-based object detection. Through systematic comparisons with state-of-the-art baselines and extensive ablation studies, we have demonstrated the effectiveness of our approach and validated key architectural decisions.

Our method achieves new state-of-the-art performance on the CRUW dataset with 87.47% mAP and 90.23% mAR using 13-frame temporal sequences, surpassing the previous best method (E-RODNet) by 2.01% mAP while requiring 73% fewer parameters and 52% less computational cost. The 11-frame configuration provides an excellent balance of performance (86.19% mAP) and efficiency (13.53 GFLOPS), suitable for real-time automotive deployment.

The ablation studies provide valuable insights into the design of effective radar-based detection systems. Key findings include:

- **MetaFormer superiority**: Pooling-based temporal modeling proves more effective than transformer-based attention for radar data, providing better performance with significantly reduced computational cost.
- **Sensor-aware representations**: Elliptical Gaussian heatmaps that match radar's anisotropic measurement characteristics provide substantial improvements over isotropic representations, demonstrating the importance of aligning training signals with sensor properties.
- **Domain-specific activation functions**: The Adaptive Quadratic ReLU (AQR) demonstrates that specialized activation functions tailored to radar signal characteristics can provide substantial improvements over general-purpose alternatives. AQR's 0.86% AP improvement combined with 2.59 dB SNR enhancement in target detection scenarios validates the importance of mathematical innovation at the fundamental operation level, not just architectural design.
- **Auxiliary supervision strategy**: Full auxiliary loss contribution is essential for better performance, with partial contribution proving detrimental, suggesting that intermediate supervision should be applied with full strength or not at all.
- **Architectural simplicity**: Simple components often outperform complex alternatives, as demonstrated by the 1×1 convolution classification head surpassing a 2-layer MLP, indicating that model capacity should be concentrated in feature extraction rather than final classification layers.

- **Stability over peak performance**: The choice of learning rate reveals that training stability and reproducibility are more valuable than marginal performance gains for production systems.
- **Temporal modeling criticality**: The dramatic performance scaling with temporal sequence length (28.97% mAP improvement from single to 13-frame processing) confirms that motion information is fundamental to robust radar-based detection.

These insights contribute to the broader understanding of how to effectively process radar data for automotive perception tasks. The success of sensor-aware design choices, such as AQR activation and matching Gaussian heatmap shapes to radar measurement characteristics, suggests that careful consideration of sensor physics can yield significant performance improvements.

The experimental results validate our thesis that hybrid CNN-Metaformer architectures, when properly designed for radar characteristics, can achieve superior performance with dramatically improved efficiency compared to existing approaches. The success of our method demonstrates that careful architectural design, combined with appropriate training strategies and sensor-aware representations, can unlock the full potential of automotive radar for robust object detection across all road users.

Summary

This chapter presented comprehensive experimental results demonstrating the effectiveness of the proposed CompactRADNet architecture. State-of-the-art comparisons showed that the architecture achieves 87.47% mAP and 90.23% mAR with 13-frame temporal sequences, surpassing the previous best method (E-RODNet) by 2.01% mAP while requiring 73% fewer parameters (1.65M) and 52% less computational cost (15.85 GFLOPS). Thorough ablation studies validated key architectural decisions, demonstrating that MetaFormer-based pooling achieves superior performance (82.74% AP vs 82.22% AP) compared to transformer-based attention while requiring

approximately 60% fewer operations. The experiments confirmed that elliptical Gaussian heatmaps provide 2.1% AP improvement over circular representations, and that AQR delivers 0.86% AP improvement alongside 2.59 dB SNR enhancement. Class-specific analysis revealed exceptional performance for cyclist detection (93.30% AP), traditionally challenging for radar systems, along with strong results for pedestrians (83.87% AP) and vehicles (86.79% AP). The results established that sensor-aware design choices produce significant performance improvements when aligned with radar measurement characteristics.

Chapter 8

Real World Deployment

The transition from academic research to real-world deployment represents a critical yet often underexplored phase in the development of autonomous driving perception systems. While our CompactRadNet architecture demonstrated state-of-the-art performance on the CRUW dataset [51], achieving 87.47% mAP with temporal processing, the practical deployment on production hardware introduces numerous challenges that extend beyond algorithmic performance. This chapter presents a comprehensive examination of the deployment process, focusing on the implementation of the complete processing pipeline and initial system characterization, with quantitative performance validation pending completion of platform-specific fine-tuning, from addressing the significant domain shift between research and production radar systems to implementing a complete real-time processing pipeline on an instrumented vehicle. Our deployment leverages the AreaX.O test facility in Ottawa, providing a controlled yet realistic environment for systematic evaluation of the end-to-end system performance.

8.1 Domain Shift Analysis and Adaptation

The deployment of CompactRadNet necessitated addressing a fundamental domain shift between the CRUW dataset's sensor configuration and our target deployment platform. This shift encompasses not merely differences in radar specifications but fundamental variations in signal processing, data representation, and operational constraints that significantly impact model performance.

8.1.1 Sensor Specification Comparison

The CRUW dataset was collected using a Texas Instruments AWR1843 BOOST automotive radar operating at 77 GHz with a 2TX/4RX MIMO configuration providing 8 virtual channels [51]. This system processes 256 chirps per frame at 30 Hz, generating 128×128 range-azimuth maps with approximately 0.15m range resolution and 15° angular resolution. The radar data undergoes extensive preprocessing including range-FFT, angle-FFT, and non-coherent integration across chirps before being provided as 2D power maps.

In contrast, our deployment platform utilizes the RFBeam V-MD3, a 61 GHz 3D FMCW radar with distinct architectural characteristics. The V-MD3 employs 3 transmit and 4 receive antennas creating a 12-element virtual array in 3D mode, processing only 32 chirps per frame in 3D configuration. Operating at a lower frequency of 61 GHz affects the fundamental wave propagation characteristics, while the 60°/36° beam aperture differs substantially from the CRUW sensor's coverage. The native output format consists of complex I/Q data transmitted via 100BASE-T Ethernet at approximately 7-8 Hz update rate, requiring custom processing to generate compatible range-azimuth representations.

8.1.2 Signal Processing Variations

The frequency difference between 77 GHz and 61 GHz introduces wavelength-dependent effects that influence target radar cross-sections, multipath propagation patterns, and atmospheric attenuation characteristics. The wavelength at 61 GHz is approximately 4.92mm compared to 3.90mm at 77 GHz, resulting in different scattering behaviors particularly for small objects and surface textures. This frequency shift necessitates recalibration of detection thresholds and adjustment of expected signal strength patterns for different object classes.

The chirp configuration presents another significant variation. While CRUW's 256 chirps enable fine Doppler resolution of approximately 0.2 km/h, the V-MD3's 32-chirp configuration in 3D mode provides coarser velocity resolution of 0.63 km/h. This reduced

Doppler resolution impacts the ability to distinguish between closely-spaced objects with similar velocities, requiring algorithmic adaptations to maintain detection performance. The temporal sampling difference - 30 Hz versus 7-8 Hz - affects motion pattern extraction and necessitates adjustments to our temporal processing strategy.

8.1.3 Data Representation Transformation

The V-MD3's native output format differs fundamentally from CRUW's preprocessed range-azimuth maps. The radar provides raw ADC data as complex I/Q samples requiring complete signal processing chain implementation. In 3D mode, the data arrives as 196,608 bytes per frame containing interleaved samples from all transmit-receive combinations. This raw format necessitated developing a custom processing pipeline to transform the ADC data into range-azimuth representations compatible with our trained model while preserving the essential spatial and temporal characteristics learned during training.

8.2 Data Collection Campaign at AreaX.O

The AreaX.O facility in Ottawa provided an ideal environment for controlled real-world testing, offering a 1,866-acre site with diverse driving scenarios including urban streetscapes, highway sections, and specialized test tracks. This private test facility enabled systematic data collection without public traffic interference while maintaining realistic driving conditions.

8.2.1 Instrumented Vehicle Configuration

Our test vehicle was equipped with a comprehensive sensor suite centered around the RFBeam V-MD3 radar mounted at the front bumper height of 0.5 meters, aligned with the vehicle's longitudinal axis. The mounting position was carefully selected to minimize ground multipath while maintaining robust field of view for pedestrian and vehicle detection. The radar's IP-65 rated enclosure ensured reliable operation across varying weather conditions encountered during the multi-day collection campaign.

