

A Deep Weighted CFAR Detector for Automotive Radar

Shujian Jia

Delft University of Technology

A Deep Weighted CFAR Detector for Automotive Radar

by

Shujian Jia

Supervisor: Dr F. Fioranelli
Daily Supervisor: Dr S. Zhu

Abstract

Automotive radar is a key all-weather sensor for autonomous driving, providing range and radial-velocity measurements under adverse illumination and weather conditions. Target detection and subsequent point-cloud generation are therefore critical stages in automotive radar signal processing. However, conventional constant false alarm rate (CFAR) detectors rely on fixed local reference rules and statistical assumptions that are often violated by extended targets and non-homogeneous backgrounds in road scenes, resulting in sparse and spatially inaccurate point clouds. End-to-end deep-learning detectors can exploit richer radar context, but may require substantial computational resources and provide limited interpretability and post-training control of the detection threshold.

In this work, we propose a hybrid neural CFAR-like detector in which a three-level U-Net estimates a background-reliability map from radar data, and an explicit local thresholding stage uses this map to form cell-wise detection decisions. Evaluated experimentally on the *RaDelft* dataset, the proposed detector achieved the lowest Chamfer Distance (CD) among the evaluated methods, reaching 1.36 m compared with 3.19 m for CA-CFAR, 3.29 m for OS-CFAR, and 1.60 m for a deep-learning baseline taken from the literature. Moreover, the proposed method used 0.87 million parameters and achieved 10.58 GFLOPs, which is 13 times less than the evaluated deep-learning baseline. These results indicate that learning the reliability of local background samples can improve automotive radar point-cloud quality while retaining an explicit and adjustable detection procedure.

Contents

Abstract	i
Nomenclature	iv
1 Introduction	1
1.1 Background and Motivation	1
1.2 Thesis Objectives	3
1.3 Thesis Outline	3
2 Literature Review	4
2.1 Model-Based Radar Target Detection	4
2.2 Learning-Based Radar Target Detection	6
2.3 Comparative Analysis and Research Gap	9
2.4 Summary	10
3 Methodology	11
3.1 Signal Model and Preprocessing	11
3.1.1 FMCW range and Doppler estimation	11
3.1.2 MIMO radar and virtual array	13
3.1.3 Radar signal-processing pipeline	15
3.2 Baseline Approaches	17
3.2.1 CFAR	17
3.2.2 Deep Learning Baseline	21
3.3 Proposed Method	23
3.4 Summary	27
4 Results and Analysis	28
4.1 RaDelft Dataset	28
4.2 Performance Metrics	29
4.3 Baseline Results	30
4.3.1 CFAR Baselines	32
4.3.2 Deep Learning Baseline	36
4.3.3 Summary of Baseline Results	38
4.4 Proposed Method Results	38
4.4.1 Backbone analysis	43
4.4.2 Ablation Study	43
4.4.3 Leave One Scene Out experiment	44
4.5 Summary	45
5 Conclusion and Future Work	47
5.1 Conclusion	47
5.2 Future Work	49

Contents	iii
5.2.1 Alternative Directions Explored During This Study	50
References	52

Nomenclature

List of abbreviations

Abbreviation	Definition
P_d	Probability of Detection
P_{fa}	Probability of False Alarm
<i>ADAS</i>	Advanced Driver Assisted System
<i>CD</i>	Chamfer Distance
<i>CUT</i>	Cell Under Test
<i>CFAR</i>	Constant False Alarm Rate
<i>RCS</i>	Radar Cross Section
<i>ATR</i>	Automatic Target Recognition
<i>ADC</i>	Analog to Digital Converter
<i>SNR</i>	Signal to Noise Ratio
<i>IID</i>	Identically Independent Distributed
<i>DL</i>	Deep Learning
<i>TDMA</i>	Time Division Multiple Access
<i>CNN</i>	Convolutional Neural Network
<i>ROC</i>	Receiver Operating Characteristic
<i>DOA</i>	Direction of Arrival
<i>FFT</i>	Fast Fourier Transform

List of Figures

3.1	The block diagram of the radar signal processing steps.	16
3.2	Structure of a CFAR detector. The cell under test (CUT) is compared with a threshold derived from the local background estimate. Guard cells reduce the influence of target energy on the background estimate, while reference cells are used to estimate the local background level [3].	17
3.3	Theoretical distributions of background-only and target-plus-background power. The power of a background-only signal follows an exponential distribution, whereas the target-plus-background power follows a non-central chi-squared distribution.	20
3.4	Empirical histograms of background-only and target-plus-background power derived from the RaDelft dataset. The distributions overlap, indicating that a single, fixed power threshold may not separate them effectively.	20
3.5	Architecture of the deep learning baseline. Three consecutive radar cubes are processed by a shared Doppler encoder and FPN/ResNet-18 backbone. Their features are then combined by a temporal-coherence network to predict one occupancy grid per frame. Adapted from Roldan et al. [6].	22
3.6	Pipeline of the proposed hybrid detector. Max and argmax operations compress the Doppler dimension into energy and Doppler-index maps. A U-Net predicts a background-reliability map, which suppresses probable target cells before local background estimation and cell-wise thresholding.	23
3.7	U-Net used to estimate the background-reliability map. The encoder extracts features at three spatial scales, and the decoder uses bilinear interpolation and skip connections to recover the input resolution. A one-channel convolution and sigmoid activation produce the final map. The four blocks at the bottom of the figure shows the operations represented by the respective coloured arrow.	25
3.8	This figure from the left to right shows the example of original radar energy map, network output $Y_{r,a}$ and weighted background map B^w	26
4.1	Antenna array topology for the radar used to collect the <i>RaDelft</i> dataset.	29
4.2	Overall point-cloud comparison. In the top left corner is the camera image, and in the top right corner is the radar energy map in XY view. The second row from the left to the right shows the point clouds generated by CA-CFAR, the proposed method, and Roldan et al. [6] respectively, as well as the LiDAR ground truth.	31
4.3	Point-cloud comparison of four CFAR configurations for an example of data frame. CFAR point clouds are shown in the second row.	33
4.4	Example point clouds generated by CA-CFAR under the configuration $T = 5$, $G = 5$, and $\alpha = 1.5$. The frames are selected according to their CD values. Good frames are selected based on the lowest CD values, while bad frames are selected based on the highest CD values.	34

4.5	Example point clouds generated by OS-CFAR under the configuration $T = 5$, $G = 5$, and $\alpha = 1.5$. The frames are selected according to their CD values.	35
4.6	Bar plot of CD for all test frames using CA-CFAR with $T = G = 5$ and $\alpha = 1.5$	36
4.7	CD bar plot for all test frames using the deep learning baseline proposed by Roldan et al. [6].	36
4.8	Point clouds of good and bad frames generated by the deep learning baseline. The first row is the camera and point cloud generated by network in XY view. The second row from left to right is the LiDAR point cloud and the True Positives, False Positives and False Negatives points of network.	37
4.9	Point clouds of good and bad frames generated by the proposed method. In the first row is the camera and radar energy map in XY view. In the bottom row from left to right are network output, final generated point cloud and LiDAR point cloud.	39
4.10	ROC curve comparison between CA-CFAR and the proposed method on test Scenes 2 and 6. The curve shows method performance in terms of P_d and P_{fa} under different threshold factors. The factors are swept across a range of values from 1.0 to 4.0. Curves closer to the upper-left corner indicate better performance.	40
4.11	Bar plot of CD for the proposed method on the test scenes, with 4080 frames in total.	40
4.12	Violin plot of the proposed method versus CA-CFAR. The IQR is the CD range from the 25th to the 75th percentile.	41
4.13	Violin plot of the proposed method versus the deep learning baseline. The IQR is the CD range from the 25th to the 75th percentile.	42
4.14	Validation loss and training loss for different U-Net depths, whose performance is analyzed in Table 4.4	44
4.15	Leave-one-scene-out CD bar plot. The bars show the best- α CD values for each leave-one-scene-out configuration, with the corresponding testing scenes indicated. Each value is the average across the corresponding test scenes; red indicates a value above the overall average, whereas yellow indicates a value below it.	45

List of Tables

2.1	A summary of the discussed learning-based methods.	8
4.1	Overall results for the baselines and the proposed method. The values are mean values over all the samples of the test dataset. The CA-CFAR and OS-CFAR and proposed method results here are shown with the best configurations, threshold factor and number of training cells and guardcells.	31
4.2	CFAR results under different configurations. T denotes the number of training cells and G denotes the number of guard cells. Each result is reported using the threshold factor that achieved the best CD.	32
4.3	Results for different U-Net variants using the default number of layers reported in the original papers [54], [55], [56]. A different layer configuration is used in the final results.	43
4.4	Results for different U-Net depths. Level denotes the number of feature scales in the U-Net. Here the 3-level version is used for the final implementation as it achieves the best CD.	43
4.5	Comparison of soft and hard targets with focal loss.	44
4.6	Best- α leave-one-scene-out results. Each row reports the operating point with the lowest CD for that split.	45

Introduction

1.1. Background and Motivation

A reliable perception system is essential for autonomous driving because it provides information about the surrounding environment, including other vehicles, pedestrians, and static objects. This information supports subsequent tasks such as localisation and mapping, prediction, planning, and control.

To operate reliably, a perception system must remain robust under varying illumination, adverse weather, and complex traffic conditions. Cameras, LiDAR, and radar are among the sensors most commonly used for this purpose. Radar is generally more robust than cameras and LiDAR under adverse weather and illumination conditions, such as rain, fog, and low light or glaring [1]. Moreover, cameras cannot directly measure distance or velocity, while conventional LiDAR measures distance but not velocity. Radar can measure both range and radial velocity through Doppler information, which is particularly valuable for distinguishing moving targets from stationary objects and for target tracking [2]. Reliable radar signal processing is therefore an important component of an effective automotive perception system.

Radar target detection is a fundamental step in the radar signal-processing pipeline. For each resolution cell, it determines whether target echoes are present in measurements that also contain noise, clutter, and interference. Following analogue-to-digital conversion and the subsequent signal-processing stages of range and Doppler/velocity estimation, target detection can be used to generate a radar point cloud for downstream tasks such as object detection¹ and tracking.

In general, radar target detection can be formulated as a binary hypothesis test comparing for each resolution cell the case without target (H_0) with the case of target mixed with non-target contributions (H_1) [3], [4]:

$$\begin{cases} H_0 : y = n, \\ H_1 : y = s + n, \end{cases} \quad (1.1)$$

¹The term 'detection' mentioned here should be distinguished from the detection task addressed in this thesis. In traditional radar target detection, detection refers to determining whether a target is present or if only noise (or clutter) exists within each resolution cell. In contrast, object detection, typically employed in computer vision, involves simultaneously localizing objects with bounding boxes and classifying them.

where y denotes the received signal, s denotes the target echo, and n represents the combined non-target contributions.

Classical radar detectors, including constant false alarm rate (CFAR) detectors, are based on statistical models of the received signal. These methods estimate the local background and derive a threshold for deciding between the two hypotheses. Their performance, however, can be limited in automotive scenarios. Complex road environments contain rapidly varying clutter, dense targets, and target-masking effects. Moreover, conventional CFAR configurations use fixed hyperparameters, such as numbers of guard and reference cells. Such fixed windows cannot readily accommodate targets whose apparent sizes vary with range, angle, and time, particularly when extended targets occupy multiple resolution cells [5], [6]. This is actually the case in mm-wave radar for automotive.

Deep learning provides an alternative means of extracting complex features and learning the spatial structure of radar data. Its applications in radar include radar sensing, signal processing, and automatic target recognition [7], [8], [9]. Radar target detection belongs to the signal-processing stage and provides information required by downstream recognition tasks. Several deep learning-based radar detectors have achieved better performance than traditional methods under the conditions considered in their respective studies [10]. Unlike fixed-window detectors based on predefined statistical rules, learned detectors can model (more) complex spatial patterns in the input data.

Nevertheless, deep learning-based detectors also have important limitations. Their performance depends strongly on the quantity, quality, and coverage of the training data, and their behaviour may be unpredictable for inputs outside the training distribution. Radar annotations are also commonly transferred from LiDAR or cameras, which may introduce inaccuracies because the sensors operate according to different physical principles [7], and there can be issues of residual calibration errors for the heading and field of view of the different sensors. In addition, end-to-end models can be computationally expensive and difficult to interpret or adjust after training [11]. These limitations are especially relevant to safety-critical automotive systems and resource-constrained edge devices.

The availability of suitable datasets presents a further challenge. Large-scale, real-world automotive radar datasets are scarce, particularly for target-detection studies that require pre-detection radar measurements, such as analogue-to-digital converter data or radar cubes, together with detection-level ground truth [7], [12]. Some existing datasets are not publicly accessible, and many studies therefore rely on simulated data [7], [13]. This reliance makes it difficult to assess reliability in real-world scenarios and to compare detectors under consistent experimental conditions [10].

Existing studies often treat statistical detectors, such as CFAR, and deep learning-based detectors as separate approaches, although their characteristics are complementary. CFAR provides a structured and interpretable mechanism for local background estimation and threshold-based detection that is meant to guarantee an expected level of false alarms. Deep learning, by contrast, can extract complex features that cannot be represented adequately by a fixed-size window. Motivated by this complementarity, ***this thesis investigates a hybrid framework in which a neural network derives background-reliability weights from radar data, while a CFAR-like procedure performs local background estimation and threshold-based detection.***

1.2. Thesis Objectives

The overall aim of this thesis is to investigate whether learned, data-dependent background weighting can address limitations of fixed-window CFAR detection while retaining an explicit local thresholding stage. To achieve this aim, the thesis pursues the following objectives:

1. To establish representative CFAR and deep learning-based detectors as baselines for automotive radar target detection under a common experimental setting.
2. To develop a hybrid detector in which a U-Net-like network estimates a background-reliability weight map from three-dimensional radar cubes, and then a CFAR-like procedure uses the weighted background to derive cell-wise detection thresholds.
3. To evaluate the proposed detector against the two baselines on the *RaDelft* dataset [14] and to analyse the contribution of its principal design components through ablation experiments.

Together, these objectives examine not only whether the proposed detector performs effectively, but also whether its learned weighting and explicit thresholding stages provide a useful combination for real-world automotive radar detection.

1.3. Thesis Outline

The remainder of this thesis is organised as follows. Chapter 2 reviews the literature and approaches for automotive radar target detection, and identifies the limitations that motivate the proposed method. Chapter 3 describes the baseline detectors in detail and presents the architecture and detection procedure of the proposed method. Chapter 4 introduces the *RaDelft* dataset, evaluation metrics, experimental settings, and final results. Finally, Chapter 5 summarises the main findings, discusses the limitations of the study, and suggests directions for future work.

2

Literature Review

This chapter reviews methods that convert pre-detection radar measurements into target detections or radar point clouds. The scope is limited to the detection stage operating on representations such as range profiles, range–Doppler maps, and multidimensional radar tensors. Methods that take an already generated radar point cloud as input for object classification, tracking, or sensor fusion are therefore outside the main scope. Existing detection methods are organised into two categories: model-based and learning-based methods. Model-based methods derive decision rules from explicit statistical descriptions of the measurements, whereas learning-based methods learn detection patterns from labelled data. This distinction exposes the complementary strengths and limitations that motivate the method developed in this thesis.

2.1. Model-Based Radar Target Detection

Model-based radar detectors formulate target detection as a statistical decision problem. A target is declared when the response in a cell under test is inconsistent with a background model estimated from neighbouring measurements. Constant false alarm rate (CFAR) processing is the most widely used realisation of this principle because it adapts the detection threshold to the local “interference” level (i.e., the combination of thermal noise, clutter, and actual interference) rather than applying one global threshold [3], [15]. Its practical appeal follows from three properties: the decision process is interpretable, the design false-alarm probability can be specified explicitly under the assumed model, and the detector can be implemented efficiently with a sliding reference window. These properties have made CFAR a standard point-generation stage in automotive radar processing [6], [15], even though its classical statistical assumptions can be violated in non-homogeneous scenes [16].