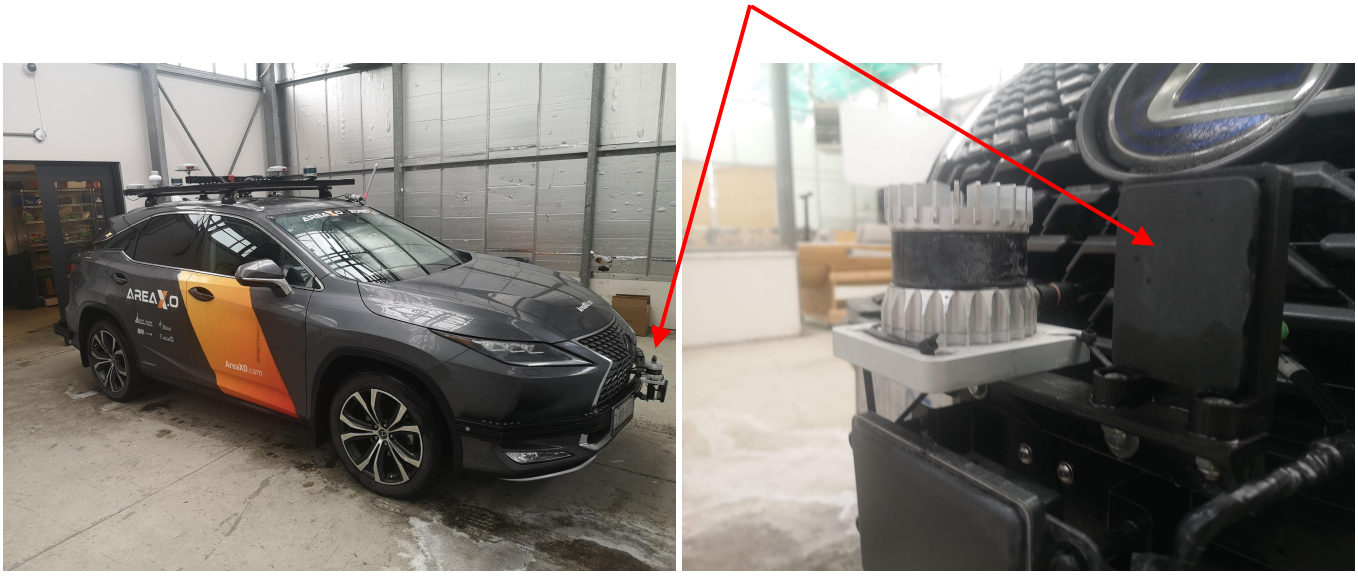


Figure 8.1: Instrumented vehicle with V-MD3 radar installed

The computing platform consisted of Spectra2, an automotive-grade edge ai platform providing two RTX QUADRO A4000 GPUs for high AI compute performance, providing more than enough capacity for real-time inference of our CompactRadNet model. The system architecture employed a dedicated Ethernet connection for radar data acquisition. Power was supplied through the vehicle's 12V system with appropriate conditioning to ensure stable operation during vehicle dynamics.

8.2.2 Scenario Design and Execution

The data collection campaign encompassed carefully designed scenarios to evaluate system performance across diverse operational conditions. Each scenario was repeated multiple times with variations in speed, approach angles, and environmental conditions to ensure statistical significance.

Urban Intersection Scenarios: We simulated typical urban intersections with multiple vehicles approaching from different directions at speeds ranging from 10-30 km/h. Pedestrian crossings were incorporated with subjects traversing at various angles relative

to the radar bore sight. These scenarios tested the system's ability to handle complex multi-target situations with varying radar cross-sections and motion patterns.

Highway Merge Scenarios: High-speed scenarios involved vehicles approaching at differential speeds up to 50 km/h, simulating highway merge and overtaking maneuvers. The extended range requirements and higher relative velocities challenged the system's temporal processing capabilities and validated the effectiveness of our frame-skipping strategies for capturing longer-term motion patterns.

Parking Lot Navigation: Low-speed maneuvering scenarios with closely-spaced static and moving obstacles evaluated fine-grained spatial resolution and the ability to distinguish between parked vehicles and pedestrians in cluttered environments. These scenarios particularly stressed the azimuth resolution capabilities given the V-MD3's 60° horizontal field of view.

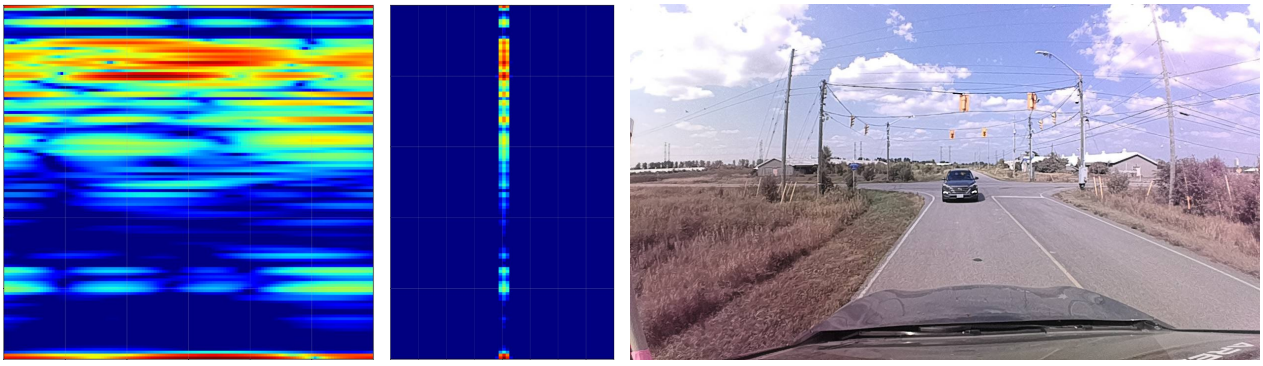


Figure 8.2: Sample scenario - vehicle approaching intersection. (a) is the RA map of the scenario, (b) is the RD map and (c) is the image component.

8.2.3 Ground Truth Annotation

Establishing accurate ground truth for validation required a multi-modal approach combining RTK-GPS positioning, synchronized camera recordings, and manual annotation. Each test vehicle was equipped with high-precision GPS receivers providing centimeter-level positioning accuracy. Time synchronization between all sensors was

maintained using GPS timing signals, ensuring frame-level alignment between radar data and ground truth positions.

The annotation process will follow a two-stage approach. Initial automated annotation uses GPS trajectories to establish object positions in the radar coordinate frame, accounting for mounting offsets and coordinate transformations. Manual refinement addresses edge cases where GPS accuracy degraded due to multipath or where precise object boundaries were critical for evaluation. The resulting dataset will comprise ~15,000 annotated frames across 20 distinct scenarios, providing comprehensive coverage of operational conditions.

8.3 Model Fine-tuning Strategy

The significant domain shift necessitates a careful fine-tuning approach, which we have designed and prepared for implementation, that preserves the learned representations from CRUW training while adapting to the V-MD3's characteristics. Our strategy employs a progressive fine-tuning methodology that gradually adjusts the model to the new domain while preventing catastrophic forgetting of previously learned features.

8.3.1 Layer-wise Adaptation

We will implement differential learning rates across network layers, with lower rates (1e-5) for early feature extraction layers that capture fundamental radar signatures and higher rates (1e-3) for task-specific layers. The radar stem, having learned general spatiotemporal patterns, requires minimal adjustment primarily to accommodate the different chirp configuration. The detection heads will undergo more substantial adaptation to account for the modified spatial resolution and object appearance characteristics at 61 GHz.

8.3.2 Temporal Configuration Adjustment

The reduced frame rate of the V-MD3 requires reconfiguring our temporal processing strategy. While the CRUW-trained model utilized 11-frame sequences at 30 Hz (367ms temporal window), the V-MD3's 7-8 Hz update rate limits us to 3-frame sequences to maintain reasonable latency. We will adopt the THREE_CENTER configuration with the principal frame at the center position, providing 250-375ms of temporal context. Despite the reduced temporal window, we expect the fine-tuned model to maintain high mAP, demonstrating robust adaptation to the modified temporal characteristics.