The principal CFAR variants differ in how they estimate the background from the reference window. Cell-averaging CFAR (CA-CFAR) combines all reference cells and uses their mean as the background estimate [17]. This method relies on two fundamental assumptions: the interference samples in the reference cells and the Cell-Under-Test (CUT) are independent and identically distributed (i.i.d.), and the reference window is free of interfering targets. Environments satisfying these conditions are termed “homogeneous environments”. Gandhi and Kassam [16] demonstrated that CA-CFAR provides baseline performance in homogeneous backgrounds. For a specified false-alarm probability, it requires the minimum additional signal-to-

noise ratio (SNR) to achieve the same detection probability as an optimal detector with perfect knowledge of the background power. However, these conditions are readily violated in practice. Target echoes may contribute energy to the reference window through nearby targets or extended targets occupying multiple cells, which are common in automotive radar [5], [6]. This contamination biases the background estimate upwards and raises the detection threshold, potentially suppressing the target. This phenomenon is known as target masking. Performance also degrades when the interference level changes abruptly at a clutter edge, which in automotive scenario could appear as the boundary of the free road and the edge of the building. An inflated estimate can cause missed detections on the low-interference side, while an underestimated background can cause excessive false alarms on the high-interference side. Greatest-of and smallest-of CFAR modify the averaging strategy by selecting between estimates from the two sides of the CUT, rather than using all samples. Greatest-of CFAR is more conservative at clutter transitions but more susceptible to target masking. Smallest-of CFAR can improve performance in multiple-target conditions but is more susceptible to clutter edges [18], [19]. These methods show that the composition of the reference set, rather than threshold scaling alone, affects detection in non-homogeneous environments.

The methods discussed above exhibit a trade-off between robustness to clutter edges and resistance to target masking. Ordered-statistic CFAR (OS-CFAR) addresses this trade-off more directly by sorting the reference samples and estimating the background from a selected order statistic. The detector can therefore tolerate a specified number of high-energy interfering cells without allowing them to dominate the estimate [20], [21]. If fewer than $N - k$ high-power samples contaminate the reference set, they are placed above $x_{(k)}$ and do not directly determine the selected value. Under the assumed step-clutter model, choosing $k > N/2$ also allows the estimate to track the higher clutter level once most reference cells enter that region [16]. OS-CFAR can therefore be more robust than mean-level processing under suitable heterogeneous conditions. This robustness comes with additional computational cost because sorting typically requires $O(N \log N)$ operations, compared with $O(N)$ for mean-level summation. Moreover, performance depends on the selected rank and the number of contaminated cells. A fixed order statistic cannot guarantee robustness when the fraction or spatial extent of interfering responses varies. This limitation is particularly relevant to automotive targets, which may occupy several cells and change in apparent extent with range, aspect angle, and radar resolution. A detector configured for one target extent may exclude useful reference cells in one scene, while retaining target-contaminated cells in another.

Further model-based work has sought greater robustness by censoring suspected outliers or replacing the arithmetic mean with less contamination-sensitive statistics. Censored-mean and related multiple-target detectors exclude selected high-energy reference cells before estimating the background [22], [23]. More recently, censored harmonic averaging has been used to reduce the influence of outliers without imposing the same rigid rejection rule as conventional trimmed estimators. By combining censoring with a harmonic mean of the retained samples, this approach further limits the influence of strong interfering targets. Zefreh et al. reported improved robustness under heterogeneous clutter and interfering-target conditions [24]. Other approaches modify the standard CFAR model more substantially. For example, Safa et al. mapped the input and threshold into a high-dimensional feature space and applied a kernel-based decision rule. Their nonlinear, low-complexity detector outperformed OS-CFAR under the indoor multipath conditions examined in that study [25]. This result shows that the

limitations of local averaging are not unique to road scenes and can be addressed through model design without learning. However, the method was applied only to range profiles from indoor scenes and did not exploit the additional dimensions available from millimetre-wave radar. Its effectiveness for multidimensional automotive radar therefore remains uncertain. Evidence from indoor sensing alone cannot establish performance in automotive settings, where target geometry, clutter statistics, measurement representations, and operational constraints differ [15], [26].

Automotive-specific clutter processing provides additional evidence that a single homogeneous-background assumption is often inadequate. Yoon et al. showed that dense roadside clutter can be recognised and suppressed by exploiting structure in FMCW radar measurements rather than treating every high-energy response as an isolated target [26]. Nevertheless, most model-based detectors still require the designer to carefully choose the reference-window geometry, guard region, background statistic, and decision parameters in advance. These choices encode useful physical knowledge and permit efficient inference, but they also limit the detector's ability to adapt when the spatial support of a target or the composition of the background changes. The central limitation is therefore not simply that CFAR uses a threshold; rather, it is that the reliability of the samples used to construct that threshold is determined by a fixed, rigid rule.

2.2. Learning-Based Radar Target Detection

Learning-based detectors replace some or all of the predefined statistical model with a mapping learned from data. Neural networks can process a wider spatial context than a local CFAR window and can jointly exploit intensity, Doppler, angular, and temporal information. This capability is attractive for automotive radar because target responses are spatially extended and the distinction between target, clutter, and multipath cannot always be expressed by a single amplitude distribution.

DL-CFAR learns to locate target-component peaks in a range–Doppler input and clips them so that the residual provides a less contaminated estimate of the noise background [27]. This design directly addresses the masking mechanism that affects CA-CFAR in multiple-target scenes. It also illustrates how learning can modify background estimation without discarding threshold-based detection. However, the method was evaluated using simulated range–Doppler data rather than real-world measurements. Akhtar et al. proposed a broader optimised hybrid architecture that combines features derived from radar echo representations and tunes the network using a population-based optimisation procedure [28]. Although the reported results indicate that learned combinations of time- and frequency-domain information can support range–Doppler detection, the study is based primarily on simulated automotive FMCW data. Its performance therefore demonstrates feasibility rather than robustness in real traffic environments.

Other methods learn the detection decision more directly. Temporal convolutional processing has been applied to multiple radar dwells so that weak but persistent targets can be distinguished from non-homogeneous clutter [8]. By integrating evidence over time, these models use information unavailable to a frame-wise CFAR detector, but the improvement is accompanied by additional latency and sensitivity to temporal alignment. Spatial networks similarly exploit multidimensional radar structure. Lin et al. processed range–azimuth–Doppler cubes with a shared feature extractor and multiple detection heads, demonstrating the value of rea-

soning jointly across radar dimensions rather than applying independent local tests [9]. Deep fully convolutional detection has also been investigated for extended radar targets, for which the output must represent target support rather than only an isolated peak [29]. These studies broaden the detection objective, but their outputs may be cell labels, target extents, or bounding regions rather than the point clouds expected by an automotive perception stack.

Point-cloud generation methods align the learning objective more closely with the output of an automotive radar front end. Cheng et al. proposed RPDNet, which learns radar detections using LiDAR-derived supervision and then applies spatiotemporal filtering to reduce spurious points [30]. The method showed that improving the density and reliability of generated radar points can benefit downstream detection and localisation, although its evaluation includes mobile-robot environments rather than being restricted to road vehicles. DenserRadar extends the same general idea to high-resolution 4D automotive radar. It constructs dense occupancy supervision by accumulating LiDAR frames and predicts point clouds directly from radar tensors; on K-Radar, the method was evaluated against CA-CFAR, OS-CFAR, and a learning-based point-cloud detector at comparable output densities [31]. This comparison is informative because it treats density and accuracy as coupled quantities: increasing the number of retained cells is useful only if the additional points correspond to meaningful scene structure.

Roldan et al. provide the setting closest to the present thesis. Their detector uses three consecutive 4D range–azimuth–Doppler–elevation cubes and synchronised LiDAR supervision from the real-world RaDelft dataset to generate a radar-only point cloud [6]. The study compared learned and CFAR-based detectors using detection probability, false-alarm probability, and Chamfer distance. Its results showed that Doppler and temporal information contributed to representing extended targets. The method also exposes two broader limitations of end-to-end learning. First, the continuous network output must be converted into a binary point set using a decision threshold. A threshold chosen during development does not provide the same statistical interpretation or inference-time adjustability as a model-based CFAR parameter. Second, LiDAR supervision is not a neutral description of radar ground truth. LiDAR and radar respond to different physical properties and observe the scene from different sensor positions. Spatially aligned LiDAR points can therefore supervise object occupancy but cannot label every radar-specific return without ambiguity.