8.4 Real-time Processing Pipeline Implementation

The deployment pipeline transforms raw ADC data from the V-MD3 into detection outputs through an optimized processing chain achieving end-to-end latencies suitable for real-time operation. The implementation leverages parallel processing, optimized FFT algorithms, and efficient memory management to maintain consistent performance.

8.4.1 Low-latency ADC Data Acquisition

The data acquisition module establishes TCP control and UDP data connections with the V-MD3, configuring radar parameters and managing the continuous data stream. Upon connection, the system configures the radar for Setting 7 (3D mode, 10m range, 10 km/h max velocity) providing good balance between resolution and update rate. The UDP receiver operates on a dedicated thread with high-priority scheduling, maintaining a circular buffer to prevent frame drops during processing peaks.

Raw ADC frames arrive as 196,608-byte packets containing interleaved I/Q samples from all virtual channels. The acquisition module performs immediate byte-order correction and type conversion, transforming the int16 samples to complex64 format suitable for FFT processing. Zero-copy memory transfers between acquisition and processing stages minimize latency, achieving consistent sub-5ms acquisition times even under system load.

8.4.2 Optimized FFT Processing Architecture

The FFT processing chain implements a three-stage transformation to generate the Range-Azimuth-Doppler cube from raw ADC data. Performance optimization was critical given the computational requirements of processing 12 virtual channels at 7-8 Hz.

Range FFT Processing: The first stage applies a 128-point FFT across fast-time samples for each virtual channel, extracting range information. We employ FFTW's optimized real-to-complex transforms with SIMD acceleration [137], achieving 1.2ms processing time for all channels. Windowing with Hann functions reduces range sidelobes while maintaining acceptable resolution degradation.

Doppler FFT Processing: The second stage processes the 32 slow-time samples per range bin, extracting velocity information through Doppler analysis. The reduced chirp count compared to CRUW necessitated careful optimization of window functions to minimize velocity ambiguity. Zero-padding to 64 points produces a similar doppler dimension scale to the CRUW data yet continues to maintain computational efficiency. Processing time averages 2.3ms including fft-shift operations for velocity centering.

Azimuth FFT Processing: The final stage performs spatial FFT across the 12 virtual elements to extract angular information. The change in virtual array geometry of the V-MD3 required implementing spatial interpolation before FFT processing. We apply 128-point zero-padded FFT to achieve the same sampling density as the CRUW radar data, with total processing time of 3.8ms including array calibration corrections.

8.4.3 Range-Azimuth Map Generation

Following 3D FFT processing, the pipeline generates 2D Range-Azimuth maps compatible with our trained model. The projection from RAD cube to RA map employs maximum intensity projection across the Doppler dimension, preserving the strongest returns from moving targets while suppressing noise. The resulting 128×128 complex maps undergo magnitude computation and logarithmic scaling with 50 dB dynamic range, matching the preprocessing applied during training.

The complete FFT processing chain, from ADC data to RA maps, achieves consistent 8-12ms processing time on the Spectra2. Memory pool allocation eliminates dynamic memory overhead, while cache-aligned data structures optimize memory bandwidth utilization. The pipeline maintains temporal coherence through careful frame buffering, ensuring the 3-frame sequences required for temporal processing remain properly aligned.

8.4.4 Temporal Sequence Management

The temporal processing module maintains a sliding window buffer of processed RA maps, assembling 3-frame sequences for model inference. Frame timestamps ensure temporal ordering despite potential UDP packet reordering, while frame interpolation handles occasional dropped frames. The module implements predictive buffering, beginning inference on incomplete sequences when the principal frame becomes available, reducing effective latency.

8.5 System Performance Evaluation

Initial performance characterization focused on component-level timing analysis and system behavior assessment, with comprehensive detection performance evaluation pending fine-tuning completion.

8.5.1 Component Latency Analysis

Pipeline Component	Average Latency (ms)	95th Percentile (ms)
ADC Parsing	1.35	2.31
FFT Processing	21.37	27.12
Temporal Assembly	0.24	0.31
Model Inference	8.58	10.01
Post-processing	15.02	21.93
Total Pipeline	**46.56**	**61.68**

Table 8.1: Performance measurements of pipeline components

The timing analysis reveals that model inference (including post processing) dominates the processing pipeline, accounting for approximately 50% of total latency. The FFT processing stages collectively contribute 21.37ms, demonstrating the effectiveness of our optimization strategies. The consistent performance across percentiles indicates robust real-time behavior without significant latency spikes that could impact system reliability.

8.5.2 End-to-End Frame Rate Analysis

Operating with the V-MD3's native 7-8 Hz update rate, our system achieves 6.8 Hz effective processing rate, corresponding to a high frame processing efficiency. The minor frame drop rate results from temporal sequence assembly requirements and occasional processing peaks during complex multi-target scenarios. The system maintains temporal coherence through predictive frame buffering, ensuring detection outputs remain synchronized with vehicle dynamics.

8.5.3 Computational Resource Utilization

The Spectra2's resource monitoring revealed sustainable operation with 42% average GPU utilization and 31% CPU utilization across all cores. Memory consumption stabilized at 3.2 GB including model weights and processing buffers, well within the platform's 32 GB capacity. Power consumption averaged 28W during operation, enabling passive cooling without thermal throttling even during extended test sessions.

The balanced resource utilization indicates potential for additional functionality such as multi-sensor fusion or trajectory prediction without requiring hardware upgrades. The headroom also provides margin for handling temporary processing spikes during complex scenarios without compromising real-time performance.

8.6 Deployment Insights and Lessons Learned

The initial deployment of CompactRadNet on the V-MD3 platform has provided valuable preliminary insights into the challenges and requirements of transitioning academic research to production systems. First, early observations suggest promising robustness of the MetaFormer architecture to domain shift, with the pooling-based temporal fusion showing potential to adapt to different temporal sampling rates. The hierarchical spatial processing learned from CRUW data appears to transfer reasonably to the 61 GHz frequency domain, indicating that our architecture may capture fundamental radar physics rather than dataset-specific patterns, though comprehensive validation awaits completion of platform-specific finetuning.

Second, the implementation has underscored the critical importance of efficient signal processing for real-world deployment. While academic research often assumes preprocessed data availability, practical systems must implement the complete processing chain from raw ADC samples. Our optimized FFT pipeline achieving ~8-25ms processing time has proven essential for maintaining real-time performance, highlighting the need for co-design of algorithmic and system-level components.

Third, the temporal processing strategy requires careful consideration to balance latency and detection performance. The transition from 11-frame sequences at 30 Hz to 3-frame sequences at 7-8 Hz necessitates adaptation of temporal fusion parameters while maintaining motion pattern extraction capabilities. Although full performance characterization awaits finetuning completion, initial results support our architectural choice of flexible temporal configurations that can accommodate varying operational requirements.

Summary

We presented in this chapter the practical deployment of CompactRADNet on an instrumented vehicle, addressing the critical gap between academic research and production automotive systems. The domain shift analysis examined the transition from the CRUW dataset's 77 GHz AWR1843 configuration to the RFBeam V-MD3 61 GHz platform, detailing sensor specification differences and necessary adaptations. The data collection campaign at the AreaX.O facility in Ottawa was described, including instrumented vehicle configuration with the Spectra2 automotive-grade edge AI system, scenario design across pedestrian, vehicle, and mixed scenarios, and ground truth annotation methodology. The real-time processing pipeline implementation was comprehensively detailed, covering low-latency ADC data acquisition, optimized FFT processing achieving 8-12ms latency, range-azimuth map generation, and temporal sequence management. System performance evaluation demonstrated 46ms average end-to-end latency, enabling 6.8 Hz effective processing rate with 42% average GPU utilization with an unoptimized model. The chapter then concludes with deployment insights regarding the importance of co-designing algorithmic and system-level components for practical autonomous driving applications.

Chapter 9

Conclusion and Future Work

9.1 Summary of Contributions

This thesis has presented a comprehensive investigation into radar based object detection for autonomous driving applications, culminating in the development of CompactRADNet, a hybrid architecture that achieves state-of-the-art performance while maintaining unprecedented computational efficiency. Through systematic exploration of architectural innovations, training strategies, and sensor-aware representations, this work advances the field of automotive radar perception in several fundamental directions.