Recent work has begun to address this cross-modal supervision problem. Schwarz et al. augmented LiDAR-derived point supervision with a downstream perception-task loss so that a radar detector would not be rewarded solely for reproducing the LiDAR point distribution [32]. This direction recognises that a useful radar point cloud should preserve information relevant to perception, including radar-specific velocity and reflection cues, rather than merely resemble a sparse version of LiDAR. At the same time, task-aware supervision makes the learned output dependent on the selected downstream model and annotation scheme. The literature therefore shows a progression from learning a cleaner background, through direct cell classification, to generating task-oriented point sets. Across this progression, greater representational flexibility is obtained at the cost of stronger dependence on training data, labels, computational resources, and post-training calibration.

A summary of above mentioned method is in Table 2.1

Methods	Main advantage	Main disadvantage
DL-CFAR [27]	Use deep learning to deal with masking effect	Use only range–Doppler data as input. Verified only on simulated data.
OHDL [28]	Design a hybrid deep learning model to effectively use the information from time and frequency domain	Simulated data and no angle estimation considered.
Deep temporal detection [8]	Use temporal information to enhance detection performance	More latency involved.
TriHeadDet [9]	Decompose 3D data to 2D by a three head network with shared features to learn effectively	High complexity, need to be verified by richer data.
Blackman et. al [29]	Use Fully-convolutional network to specifically deal with the extended targets	Simulated data for Wide-area surveillance (WAS) radar.
RPDNet [30]	Processing LiDAR points cloud for supervision of the network, combining segmentation network with GAN network to improve the performance	High complexity, the angle information is not included.
DenserRadar [31]	Directly extract feature from 4D radar cube using 3D convolution network. Cross-attention mechanism enhanced feature fusion.	High complexity,
Ignacio et. al [6]	Combined multi-frame features to use temporal information in 4D radar cube	High complexity.
Schwarz et. al [32]	Jointly train learn radar detection and perception to preserve radar specific characteristics	High complexity.

Table 2.1: A summary of the discussed learning-based methods.

2.3. Comparative Analysis and Research Gap

Model-based and learning-based detectors make different commitments about where detection knowledge is stored. In model-based methods, the assumptions are explicit in the background distribution, reference statistic, window geometry, and false-alarm design. These choices support transparent operation and efficient implementation, but they restrict adaptation to extended targets and heterogeneous road scenes. In learning-based methods, the assumptions are encoded implicitly by the architecture, training distribution, supervision, and loss function. Such detectors can use spatial and temporal context to identify patterns beyond the reach of a fixed local window, but their behaviour outside the training distribution is less predictable and their output threshold is not necessarily associated with a controlled false-alarm probability. Neither approach is therefore uniformly preferable; their relative performance depends on the scene, radar representation, supervision, and operating point.

Comparison across the published literature is further complicated by inconsistent evaluation objectives. Classical detection studies typically report probability of detection, probability of false alarm, or receiver-operating-characteristic curves [8], [21]. Learning-based point-generation studies additionally use point density, point accuracy, Chamfer distance, intersection over union, or downstream average precision [6], [30], [31]. These metrics answer different questions. Cell-level detection metrics measure the ability to separate labelled target and background cells, whereas geometric metrics assess whether the resulting point set represents the spatial support of an object. Downstream metrics additionally depend on the perception model receiving the point cloud. Reported gains are consequently not directly comparable when the methods use different output definitions, radar resolutions, datasets, or numbers of generated points.

The available evidence also reveals a tension between point-cloud completeness and false detections. A conservative CFAR operating point can suppress clutter but produce a sparse representation of extended objects. Relaxing the threshold increases point density while admitting additional background and multipath responses. Learning-based methods can shift this trade-off by using contextual evidence, as demonstrated by RPDNet, DenserRadar, and the RaDelft detector [6], [30], [31]; however, their apparent advantage depends on the completeness and alignment of the LiDAR-derived labels. A prediction that disagrees with LiDAR may be a false radar detection, a radar-specific response absent from LiDAR, or a cross-sensor calibration error. This ambiguity limits claims that a learned detector has recovered a uniquely correct radar point cloud.

These observations identify a narrower gap than simply using more deep learning. CA-CFAR and OS-CFAR determine the contribution of reference cells through fixed averaging or ranking rules. End-to-end neural detectors learn richer contextual decisions, but ordinarily replace the explicit local background estimate with a score map followed by a fixed output threshold. Comparatively less attention has been directed towards learning the reliability of the reference data while retaining an explicit, locally adaptive thresholding stage. Such a hybrid formulation could use spatial context to reduce target contamination in the background estimate while preserving a CFAR-like decision mechanism that remains interpretable and adjustable at inference time. Whether this combination improves point-cloud generation on real automotive radar data, without simply transferring the limitations of LiDAR supervision into the detector, is the research question addressed by this thesis.

2.4. Summary

Model-based detectors provide an efficient and interpretable route from radar measurements to detection points, but their performance depends on fixed assumptions about the background and the suitability of the reference window. Robust CFAR variants reduce sensitivity to clutter transitions and interfering targets, yet they cannot fully accommodate target extents and background structures that vary across automotive scenes. Learning-based detectors use spatial, Doppler, and temporal information to generate denser and more structured outputs, but introduce dependence on training coverage, cross-modal supervision, computation, and output calibration.

The two families are therefore complementary rather than mutually exclusive. The literature motivates a detector in which learning is used to estimate which measurements provide reliable evidence of the local background, while an explicit CFAR-like stage converts that estimate into an adaptive detection threshold. The following chapter describes the model-based and learning-based baselines used to evaluate this idea, and then presents the hybrid method developed in this thesis.

3

Methodology

This chapter first introduces the automotive-radar signal model and the preprocessing steps used to construct the radar cube. It then describes the three baselines used for comparison: CA-CFAR, OS-CFAR, and a deep learning detector from the literature. Finally, the chapter presents the proposed hybrid method, which combines learned background estimation with a CFAR-like detection procedure.

3.1. Signal Model and Preprocessing

This section establishes the signal model used to describe the radar data in this thesis. Automotive radars commonly operate at millimetre-wave frequencies and employ frequency-modulated continuous-wave (FMCW) signals because they support joint range and radial-velocity estimation with a substantially lower sampling rate than the transmitted radio-frequency bandwidth [2], [3]. The derivation below first considers a single-input single-output channel and then extends the model to a multiple-input multiple-output (MIMO) array. Unless stated otherwise, the model assumes point targets, constant radial velocity during one coherent processing interval (CPI), narrow-band propagation, and far-field observations.

3.1.1. FMCW range and Doppler estimation

An FMCW radar transmits a sequence of linear-frequency-modulated chirps. In the rising sawtooth modulation commonly used by automotive radar, the instantaneous frequency increases linearly during each chirp and then returns to its initial value before the next chirp begins. A reflected chirp is delayed by the target's propagation time and shifted in frequency by its radial motion via the Doppler effect. Comparing this echo with the currently transmitted chirp therefore provides the information required to estimate both range and radial velocity. Let f_c be the start frequency, B the swept bandwidth, T_c the chirp duration, and $\kappa = B/T_c$ the chirp slope. The complex transmitted signal of one chirp is:

$$x(t) = \exp \left[j2\pi \left(f_c t + \frac{\kappa}{2} t^2 \right) \right], \quad 0 \leq t < T_c. \quad (3.1)$$

The quadratic phase in Eq. (3.1) produces the desired linear increase in instantaneous frequency. Consider first a stationary point target at range R . Because the signal travels from the radar

to the target and back, its round-trip delay is $\tau = 2R/c$. Apart from propagation loss and the target reflection coefficient, represented by the complex amplitude α , the received echo is a delayed copy of the transmitted chirp, $r(t) = \alpha x(t - \tau)$. The receiver then performs dechirping by multiplying this echo by the conjugate of the current transmitted signal:

$$\begin{aligned} z(t) &= r(t)x^*(t) \\ &= \tilde{\alpha} \exp(-j2\pi\kappa\tau t), \end{aligned} \quad (3.2)$$

where all terms that are constant with respect to fast time are absorbed into $\tilde{\alpha}$. The remaining complex-valued exponential is in real terms a sinusoid whose frequency is proportional to the propagation delay. This frequency is referred to as the range beat frequency and satisfies:

$$f_r = \kappa\tau = \frac{2\kappa R}{c}, \quad R = \frac{cf_r}{2\kappa}. \quad (3.3)$$

Thus, a larger range produces a longer delay and, because the transmitted frequency changes continuously, a larger difference between the transmitted and received frequencies. This relationship converts the delay measurement into a frequency-estimation problem. Importantly, the beat frequency is far below the millimetre-wave carrier frequency, allowing the range information to be digitally sampled and processed at a substantially lower rate [3].