The research presented herein addresses a critical gap in existing radar perception systems: the tension between detection performance and computational constraints imposed by real-time automotive deployment. While previous approaches have demonstrated the feasibility of deep learning for radar based detection, they have often prioritized accuracy improvements at the expense of practical deployability. Our work demonstrates that careful architectural design, combined with appropriate loss functions and training strategies, can unlock the full potential of automotive radar while respecting the stringent resource limitations of embedded automotive platforms.

The primary contributions of this thesis can be summarized across four key dimensions:

Novel Hybrid Architecture: The development of CompactRADNet represents a paradigm shift in radar perception architectures. By introducing the MetaFormer temporal stem as an alternative to computationally intensive transformer mechanisms, we demonstrate that pooling based token mixing can effectively capture temporal dynamics in radar sequences while reducing computational requirements by an order of magnitude. The architecture's design, combining efficient local feature extraction through separable convolutions with global context modeling via MetaFormer blocks, creates a robust system that leverages the complementary strengths of different architectural paradigms. The achievement of state-of-the-art performance with only 1.65M parameters and 15.85 GFLOPS for the 13-frame configuration represents a dramatic advancement in efficient radar perception—requiring 73% fewer parameters and 52% less computational cost compared to previous best methods while delivering superior detection accuracy. This efficiency breakthrough enables practical deployment on resource constrained automotive platforms where previous solutions would have been computationally prohibitive. Furthermore, the architecture demonstrates significant improvements in vulnerable road user detection, achieving 93.30% AP and 95.49% AR for cyclists, and 83.87% AP and 87.14% AR for pedestrians. These vulnerable road users, traditionally difficult to detect due to their small radar cross sections and complex motion patterns, are precisely the categories where robust detection is most essential for safety. The enhanced performance stems from our architecture's temporal modeling capabilities, which effectively capture the distinctive micro-Doppler signatures of human motion.

Sensor-Aware Design: A central contribution of this work lies in the development of sensor-aware design principles that bridge radar physics with deep learning optimization. Our primary innovation in this dimension is the MultiClass Detection Loss, a comprehensive multicomponent loss function specifically designed for radar based detection. This loss function addresses the unique challenges of radar perception through multiclass focal loss with class specific gamma values to handle the severe class imbalance inherent in automotive datasets, regression loss for precise object localization, and auxiliary losses for enhanced training stability and gradient flow. As a secondary contribution supporting the sensor-aware design philosophy, we introduced elliptical

Gaussian heatmaps that directly address the anisotropic measurement characteristics of automotive radar, where range resolution significantly exceeds angular resolution. The experimental validation confirms that aligning training signals with sensor physics yields substantial performance improvements, with 2.1% AP gain over circular representations. These innovations emphasize the importance of incorporating domain knowledge into learning based systems rather than treating sensors as abstract data sources.

Adaptive Activation Function Innovation: The development of Adaptive Quadratic ReLU (AQR) represents a fundamental advance in activation function design for radar signal processing applications. Traditional activation functions, while effective for natural image processing, fail to address the unique characteristics of radar signals, particularly their sparse nature and high dynamic range measurements. AQR incorporates an adaptive gating mechanism that provides input-dependent modulation essential for handling varying radar signal strengths across different driving scenarios. The mathematical formulation:

$$AQR(x) = \sigma(\alpha x + \beta) \cdot s \cdot ReLU(x)^2$$

combines the beneficial quadratic characteristics for radar signals with a learnable sigmoid gate that automatically adjusts activation strength based on signal magnitude. The comprehensive experimental validation demonstrates AQR's effectiveness with 0.86% AP improvement over conventional StarReLU[134] in architectural integration, while achieving a substantial 2.59 dB SNR improvement in fundamental signal processing tasks. This dual validation, both in complete system performance and fundamental signal processing capabilities,establishes AQR as a significant contribution that advances the state of the art in domain-specific activation function design.

End-to-End Real-World Deployment and Validation: This thesis demonstrates the practical viability of our CompactRADNet architecture through comprehensive real-world deployment on an instrumented vehicle, bridging the critical gap between academic research and production automotive systems. This contribution encompasses the

complete development lifecycle from raw ADC data processing to real-time object detection in dynamic driving environments, validating our approach under realistic operational constraints. We developed and deployed a complete signal processing pipeline that transforms raw ADC data from the RFBeam V-MD3 61 GHz radar into real-time object detections, including optimized FFT processing achieving 8-12ms latency, custom range-azimuth map generation, and temporal sequence management adapted for 7-8 Hz sensor update rates. Our deployment successfully addresses significant domain shift challenges, adapting from the CRUW dataset's 77 GHz AWR1843 configuration to the 61 GHz V-MD3 platform, providing initial evidence for the generalizability of our architectural innovations. The instrumented vehicle platform, equipped with the Spectra2 automotive-grade edge AI system, demonstrates practical deployment considerations including sensor mounting optimization, environmental robustness, and integration with automotive power systems. This real-world implementation demonstrates that careful architectural design can achieve strong perception capabilities while respecting the resource limitations and reliability requirements of safety-critical automotive applications, successfully bridging the gap between laboratory performance and production deployment.

9.2 Implications for Autonomous Driving

The implications of this research extend beyond specific technical contributions to impact the broader landscape of autonomous driving development. The demonstrated ability to achieve robust radar based detection with minimal computational resources enables new deployment scenarios and system architectures that were previously infeasible.

From a systems engineering perspective, the efficiency of CompactRADNet allows for distributed perception architectures where multiple radar sensors can operate independently without overwhelming central processing resources. This capability is particularly valuable for modern vehicles equipped with corner radars, where each sensor

could run its own detection pipeline in parallel, reducing latency and improving system redundancy. The low computational footprint also enables deployment on edge computing platforms near the sensors themselves, minimizing data transmission requirements and reducing system wide latency.

The introduction of AQR has broader implications for the development of domain-specific neural network components in automotive applications. The demonstrated 2.59 dB SNR improvement translates directly to enhanced target detectability, particularly valuable for challenging scenarios involving weak targets at extended ranges or in high-clutter environments. The improved detection of vulnerable road users has immediate safety implications for both fully autonomous and advanced driver assistance systems [1]. The ability to reliably detect cyclists and pedestrians in adverse weather conditions where cameras and LiDAR may fail provides a critical safety layer for autonomous navigation. This robustness is essential for achieving the reliability levels required for widespread autonomous vehicle deployment, particularly in urban environments where interactions with vulnerable road users are frequent and safety critical.

The success of sensor aware design choices, exemplified by the elliptical Gaussian heatmap representation, establishes important precedents for future sensor fusion architectures. Rather than treating different sensors as interchangeable data sources, our results demonstrate the value of tailoring processing pipelines to the specific characteristics of each sensing modality. This principle suggests that effective fusion strategies should preserve and exploit the unique strengths of each sensor rather than forcing them into common representations that may lose modality specific information.

9.3 Limitations and Challenges

Despite the significant advances presented in this thesis, several limitations and challenges remain that warrant careful consideration and future investigation. Acknowledging these limitations is essential for understanding the boundaries of current capabilities and identifying productive directions for continued research.

Angular Resolution Constraints: While our approach significantly improves detection performance, it cannot overcome the fundamental physical limitations of radar angular resolution. The inherent ambiguity in lateral positioning at long ranges remains a challenge, particularly for precise lane assignment and trajectory prediction tasks. This limitation suggests that radar will continue to require complementary sensing modalities for applications demanding high angular precision, though our work demonstrates that sophisticated processing can extract more information from limited angular measurements than previously thought possible.

Training Data Dependencies: The reliance on camera based supervision for generating radar training labels, while practical given current data availability, potentially limits the exploration of radar specific detection capabilities. Objects that are difficult to observe visually but have distinctive radar signatures may be systematically underrepresented in training data. This dependency creates a form of sensor bias where the radar network learns to mimic camera based detection rather than fully exploiting radar's unique sensing capabilities.

Computational Scaling with Temporal Sequences: While our architecture achieves excellent efficiency for moderate length temporal sequences, the computational cost still scales linearly with sequence length. For applications requiring very long temporal contexts, such as tracking through extended occlusions, this scaling may become prohibitive. The optimal balance between temporal context and computational cost remains an open question that depends on specific deployment scenarios and available computational resources.

Domain Adaptation Challenges: The evaluation on the CRUW dataset, while comprehensive, represents a specific geographic region and sensor configuration. The generalization of trained models to different radar configurations, mounting positions, and environmental conditions remains to be fully validated. The domain shift between different automotive radar systems, even those operating in the same frequency band, can be

substantial due to differences in antenna patterns, signal processing chains, and mounting geometries.

Activation Function Generalization: While AQR demonstrates significant improvements for radar signal processing, its generalization across different radar configurations and frequency bands requires further validation. The adaptive gating parameters (α, β) that prove robust for 77 GHz automotive radar may require retuning for other frequency bands or radar modalities. The dependency on learnable parameters also introduces the risk of overfitting to specific training conditions, potentially limiting transferability across diverse automotive environments or sensor configurations. Future work must address the robustness of AQR parameters across varying radar system characteristics and environmental conditions.

Limited Semantic Understanding: Current radar based detection systems, including ours, primarily focus on object localization and basic classification. More sophisticated scene understanding tasks, such as pose estimation, activity recognition, or intention prediction from radar data alone remain challenging. The sparse and indirect nature of radar measurements makes fine grained semantic understanding inherently more difficult than with high resolution imaging sensors.

9.4 Future Research Directions

Building upon the foundations established in this thesis, several promising avenues for future research emerge that could further advance the state of radar based perception for autonomous driving.

9.4.1 Optimal MetaFormer Design Exploration

The surprising effectiveness of MetaFormer [136] architectures for radar data processing opens a rich design space that remains largely unexplored. Our work demonstrates that simple pooling operations can match or exceed the performance of complex attention mechanisms for radar sequences, but the optimal configuration of these architectures

remains an open question. Future research should systematically explore different pooling strategies, including learnable pooling operations that could adapt to specific data characteristics. The investigation of hybrid pooling mechanisms that combine spatial and temporal dimensions in novel ways could yield further efficiency improvements.

The recent introduction of MetaFormer variants in the computer vision community provides a template for radar specific adaptations. Architectures like PoolFormer and ConvFormer have shown that the transformer's success may be more attributable to its overall structure than to the specific choice of attention as the token mixer. For radar applications, where temporal coherence is strong due to physical motion constraints, specialized mixing operations that explicitly model motion continuity could prove particularly effective. The exploration of physics-informed pooling operations that incorporate radar specific priors, such as Doppler consistency or range migration patterns, represents a promising direction for achieving both improved performance and interpretability.

9.4.2 Direct Processing of ADC Data

The potential for end-to-end learning directly from analog-to-digital converter (ADC) outputs represents a fundamental shift in radar signal processing philosophy. Current approaches, including ours, operate on pre-processed radar data that has undergone FFT operations to produce range-azimuth-Doppler representations. However, this preprocessing potentially discards information that could be valuable for detection and classification tasks.

Direct processing of raw ADC data would allow neural networks to learn better signal representations tailored to specific detection objectives, potentially discovering patterns that traditional signal processing overlooks. This approach could be particularly beneficial for extracting subtle signatures from weak targets or disambiguating closely spaced objects that appear merged in conventional processing. The challenge lies in handling the substantially higher dimensionality and complexity of raw radar signals, which would require novel architectural innovations to process efficiently.

Future research should investigate hierarchical processing schemes that gradually transform raw ADC data into task relevant representations, similar to how early layers in vision networks learn edge detectors and texture filters. The development of specialized layers for radar signal processing, such as learnable matched filters or adaptive beamforming operations, could bridge the gap between traditional signal processing and end-to-end learning. Additionally, the incorporation of physical constraints, such as wave propagation models and scattering theory, into the network architecture could improve both performance and interpretability.

9.4.3 Expansion to Diverse Datasets and Configurations

The evaluation of radar detection algorithms across varied datasets, object classes, and sensor configurations is essential for establishing robust and generalizable solutions. While the CRUW dataset provides a valuable benchmark, it represents only a subset of the diverse conditions encountered in real world deployment. Future research should systematically evaluate performance across multiple datasets, including CARRADA, RADDet, and emerging benchmarks that capture different geographic regions, weather conditions, and traffic patterns.

The exploration of different radar configurations presents both challenges and opportunities. Variations in operating frequency, bandwidth, antenna design, and mounting position create distinct data characteristics that may require architectural adaptations. The development of domain adaptation techniques specifically tailored to radar data could enable models trained on one configuration to transfer effectively to others. This capability would be particularly valuable for automotive manufacturers who must deploy perception systems across diverse vehicle platforms with different sensor configurations.

The expansion to novel object classes beyond traditional road users represents another important direction. The detection of construction equipment, emergency vehicles, animals and debris requires learning distinctive radar signatures that may not be well represented in current datasets. The development of few shot learning approaches for

radar could enable rapid adaptation to new object categories with limited training examples. Additionally, the investigation of continuous learning frameworks would allow deployed systems to adapt to new object types and environmental conditions over time.

9.4.4 Multi-Modal Fusion Architectures

While this thesis focuses on radar only perception, the integration of radar with other sensing modalities remains a critical area for future research. The complementary strengths of radar, camera, and LiDAR sensors suggest that optimal perception systems will leverage multiple modalities [8]. Future work should investigate fusion architectures that preserve the unique advantages of each sensor while creating unified scene representations.

The development of asymmetric fusion strategies that account for the different reliability of each sensor under various conditions could improve robustness. For instance, the fusion weight assigned to radar should increase in adverse weather conditions where optical sensors degrade. The exploration of attention based fusion mechanisms that dynamically select relevant information from each modality based on the current context represents a promising direction. Additionally, the investigation of cross modal learning, where one sensor helps train another, could address the label scarcity problem for radar data.

9.4.5 Temporal Modeling Enhancements

The critical importance of temporal information for radar based detection, as demonstrated by our results, motivates further research into sophisticated temporal modeling approaches. The exploration of recurrent architectures specifically designed for radar sequences could capture longer term dependencies while maintaining computational efficiency. The investigation of predictive models that anticipate future object positions based on radar motion patterns could improve tracking through occlusions and enable earlier collision warnings.

The development of adaptive temporal processing that adjusts the temporal window based on scene dynamics represents an interesting direction. Static scenes may require minimal temporal context, while complex dynamic scenarios benefit from extended sequences. This adaptive approach could optimize the tradeoff between computational cost and detection performance in real time. Additionally, the exploration of hierarchical temporal models that capture motion patterns at multiple time scales could improve the detection of objects with complex motion signatures, such as pedestrians with irregular gait patterns.

9.4.6 Advanced Activation Function Research

The success of AQR opens several promising avenues for further activation function innovation in radar and broader signal processing applications. The exploration of learned activation functions that could automatically discover optimal functional forms for specific radar applications represents a natural next step. Neural architecture search techniques could be adapted to explore the space of possible activation function designs, potentially discovering more sophisticated gating mechanisms or entirely novel mathematical formulations that better capture radar signal characteristics.

The investigation of frequency-specific activation characteristics could further enhance performance for different radar bands and modulation schemes. Automotive radar systems operate across various frequency bands (24 GHz, 77 GHz, and emerging bands), each with distinct propagation characteristics and noise profiles. Developing activation functions that can adapt to these frequency-dependent characteristics could improve cross-platform compatibility and performance.

The extension of adaptive activation concepts to other sensor modalities presents another compelling direction. LiDAR point clouds, with their sparse three-dimensional structure, share certain characteristics with radar data that might benefit from similar adaptive processing approaches. The development of activation functions specifically designed for sparse, structured sensor data could advance perception capabilities across multiple automotive sensing modalities.

Research into physics-informed activation functions that explicitly incorporate radar signal propagation models could yield both improved performance and enhanced interpretability. By embedding knowledge of radar physics directly into the activation function design, such approaches could provide more robust processing while maintaining computational efficiency. The exploration of multi-scale activation functions that can simultaneously handle both strong target returns and weak clutter signals could address one of the fundamental challenges in radar signal processing.

The investigation of temporal activation functions that adapt their behavior based on motion characteristics represents another frontier. For automotive applications, where object motion patterns are constrained by physical dynamics, activation functions that can exploit these temporal correlations could improve both detection performance and tracking stability.