The finite chirp duration limits the ability to distinguish closely spaced beat frequencies. Over an observation interval T_c , two sinusoidal components require a frequency separation of approximately $1/T_c$ to be resolved. Mapping this frequency separation through Eq. (3.3), and using $B = \kappa T_c$, gives the theoretical range resolution

$$\Delta R = \frac{c}{2\kappa T_c} = \frac{c}{2B}. \quad (3.4)$$

Whereas the range resolution is controlled by the swept bandwidth, the practical measurement range is mainly controlled by the receiver bandwidth. After dechirping, an analogue low-pass filter suppresses high-frequency mixing products and limits the beat signal presented to the ADC. If its cut-off frequency is f_{LPF} , the largest range-dependent beat frequency that can pass through the receiver determines the practical maximum range. Neglecting Doppler at this stage gives:

$$R_{\text{max}} = \frac{cf_{\text{LPF}}}{2\kappa}. \quad (3.5)$$

This practical limit should be distinguished from the unambiguous range imposed by discrete-time sampling. After ADC sampling at frequency f_s , beat frequencies separated by an integer multiple of f_s produce identical sample sequences because

$$e^{-j2\pi(f_r + kf_s)n/f_s} = e^{-j2\pi f_r n/f_s}, \quad k \in \mathbb{Z}. \quad (3.6)$$

The sampled data can therefore identify the beat frequency only within one interval of width f_s . Under the positive beat-frequency convention used here, mapping this interval to range gives:

$$R_u = \frac{cf_s}{2\kappa}. \quad (3.7)$$

The receiver filter is normally designed with $f_{\text{LPF}} < f_s$ and may occupy only a limited fraction of the available sampling interval. Consequently, the usable range R_{max} is smaller than

the sampling-defined unambiguous range R_u , and it can be substantially smaller when the analogue passband is narrow. In practice, echoes outside $R_{\max}$ are attenuated by the receiver before sampling, even though the sampling process would in principle permit a larger unambiguous interval.

The stationary-target derivation attributes the entire beat frequency to propagation delay. For a moving target, however, the Doppler effect introduces an additional frequency shift. If v denotes radial velocity and λ the carrier wavelength, the Doppler frequency is [3]:

$$f_D = \frac{2v}{\lambda}, \quad \lambda = \frac{c}{f_c}. \quad (3.8)$$

For a rising chirp and the sign convention adopted here, this term adds to the range beat frequency, so the measured beat frequency becomes:

$$f_b = f_r + f_D = \frac{2\kappa R}{c} + \frac{2v}{\lambda}. \quad (3.9)$$

Equation (3.9) contains two unknowns, R and v , and a single fast-time beat frequency cannot separate them. FMCW radar resolves this ambiguity by transmitting repeated chirps. Although the target range changes only slightly between adjacent chirps, its Doppler frequency produces a systematic phase progression in slow time. With the dechirping convention used above, observations of the same transmit channel separated by T_{slow} have the phase difference $\Delta\phi = -2\pi f_D T_{\text{slow}}$. The Doppler frequency and radial velocity can therefore be estimated as:

$$\hat{f}_D = -\frac{\Delta\phi}{2\pi T_{\text{slow}}}, \quad \hat{v} = -\frac{\lambda \Delta\phi}{4\pi T_{\text{slow}}}. \quad (3.10)$$

Because phase is observed modulo 2π , phase increments outside $[-\pi, \pi)$ are indistinguishable from wrapped values. Applying this condition to the slow-time phase progression gives the maximum unambiguous radial velocity [3]

$$|v| < v_{\max} = \frac{\lambda}{4T_{\text{slow}}}. \quad (3.11)$$

The estimated Doppler frequency can finally be subtracted from the measured beat frequency. The remaining component is associated with propagation delay alone, yielding the Doppler-corrected range:

$$\hat{R} = \frac{c}{2\kappa} (f_b - \hat{f}_D). \quad (3.12)$$

3.1.2. MIMO radar and virtual array

Range and Doppler can be estimated from a single antenna channel, but angle estimation requires a spatial comparison. The basic principle is that a wave arriving from an oblique direction does not reach all antenna elements at exactly the same time. Consider a uniform linear receive array with N_R elements separated by distance d , and let θ denote the direction of arrival (DoA) measured from broadside. Under the far-field assumption, the wavefront is approximately planar. Relative to the first element, the additional path length and arrival-time difference at element r are:

$$\Delta l_r = rd \sin \theta, \quad \Delta \tau_r = \frac{rd \sin \theta}{c}, \quad r = 0, \dots, N_R - 1. \quad (3.13)$$

These time differences are extremely small for automotive-radar array dimensions and are not usually measured directly. Instead, they appear as phase differences at the carrier frequency. Since $\lambda = c/f_c$, the phase of element r relative to the first element is:

$$\Delta\phi_r = -2\pi f_c \Delta\tau_r = -\frac{2\pi r d}{\lambda} \sin \theta. \quad (3.14)$$

The DoA is therefore recovered by finding the angle whose predicted phase progression best matches the phases measured across the array. Collecting these predicted responses gives the steering vector

$$\mathbf{a}(\theta) = \left[1, e^{-j\frac{2\pi d}{\lambda} \sin \theta}, \dots, e^{-j\frac{2\pi(N_R-1)d}{\lambda} \sin \theta} \right]^T. \quad (3.15)$$

Conventional digital beamforming evaluates this steering vector over a grid of candidate angles and identifies directions that produce coherent addition across the antenna channels.

The ability to distinguish two nearby directions is governed primarily by the array aperture. For the receive array above, the aperture is $D = (N_R - 1)d$, and the angular resolution scales approximately as [3], [33]:

$$\Delta\theta \approx \frac{\lambda}{D \cos \theta}. \quad (3.16)$$

Hence, improving angular resolution requires a larger effective aperture. One direct approach is to add physical receive antennas or increase their span. However, every additional receive channel requires radio-frequency hardware, calibration, ADC resources, and data transfer, so a large physical array rapidly increases system cost and complexity.

MIMO radar provides a more hardware-efficient way to enlarge the effective aperture [33]. It transmits distinguishable (i.e., ideally orthogonal) signals from N_T transmit antennas and records each signal at N_R receive antennas. Let p_t and q_r denote the one-dimensional positions of transmit antenna t and receive antenna r , respectively, relative to a common array reference, where $t = 0, \dots, N_T - 1$ and $r = 0, \dots, N_R - 1$. In a colocated far-field MIMO radar, the spatial phase of the transmit–receive pair depends on the sum of these positions. Defining $u_{t,r}$ as the position of the corresponding virtual array element gives:

$$u_{t,r} = p_t + q_r, \quad [\mathbf{a}_V(\theta)]_{t,r} = \exp\left(-j\frac{2\pi u_{t,r}}{\lambda} \sin \theta\right). \quad (3.17)$$

If all position sums are distinct, N_T transmitters and N_R receivers generate up to $N_T N_R$ virtual antenna channels. Appropriate placement of the physical antennas can spread these virtual positions over a substantially larger aperture than the receive array alone. According to Eq. (3.16), this larger virtual aperture improves angular resolution while using fewer physical receive chains than an equivalent conventional array.

The transmitted signals must be separated before the virtual channels can be formed. The radar used in this thesis employs time-division multiple access (TDMA), in which the transmit antennas are activated sequentially in different chirps [6], [34], [35]. This sequential transmission incurs two motion-related costs. First, when the chirp repetition interval for the overall radar schedule is fixed, a given transmitter is revisited only once every N_T chirps. The effective slow-time interval for that transmitter is therefore increased by a factor of N_T , which reduces

the unambiguous Doppler interval and the corresponding maximum unambiguous radial velocity by the same factor:

$$v_{\max, \text{TDMA}} = \frac{\lambda}{4N_{\text{T}}T_{\text{PRI}}} = \frac{c}{4f_{\text{c}}N_{\text{T}}T_{\text{PRI}}}, \quad (3.18)$$

where T_{PRI} is the interval between consecutive transmitted chirps. Second, a moving target changes range between transmissions from different antennas. For transmit events separated by Δt , this motion introduces an additional virtual-array phase term

$$\phi_{\text{mig}} = \frac{4\pi v \Delta t}{\lambda}. \quad (3.19)$$

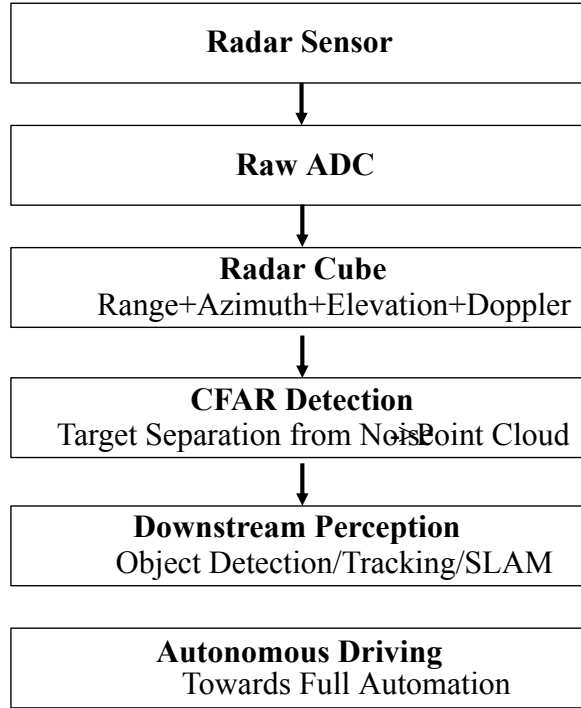
Because this term depends on target velocity as well as direction of arrival, it can bias angle estimation and produce artefacts unless Doppler-dependent motion compensation is applied before beamforming [33], [34], [35], [36]. After the samples are assigned to their transmit–receive pairs, range is estimated along fast time, Doppler along slow time, and angle across the virtual antenna channels. These processing steps produce the range–Doppler–azimuth radar cube used by the detection methods in the remainder of this chapter.

3.1.3. Radar signal-processing pipeline

The signal-processing steps are shown in Figure 3.1. After dechirping and analogue-to-digital conversion, the complex baseband samples are organised along three acquisition dimensions: fast time, slow time, and antenna channel. Fast-time samples describe the beat signal within one chirp, slow-time samples describe its evolution across repeated chirps, and the channel dimension contains the measurements from the physical or virtual antenna elements. This organisation preserves the information required to estimate range, radial velocity, and angle in successive processing stages [6], [15].

The first stage converts the sampled waveforms into range–Doppler spectra. A window is normally applied before spectral estimation to reduce leakage and sidelobes. A fast Fourier transform (FFT) along fast time then maps the range-dependent beat frequencies to range bins, while a second FFT along slow time maps the inter-chirp phase progression to Doppler bins. These operations are performed for each antenna channel, producing one range–Doppler matrix per channel. When the displacement of the ego vehicle or a target is large relative to the radar resolution during the coherent processing interval, compensation may additionally be applied to reduce range or Doppler migration [15].

Angle estimation follows by comparing the complex responses of the antenna channels at each range–Doppler location. For a one-dimensional array, this stage estimates azimuth; an array with sufficient aperture in two spatial dimensions can additionally estimate elevation. Digital beamforming is frequently used because the steering-vector search can be implemented efficiently using a spatial FFT. Combining the range, Doppler, and angular estimates produces a radar cube whose cells contain the reflected signal power at discrete spatial and velocity coordinates. In computationally constrained implementations, detection may be performed before angle estimation so that beamforming is evaluated only for selected range–Doppler cells. This reduces considerably the processing load, although it also prevents coherent spatial integration from contributing to the detection SNR.

Schematic of Automotive Radar Full -Link**Figure 3.1:** The block diagram of the radar signal processing steps.

The detection stage converts the dense radar representation into a sparse set of target hypotheses. As mentioned in previous chapters, conventional processing typically combines CFAR thresholding with local-maximum or peak selection. Data-driven detectors provide an alternative by learning target structure directly from range–Doppler, range–azimuth, or range–azimuth–Doppler representations [6], [8], [9], [37], [38]. Because most radar-cube cells correspond to empty space or weak background returns, the detector output is naturally sparse. Detected cells can be represented as an occupancy grid, but they are commonly transformed to Cartesian coordinates and stored as a radar point cloud for subsequent processing, visualisation, and comparison with other sensors.

Downstream perception modules can attach task-level meaning to these detections. Semantic processing may classify or segment radar returns into road-user and infrastructure categories, while learned radar networks can jointly estimate object locations and class confidence [39], [40], [41]. Tracking then associates detections over time to smooth noisy estimates, suppress short-lived false alarms, and estimate target trajectories. This stage is particularly important for automotive radar because a vehicle or pedestrian can generate several detections rather than a single point; tracking methods must therefore account for the spatial extent and motion of the target [42], [43].

This thesis focuses on the detection stage: its input is the processed radar representation, and its output is the sparse point cloud evaluated in the following chapters.

3.2. Baseline Approaches

This thesis implements two model-driven CFAR baselines and one data-driven deep learning baseline. As previously discussed, CA-CFAR and OS-CFAR assume a statistical distribution for the input data, estimate its parameters from neighbouring cells, and derive a power threshold for a specified probability of false alarm, P_{fa} . Their principal advantages are low computational cost and interpretable operation. However, their assumptions become less appropriate when targets occupy multiple cells or when the observed background does not follow the assumed distribution, which is often the case in automotive radar. The deep learning baseline instead predicts whether each cell is occupied directly from the input data. It can capture more complex features than a fixed statistical model, but it entails greater training and computational costs and is less interpretable.

3.2.1. CFAR

Two CFAR methods are used as baselines in this thesis: cell-averaging CFAR (CA-CFAR) and ordered-statistic CFAR (OS-CFAR). CFAR is a conventional approach to radar target detection. Its core principle is to derive a detection threshold from locally estimated distribution parameters while maintaining a specified P_{fa} . The structure of a CFAR detector is shown in Fig. 3.2 [3]. For each cell under test (CUT), guard and reference cells are defined around the CUT. The guard cells reduce the influence of potential target-energy leakage on the background estimate. The reference cells are used to estimate the local background level. The CFAR window slides across the input domain and performs a binary hypothesis test at each position. The input may be one-dimensional, such as a range profile, or two-dimensional, such as a range–Doppler map.