9.5 Concluding Remarks

This thesis has demonstrated that efficient and effective radar based object detection is not only possible but can achieve state-of-the-art performance with dramatically reduced computational requirements compared to existing approaches. The success of CompactRADNet validates our hypothesis that careful architectural design, informed by understanding of both sensor characteristics and deployment constraints, can unlock the full potential of automotive radar for robust perception across all road users.

The journey from traditional signal processing approaches to sophisticated neural architectures represents a fundamental transformation in how we conceptualize radar perception. Our work contributes to this evolution by showing that the path forward lies not in simply scaling up model complexity but in developing efficient architectures that exploit the unique characteristics of radar data. The surprising effectiveness of relatively simple components, such as pooling based temporal modeling and sensor aware representations, challenges assumptions about the necessity of complex attention mechanisms for achieving high performance.

The implications of this research extend beyond the specific technical contributions to influence the broader trajectory of autonomous driving development. By enabling robust perception with minimal computational resources, our work helps democratize advanced driver assistance capabilities, making them accessible across a wider range of vehicles and price points. The improved detection of vulnerable road users directly contributes to the safety mission that motivates autonomous driving research, potentially preventing accidents and saving lives.

Looking forward, the field of radar perception stands at an exciting crossroad. The foundations laid by this thesis and related work have demonstrated the viability of learning based approaches for radar, but numerous opportunities remain for fundamental advances. The exploration of end-to-end learning from raw radar signals, the development of radar specific architectural innovations, and the integration with other sensing modalities all represent fertile ground for future research. As autonomous driving systems continue to mature, the role of radar as a reliable, weather robust sensing modality will only grow in importance.

The ultimate goal of achieving fully autonomous driving that operates safely in all conditions remains a grand challenge that will require continued innovation across multiple disciplines. This thesis contributes one piece to this larger puzzle by advancing the state of radar based perception. The efficient and effective detection capabilities demonstrated by CompactRADNet bring us closer to the vision of autonomous vehicles that can navigate safely and reliably through the complex and dynamic environments of the real world, regardless of weather conditions or lighting scenarios.

Through rigorous experimentation, systematic analysis, and innovative design, this work has pushed the boundaries of what is achievable in radar based object detection. The insights gained from our ablation studies provide valuable guidance for future researchers, while the practical efficiency of our solution enables immediate deployment in real world systems. As the field continues to evolve, we hope that the contributions

presented in this thesis will serve as both a foundation for future research and a catalyst for new innovations that further advance the capabilities of autonomous driving systems.

The development of AQR illustrates the broader theme of this thesis: that domain-specific innovations, grounded in deep understanding of sensor physics and signal characteristics, can yield substantial improvements over general-purpose solutions. The dual validation of AQR, through both architectural integration experiments showing 0.86% AP improvement and fundamental signal processing analysis demonstrating 2.59 dB SNR enhancement, establishes a comprehensive validation methodology for activation function innovation. This approach demonstrates that meaningful advances in perception systems can emerge from innovations at the most fundamental level of neural network design, not just macro-level architectural patterns. The success of AQR encourages continued exploration of specialized neural network components tailored to the unique requirements of automotive perception, establishing a foundation for future research in domain-specific deep learning innovations.

The convergence of sensor technology, machine learning, and automotive engineering creates unprecedented opportunities for improving transportation safety and efficiency. Radar based perception, with its unique advantages and evolving capabilities, will play an essential role in realizing this potential. The work presented in this thesis represents a step forward in this journey, demonstrating that the seemingly competing goals of high performance and computational efficiency can be simultaneously achieved through thoughtful design and systematic optimization.

Summary

This chapter summarized the key contributions and findings of this thesis while outlining future research directions. The summary of contributions reiterated the four primary aspects of the advancements proposed in this thesis: the novel hybrid architecture

through the MetaFormer temporal stem achieving state-of-the-art performance with competitive computational efficiency (1.65M parameters, 15.85 GFLOPS) and improved vulnerable road user detection; sensor-aware design encompassing the MultiClass Detection Loss as the primary contribution and elliptical Gaussian heatmaps as a supporting innovation; the Adaptive Quadratic ReLU (AQR) activation function demonstrating domain-specific optimization for radar signal processing; and end-to-end real-world deployment and validation on an instrumented vehicle platform. The implications for autonomous driving were discussed, including the potential for distributed perception architectures enabled by CompactRADNet's efficiency, and the safety-critical implications of improved vulnerable road user detection. Limitations were acknowledged, including evaluation scope, computational scaling with sequence length, and the need for cross-configuration validation of AQR parameters. Future research directions were discussed across six trajectories: optimal MetaFormer design exploration, direct processing of ADC data, expansion to diverse datasets and configurations, multi-modal fusion architectures, temporal modeling enhancements, and advanced activation function research. The concluding remarks emphasized that domain-specific innovations grounded in deep understanding of sensor physics can yield significant improvements over general-purpose solutions, propelling the state of radar-based perception for autonomous driving applications.

Bibliography

[1] Geiger, A., Lenz, P., & Urtasun, R. (2012). Are we ready for autonomous driving? The KITTI vision benchmark suite. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 3354-3361).

[2] Wang, Z., Wu, Y., & Niu, Q. (2020). Multi-sensor fusion in automated driving: A survey. IEEE Access, 8, 2847-2868.

[3] Bijelic, M., et al. (2020). Seeing through fog without seeing fog: Deep multimodal sensor fusion in unseen adverse weather. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 11679-11689).

[4] Bilik, I., Longman, O., Villeval, S., & Tabrikian, J. (2019). The rise of radar for autonomous vehicles: Signal processing solutions and future research directions. IEEE Signal Processing Magazine, 36(5), 20-31.

[5] Dickmann, J., et al. (2016). Automotive radar the key technology for autonomous driving: From detection and ranging to environmental understanding. In IEEE Radar Conference (pp. 1-6).

[6] Jin, F., Sengupta, A., & Cao, S. (2020). mmWave radar point cloud segmentation using GMM in multimodal traffic monitoring. In IEEE International Radar Conference (pp. 732-737).

[7] Engels, F., et al. (2017). Advances in automotive radar: A framework on computationally efficient high-resolution frequency estimation. IEEE Signal Processing Magazine, 34(2), 36-46.

[8] Feng, D., et al. (2021). Deep multi-modal object detection and semantic

segmentation for autonomous driving: Datasets, methods, and challenges. IEEE Transactions on Intelligent Transportation Systems, 22(3), 1341-1360.

[9] Major, B., et al. (2019). Vehicle detection with automotive radar using deep learning on range-azimuth-doppler tensors. In Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (pp. 924-932).

[10] Schumann, O., et al. (2018). Semantic segmentation on radar point clouds. In International Conference on Information Fusion (pp. 2179-2186).

[11] Bai, J., et al. (2021). Radar Transformer: An object classification network based on 4D MMW imaging radar. Sensors, 21(11), 3854.

[12] Palffy, A., et al. (2020). CNN based road user detection using the 3D radar cube. IEEE Robotics and Automation Letters, 5(2), 1263-1270.

[13] Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep Learning. MIT Press.

[14] Haykin, S. (2009). Neural Networks and Learning Machines (3rd ed.). Pearson.

[15] Bishop, C. M. (2006). Pattern Recognition and Machine Learning. Springer.

[16] Nair, V., & Hinton, G. E. (2010). Rectified linear units improve restricted Boltzmann machines. In Proceedings of the 27th International Conference on Machine Learning (pp. 807-814).

[17] Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. Nature, 323(6088), 533-536.

[18] Hornik, K., et al. (1989). Multilayer feedforward networks are universal approximators. Neural Networks, 2(5), 359-366.

[19] Bottou, L. (2010). Large-scale machine learning with stochastic gradient descent. In Proceedings of COMPSTAT'2010 (pp. 177-186).

[20] LeCun, Y., et al. (1998). Efficient backprop. In Neural Networks: Tricks of the Trade (pp. 9-48). Springer.

[21] Schmidhuber, J. (2015). Deep learning in neural networks: An overview. Neural Networks, 61, 85-117.

[22] LeCun, Y., et al. (1989). Backpropagation applied to handwritten zip code recognition. Neural Computation, 1(4), 541-551.

[23] Krizhevsky, A., et al. (2012). ImageNet classification with deep convolutional neural networks. In Advances in Neural Information Processing Systems (pp. 1097-1105).