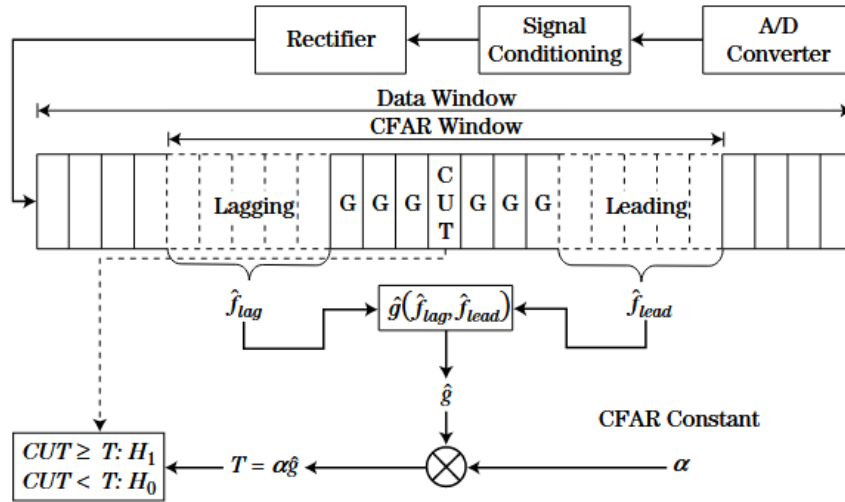


Figure 3.2: Structure of a CFAR detector. The cell under test (CUT) is compared with a threshold derived from the local background estimate. Guard cells reduce the influence of target energy on the background estimate, while reference cells are used to estimate the local background level [3].

The difference between CA-CFAR and OS-CFAR lies in how they estimate the background and compute the threshold for a specified P_{fa} . CA-CFAR estimates the background level by averaging the values of the reference cells [3], [17]. OS-CFAR instead uses a selected order statistic from those cells [3], [20]. Under the assumed background-only model, the reference-

cell power samples X_i are independent and identically distributed (i.i.d.). If the complex base-band background is circular complex Gaussian with zero mean, square-law detection yields exponentially distributed power samples [3], [4]. For CA-CFAR, a threshold corresponding to a specified P_{fa} can be derived from the mean interference power. The probability density function of the exponential distribution is:

$$p(y) = \frac{1}{\mu} \exp\left(-\frac{y}{\mu}\right), \quad y \geq 0, \quad (3.20)$$

where μ denotes the mean interference power. Integrating this distribution above the threshold T gives:

$$\begin{aligned} P_{fa} &= \int_T^\infty p(y) dy \\ &= \int_T^\infty \frac{1}{\mu} \exp\left(-\frac{y}{\mu}\right) dy \\ &= \exp\left(-\frac{T}{\mu}\right). \end{aligned} \quad (3.21)$$

The corresponding threshold is:

$$T = -\mu \ln(P_{fa}). \quad (3.22)$$

For a fixed design value of P_{fa} , the threshold T is therefore proportional to the estimated mean interference power μ .

For OS-CFAR, the reference samples are sorted in ascending order as:

$$X_{(1)} \leq X_{(2)} \leq \dots \leq X_{(N)}, \quad (3.23)$$

and the k -th order statistic $X_{(k)}$ is used as the background estimate. The detection threshold is:

$$T = \alpha_{OS} X_{(k)}, \quad (3.24)$$

where α_{OS} is the OS-CFAR scaling factor. Let the CUT sample Y follow the same exponential distribution under the null hypothesis H_0 . For a known value $X_{(k)} = x$, the conditional false-alarm probability is:

$$\begin{aligned} \Pr(Y > T \mid X_{(k)} = x) &= \Pr(Y > \alpha_{OS} x \mid X_{(k)} = x) \\ &= \exp(-\lambda \alpha_{OS} x), \end{aligned} \quad (3.25)$$

where $\lambda = 1/\mu$. Because $X_{(k)}$ is random, the overall false-alarm probability is obtained by averaging over its distribution.

$$P_{fa,OS} = \int_0^\infty \exp(-\lambda \alpha_{OS} x) f_{X_{(k)}}(x) dx. \quad (3.26)$$

For N i.i.d. samples with cumulative distribution function $F(x)$ and probability density function $f(x)$, the density of the k -th order statistic is:

$$f_{X_{(k)}}(x) = k \binom{N}{k} [F(x)]^{k-1} [1 - F(x)]^{N-k} f(x). \quad (3.27)$$

For exponentially distributed samples,

$$F(x) = 1 - \exp(-\lambda x), \quad f(x) = \lambda \exp(-\lambda x). \quad (3.28)$$

Substitution into Eq. (3.27) gives:

$$f_{X_{(k)}}(x) = k \binom{N}{k} (1 - \exp(-\lambda x))^{k-1} \exp[-\lambda x(N - k + 1)]. \quad (3.29)$$

Using this density in Eq. (3.26) yields:

$$P_{fa,OS} = \prod_{i=0}^{k-1} \frac{N - i}{N - i + \alpha_{OS}}. \quad (3.30)$$

Thus, α_{OS} can be obtained numerically for specified values of $P_{fa,OS}$, N , and k . OS-CFAR is more robust to high-valued outliers because it uses a selected order statistic instead of the mean of all reference cells. With the ascending-order convention, up to $N - k$ high-valued samples can lie above $X_{(k)}$ without directly determining the estimate. However, OS-CFAR is more computationally intensive because it requires ordering the reference samples and solving for the scaling factor [20], [21]. In practice, the choice of CFARs depends on the radar application and the characteristics of its operating environment [44], [45].

The preceding analysis relies on several assumptions that may be violated in automotive scenarios. Three assumptions are particularly relevant:

1. **Point-target assumption.** Each target in the radar field of view is assumed to occupy approximately one resolution cell.
2. **Independence assumption.** The reference-cell samples are assumed to be independent and identically distributed (i.i.d.) random variables.
3. **Distribution assumption.** The background data are assumed to follow a specified distribution from which a detection threshold can be derived.

The point-target assumption can be violated for several reasons. Automotive radar operates at relatively short and varying distances from targets, so a target's apparent size in the radar field of view changes with its position. For example, because angular resolution varies across the field of view, a target may occupy different numbers of cells at different angles. A distant target may occupy a single cell, whereas a nearby target may extend across multiple cells. Moreover, increasing radar resolution further reduces cell size and makes multi-cell target responses more likely [5], [6]. A fixed number of guard cells may therefore fail to exclude energy leakage from targets of varying sizes, leading to an inaccurate background estimate and target masking.

The independence assumption can also be violated in dense automotive scenarios. Truncated and windowed FFT processing induces side-lobes from strong targets, leading to cross-cell power leakage. These side-lobes can introduce correlations between adjacent cells, while nearby targets can contaminate several reference cells simultaneously. Consequently, a fixed window size may not provide a reliable background estimate across scenes with different target densities and spatial structures.

The distribution assumption is fundamental to the detector's ability to maintain a constant false alarm rate. Under the model adopted here, the power of a background-only signal follows

an exponential distribution. For a deterministic target return embedded in complex Gaussian background, the target-plus-background power follows a scaled non-central chi-squared distribution [3], [4], [46]. These theoretical distributions are illustrated in Fig. 3.3, and their separation depends on the signal-to-noise ratio (SNR). Figure 3.4 additionally presents histograms derived from the experimental RaDelft [14] dataset later used in this thesis. The target and non-target distributions have similar shapes and overlap substantially, indicating that a single power threshold may not separate them effectively.

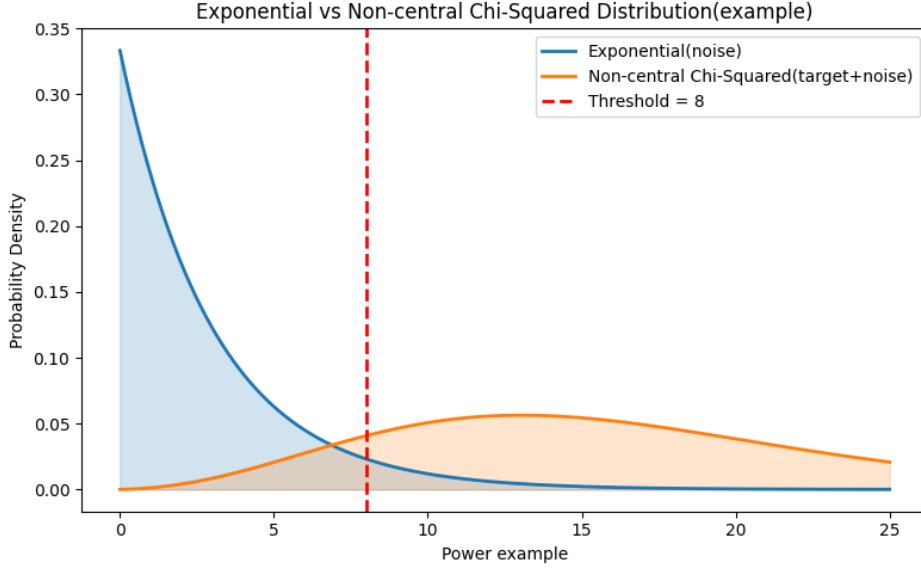


Figure 3.3: Theoretical distributions of background-only and target-plus-background power. The power of a background-only signal follows an exponential distribution, whereas the target-plus-background power follows a non-central chi-squared distribution.