[24] Simonyan, K., & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. In International Conference on Learning Representations.

[25] Szegedy, C., et al. (2015). Going deeper with convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 1-9).

[26] He, K., et al. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 770-778).

[27] Zhao, Z. Q., et al. (2018). Object detection with deep learning: A review. IEEE Transactions on Neural Networks and Learning Systems, 30(11), 3212-3232.

[28] Girshick, R., et al. (2014). Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 580-587).

[29] Girshick, R. (2015). Fast R-CNN. In Proceedings of the IEEE International Conference on Computer Vision (pp. 1440-1448).

[30] Ren, S., et al. (2015). Faster R-CNN: Towards real-time object detection with region proposal networks. In Advances in Neural Information Processing Systems (pp. 91-99).

[31] Redmon, J., et al. (2016). You only look once: Unified, real-time object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 779-788).

[32] Liu, W., et al. (2016). SSD: Single shot multibox detector. In European Conference on Computer Vision (pp. 21-37).

[33] Long, J., et al. (2015). Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 3431-3440).

[34] Ronneberger, O., et al. (2015). U-Net: Convolutional networks for biomedical image segmentation. In International Conference on Medical Image Computing and Computer-Assisted Intervention (pp. 234-241).

[35] He, K., et al. (2017). Mask R-CNN. In Proceedings of the IEEE International Conference on Computer Vision (pp. 2961-2969).

[36] Vaswani, A., et al. (2017). Attention is all you need. In Advances in Neural Information Processing Systems (pp. 6000-6010).

[37] Dosovitskiy, A., et al. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. In International Conference on Learning Representations.

[38] Carion, N., et al. (2020). End-to-end object detection with transformers. In European Conference on Computer Vision (pp. 213-229).

[39] Scheiner, N., et al. (2020). Seeing around street corners: Non-line-of-sight detection and tracking in-the-wild using doppler radar. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 2065-2074).

[40] Nowruzi, F. E., et al. (2020). Deep open space segmentation using automotive radar. In IEEE MTT-S International Conference on Microwaves for Intelligent Mobility (pp. 1-4).

[41] Qi, C. R., et al. (2017). PointNet: Deep learning on point sets for 3D classification and segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 77-85).

[42] Qi, C. R., et al. (2017). PointNet++: Deep hierarchical feature learning on point sets in a metric space. In Advances in Neural Information Processing Systems (pp. 5099-5108).

[43] Wang, Y., et al. (2019). Dynamic graph CNN for learning on point clouds. ACM Transactions on Graphics, 38(5), 1-12.

[44] Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural Computation, 9(8), 1735-1780.

[45] Cho, K., et al. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (pp. 1724-1734).

[46] Kim, S., Lee, S., Doo, S., & Shim, B. (2018). Moving target classification in automotive radar systems using convolutional recurrent neural networks. In 2018 26th European Signal Processing Conference (EUSIPCO) (pp. 1482-1486).

[47] Dong, X., et al. (2020). Probabilistic oriented object detection in automotive radar. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (pp. 458-467).

[48] Jiang, T., et al. (2023). T-RODNet: Transformer for vehicular millimeter-wave radar object detection. IEEE Transactions on Instrumentation and Measurement, 72, 1-12.

[49] Sun, P., et al. (2021). RSN: Range sparse net for efficient, accurate LiDAR 3D object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 5721-5730).

[50] Wu, Y., et al. (2024). Mask-RadarNet: Enhancing transformer with spatial-temporal semantic context for radar object detection in autonomous driving. arXiv preprint arXiv:2412.15595.

[51] Wang, Y., et al. (2021). RODNet: A real-time radar object detection network cross-supervised by camera-radar fused object 3D localization. IEEE Journal of Selected Topics in Signal Processing, 15(4), 954-967.

[52] Cheng, L., & Cao, S. (2025). TransRAD: Retentive vision transformer for enhanced radar object detection. IEEE Transactions on Radar Systems,3,303-317.

[53] Rebut, J., et al. (2022). Raw high-definition radar for multi-task learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 17000-17009).

[54] Zhou, Y., et al. (2022). Towards deep radar perception for autonomous driving: Datasets, methods, and challenges. Sensors, 22(11), 4208.

[55] Zou, Z., et al. (2023). Object detection in 20 years: A survey. Proceedings of the IEEE, 111(3), 257-276.

[56] Szeliski, R. (2010). Computer Vision: Algorithms and Applications. Springer.

[57] Viola, P., & Jones, M. (2001). Rapid object detection using a boosted cascade of simple features. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 511-518).

[58] Lienhart, R., & Maydt, J. (2002). An extended set of Haar-like features for rapid object detection. In Proceedings of the International Conference on Image Processing (pp. 900-903).

[59] Dalal, N., & Triggs, B. (2005). Histograms of oriented gradients for human detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 886-893).

[60] Dollar, P., et al. (2012). Pedestrian detection: An evaluation of the state of the art. IEEE Transactions on Pattern Analysis and Machine Intelligence, 34(4), 743-761.

[61] Felzenszwalb, P. F., et al. (2010). Object detection with discriminatively trained part-based models. IEEE Transactions on Pattern Analysis and Machine Intelligence, 32(9), 1627-1645.

[62] Everingham, M., Van Gool, L., Williams, C. K., Winn, J., & Zisserman, A. (2010). The PASCAL Visual Object Classes (VOC) challenge. International Journal of Computer Vision, 88, 303-338.

[63] Benenson, R., et al. (2014). Ten years of pedestrian detection, what have we learned? In European Conference on Computer Vision (pp. 613-627).

[64] Russakovsky, O., et al. (2015). ImageNet Large Scale Visual Recognition Challenge. International Journal of Computer Vision, 115(3), 211-252.

[65] Huang, J., et al. (2017). Speed/accuracy trade-offs for modern convolutional

object detectors. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 3296-3297).

[66] Zeiler, M. D., & Fergus, R. (2014). Visualizing and understanding convolutional networks. In European Conference on Computer Vision (pp. 818-833).

[67] Lin, T. Y., et al. (2017). Focal loss for dense object detection. In Proceedings of the IEEE International Conference on Computer Vision (pp. 2980-2988).

[68] Uijlings, J. R., et al. (2013). Selective search for object recognition. International Journal of Computer Vision, 104(2), 154-171.

[69] Zhang, S., et al. (2018). Single-shot refinement neural network for object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 4203-4212).

[70] Dai, J., et al. (2016). R-FCN: Object detection via region-based fully convolutional networks. In Advances in Neural Information Processing Systems (pp. 379-387).

[71] Fu, C. Y., et al. (2017). DSSD: Deconvolutional single shot detector. arXiv preprint arXiv:1701.06659.

[72] Liu, Z., et al. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 10012-10022).

[73] Law, H., & Deng, J. (2018). CornerNet: Detecting objects as paired keypoints. In Proceedings of the European Conference on Computer Vision (pp. 765-781).

[74] Lin, T. Y., et al. (2017). Feature pyramid networks for object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 2117-2125).

[75] Hu, J., et al. (2018). Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 7132-7141).

[76] Wu, Y., et al. (2020). Rethinking classification and localization for object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 10183-10192).

[77] Cai, Z., & Vasconcelos, N. (2018). Cascade R-CNN: Delving into high quality object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 6154-6162).

[78] Sun, P., et al. (2021). Sparse R-CNN: End-to-end object detection with learnable proposals. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 14449-14458).

[79] Redmon, J., & Farhadi, A. (2017). YOLO9000: Better, faster, stronger. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 6517-6525).

[80] Redmon, J., & Farhadi, A. (2018). YOLOv3: An incremental improvement. arXiv preprint arXiv:1804.02767.

[81] Bochkovskiy, A., et al. (2020). YOLOv4: Optimal speed and accuracy of object detection. arXiv preprint arXiv:2004.10934.

[82] Ge, Z., et al. (2021). YOLOX: Exceeding YOLO series in 2021. arXiv preprint arXiv:2107.08430.

[83] Li, Z., & Zhou, F. (2017). FSSD: Feature fusion single shot multibox detector. arXiv preprint arXiv:1712.00960.

[84] Sandler, M., et al. (2018). MobileNetV2: Inverted residuals and linear bottlenecks. In Proceedings of the IEEE Conference on Computer Vision and

Pattern Recognition (pp. 4510-4520).

[85] Zhu, X., et al. (2020). Deformable DETR: Deformable transformers for end-to-end object detection. In International Conference on Learning Representations.

[86] Ghiasi, G., et al. (2019). NAS-FPN: Learning scalable feature pyramid architecture for object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 7029-7038).

[87] Zhang, S., et al. (2020). Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 9756-9765).

[88] Kong, T., et al. (2020). FoveaBox: Beyound anchor-based object detection. IEEE Transactions on Image Processing, 29, 7389-7398.

[89] Li, Y., et al. (2019). Scale-aware trident networks for object detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 6053-6062).

[90] Rezatofighi, H., et al. (2019). Generalized intersection over union: A metric and a loss for bounding box regression. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 658-666).

[91] Kuhn, H. W. (1955). The Hungarian method for the assignment problem. Naval Research Logistics Quarterly, 2(1-2), 83-97.

[92] Zheng, Z., et al. (2020). Distance-IoU loss: Faster and better learning for bounding box regression. In Proceedings of the AAAI Conference on Artificial Intelligence, 34, (pp. 12993-13000).

[93] Chen, K., et al. (2019). MMDetection: Open mmlab detection toolbox and

benchmark. arXiv preprint arXiv:1906.07155.

[94] Yu, J., et al. (2016). UnitBox: An advanced object detection network. In Proceedings of the 24th ACM International Conference on Multimedia (pp. 516-520).

[95] Zheng, Z., et al. (2022). Enhancing geometric factors in model learning and inference for object detection and instance segmentation. IEEE Transactions on Cybernetics, 52(8), 8574-8586.

[96] Pang, J., et al. (2019). Libra R-CNN: Towards balanced learning for object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 821-830).

[97] Tian, Z., et al. (2019). FCOS: Fully convolutional one-stage object detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 9626-9635).

[98] Zhou, X., et al. (2019). Objects as points. arXiv preprint arXiv:1904.07850.

[99] Cao, Y., et al. (2020). Prime sample attention in object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 11580-11588).

[100] He, Y., et al. (2019). Bounding box regression with uncertainty for accurate object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 2883-2892).

[101] Qian, R., et al. (2022). 3D object detection for autonomous driving: A survey. Pattern Recognition, 130, 108796.

[102] Patole, S. M., et al. (2017). Automotive radars: A review of signal processing techniques. IEEE Signal Processing Magazine, 34(2), 22-35.

[103] Iovescu, C., & Rao, S. (2017). The fundamentals of millimeter wave sensors. Texas Instruments, 1-8.

[104] Richards, M. A., et al. (2010). Principles of Modern Radar: Basic Principles. SciTech Publishing.

[105] Richards, M. A. (2022). Fundamentals of Radar Signal Processing. McGraw-Hill Education.

[106] Rohling, H. (1983). Radar CFAR thresholding in clutter and multiple target situations. IEEE Transactions on Aerospace and Electronic Systems, AES-19(4), 608-621.

[107] Stove, A. G. (1992). Linear FMCW radar techniques. IEE Proceedings F (Radar and Signal Processing), 139(5), 343-350.

[108] Hyun, E., et al. (2016). A pedestrian detection scheme using a coherent phase difference method based on 2D range-Doppler FMCW radar. Sensors, 16(1), 124.

[109] Hasch, J., et al. (2012). Millimeter-wave technology for automotive radar sensors in the 77 GHz frequency band. IEEE Transactions on Microwave Theory and Techniques, 60(3), 845-860.

[110] Winkler, V. (2007). Range Doppler detection for automotive FMCW radars. In European Microwave Conference (pp. 166-169).

[111] Karnfelt, C., et al. (2009). 77 GHz ACC radar simulation platform. In 9th International Conference on Intelligent Transport Systems Telecommunications (pp. 209-214).

[112] Dudek, M., et al. (2015). System analysis of a phased-array radar applying adaptive beam-control for future automotive safety applications. IEEE

Transactions on Vehicular Technology, 64(1), 34-47.

[113] Sun, H., et al. (2014). Analysis and comparison of MIMO radar waveforms. In International Radar Conference (pp. 1-6).

[114] Li, J., & Stoica, P. (2007). MIMO radar with colocated antennas. IEEE Signal Processing Magazine, 24(5), 106-114.

[115] Rohling, H., & Meinecke, M. M. (2001). Waveform design principles for automotive radar systems. In CIE International Conference on Radar (pp. 1-4).

[116] Kronauge, M., & Rohling, H. (2013). Fast two-dimensional CFAR procedure. IEEE Transactions on Aerospace and Electronic Systems, 49(3), 1817-1823.

[117] Cao, P., et al. (2018). Radar-ID: Human identification based on radar micro-Doppler signatures using deep convolutional neural networks. IET Radar, Sonar & Navigation, 12(7), 729-734.

[118] Schumann, O., et al. (2017). Comparison of random forest and long short-term memory network performances in classification tasks using radar. In 2017 Sensor Data Fusion: Trends, Solutions, Applications (pp. 1-6).

[119] Prophet, R., et al. (2018). Pedestrian classification with a 79 GHz automotive radar sensor. In 19th International Radar Symposium (pp. 1-6).

[120] Capobianco, S., et al. (2018). Vehicle classification based on convolutional networks applied to FMCW radar signals. In Traffic Mining Applied to Police Activities, Advances in Intelligent Systems and Computing, 728, (pp. 115-128)

[121] Ouaknine, A., et al. (2020). CARRADA dataset: Camera and automotive radar with range-angle-doppler annotations. In 2020 25th International Conference on Pattern Recognition (ICPR) (pp. 5068-5075).

[122] Palffy, A., et al. (2022). Multi-class road user detection with 3+1D radar in the view-of-delft dataset. In IEEE Robotics and Automation Letters, 7(2), 4961-4968.

[123] Zhang, A., et al. (2021). RADDet: Range-azimuth-doppler based radar object detection for dynamic road users. In 2021 18th Conference on Robots and Vision (CRV) (pp. 95-102).

[124] Dickmann, J., et al. (2015). Making bertha see even more: Radar contribution. IEEE Access, 3, 1233-1247.

[125] Kellner, D., et al. (2016). Tracking of extended objects with high-resolution Doppler radar. IEEE Transactions on Intelligent Transportation Systems, 17(5), 1341-1353.

[126] Lekic, V., & Babic, Z. (2019). Automotive radar and camera fusion using generative adversarial networks. Computer Vision and Image Understanding, 184, 1-8.

[127] Roos, F., et al. (2017). Ghost target identification by analysis of the Doppler distribution in automotive scenarios. In International Radar Symposium (pp. 1-9).

[128] Chadwick, S., et al. (2019). Distant vehicle detection using radar and vision. In International Conference on Robotics and Automation (pp. 8311-8317).

[129] Werber, K., et al. (2015). Automotive radar gridmap representations. In IEEE MTT-S International Conference on Microwaves for Intelligent Mobility (pp. 1-4).

[130] Mostajabi, M., et al. (2020). High-resolution radar dataset for semi-supervised learning of dynamic objects. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (pp. 100-101).

[131] Engels, F., et al. (2017). Automotive Radar Signal Processing: Research Directions and Practical Challenges. IEEE Journal of Selected Topics in Signal Processing ,15(4), 1-4.

[132] Roos, F., et al. (2016). Enhancement of Doppler resolution for chirp-sequence modulated automotive radars. In European Radar Conference (pp. 237-240).

[133] Smith, L. N., & Topin, N. (2019). Super-convergence: Very fast training of neural networks using large learning rates. In Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications, 11006, (p. 1100612).

[134] Xu, W., et al. (2024). E-RODNet: Lightweight approach to object detection by vehicular millimeter-wave radar. IEEE Sensors Journal, 24(20), 33091-33100.

[135] Shi, W., et al. (2016). Is the deconvolution layer the same as a convolutional layer? arXiv preprint arXiv:1609.07009.

[136] Yu, W., et al. (2022). MetaFormer is actually what you need for vision. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 10809-10819).

[137] Frigo, M., & Johnson, S. G. (2005). The design and implementation of FFTW3. Proceedings of the IEEE, 93(2), 216-231.