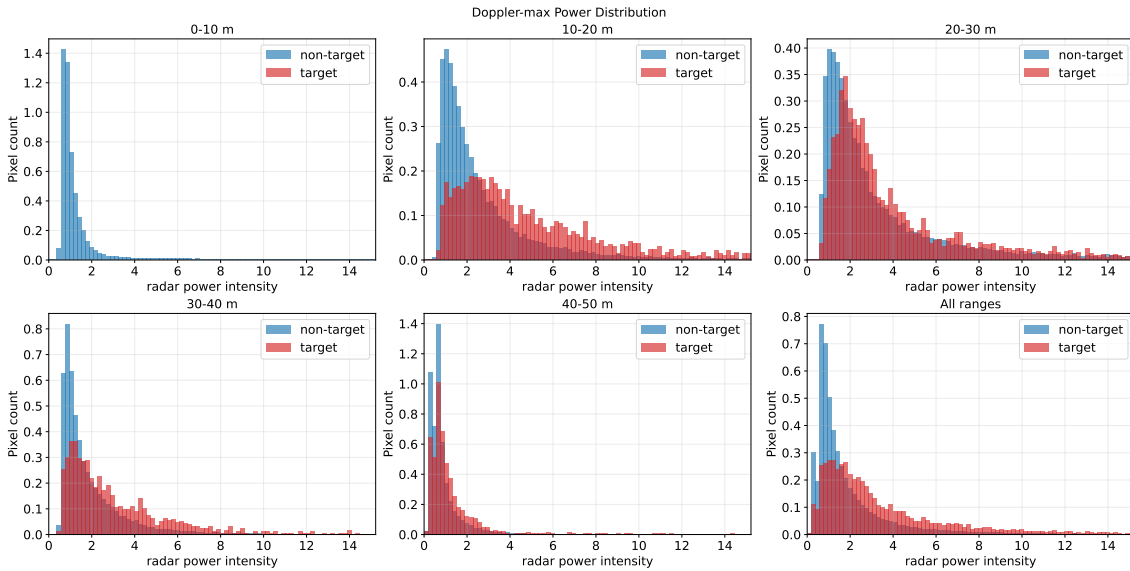


Figure 3.4: Empirical histograms of background-only and target-plus-background power derived from the RaDelft dataset. The distributions overlap, indicating that a single, fixed power threshold may not separate them effectively.

3.2.2. Deep Learning Baseline

The deep learning baseline is based on the method proposed by Roldan et al. [6]. This baseline was selected for three reasons. First, unlike CA-CFAR and OS-CFAR, it is an end-to-end data-driven detector that maps radar-cube measurements directly to an occupancy-grid prediction, providing a relevant learned alternative to the proposed hybrid design. Second, it was developed and evaluated using the publicly available RaDelft dataset, which consists of real-world synchronized automotive radar and LiDAR measurements and is therefore closely aligned with the data setting of this thesis. Third, the network is trained with LiDAR-derived supervision while using radar data alone at inference. This provides a controlled comparison because both the reproduced baseline and the proposed method can use the same data split and LiDAR-derived training targets, while differing in how the final detection decision is formed [6], [14], [47]. The original study presented several network variants, for example networks without Doppler information or without elevation information. For the configuration without elevation information, the method takes three consecutive three-dimensional radar cubes (range $\times$ azimuth $\times$ Doppler) as input and produces a detected point cloud. As shown in Fig. 3.5, each frame is first processed independently by a Doppler encoder comprising two three-dimensional convolutional layers and a three-dimensional max-pooling layer. The encoded features then pass through a feature pyramid network [48] with a ResNet-18 backbone [49]. The three branches share weights, and their outputs are concatenated and processed jointly by a temporal-coherence network comprising four three-dimensional convolutional layers. This network produces an occupancy grid for each input frame.

The architecture has two strengths that are relevant to this study. First, the Doppler encoder preserves velocity-related information before the radar cube is reduced to a representation comparable with the LiDAR-derived labels. Second, the temporal-coherence network combines three consecutive frames to capture temporal context. The ablation study by Roldan et al. confirms that both Doppler and temporal information contribute to detection performance [6]. However, the method also has three limitations in the present setting. First, its binary output is determined by a fixed decision threshold after training. A CFAR detector instead allows its detection parameters to be adjusted at inference time, although this adjustment may affect the constant-false-alarm-rate property. Second, the deep learning baseline has a relatively high computational cost. Third, thresholding discards the distinction between confidence scores above the selected operating point. For example, scores of 0.51 and 0.99 receive the same binary label when the threshold is 0.5.

In this work, this baseline was reproduced using the authors' publicly available implementation without modifying the model or training code [47]. Because pretrained weights were not released, the model was retrained for 50 epochs using the same training, validation, and test split as the proposed method. The checkpoint with the lowest validation loss was selected for evaluation. Following the original implementation, the continuous network output was converted into a binary occupancy grid using a decision threshold of 0.5. This value corresponds to the conventional midpoint of a probability-like output, but it is not statistically calibrated and cannot guarantee a specified false-alarm probability. It therefore provides a fixed operating point for the comparison rather than the inference-time adjustability offered by the proposed method.

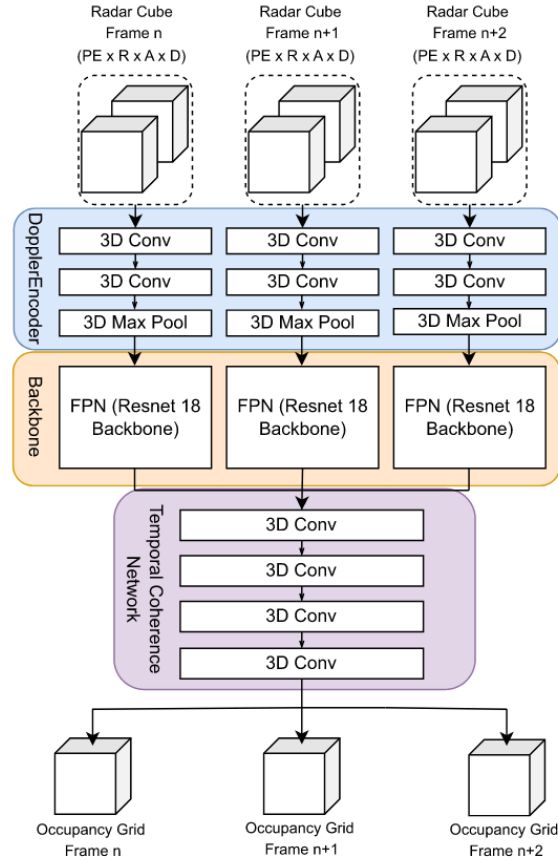


Figure 3.5: Architecture of the deep learning baseline. Three consecutive radar cubes are processed by a shared Doppler encoder and FPN/ResNet-18 backbone. Their features are then combined by a temporal-coherence network to predict one occupancy grid per frame. Adapted from Roldan et al. [6].

3.3. Proposed Method

This section presents the proposed method, which combines learned background estimation with an explicit CFAR-like detection procedure. As shown in Fig. 3.6, the network does not directly convert the radar measurements into a binary point cloud. Instead, it predicts a continuous background-reliability map that attenuates probable target responses before the local background is estimated. The final detections are obtained by comparing the original radar energy with a cell-wise threshold derived from this weighted background. This design preserves the confidence information contained in the continuous network output, while retaining an interpretable and adjustable thresholding stage.

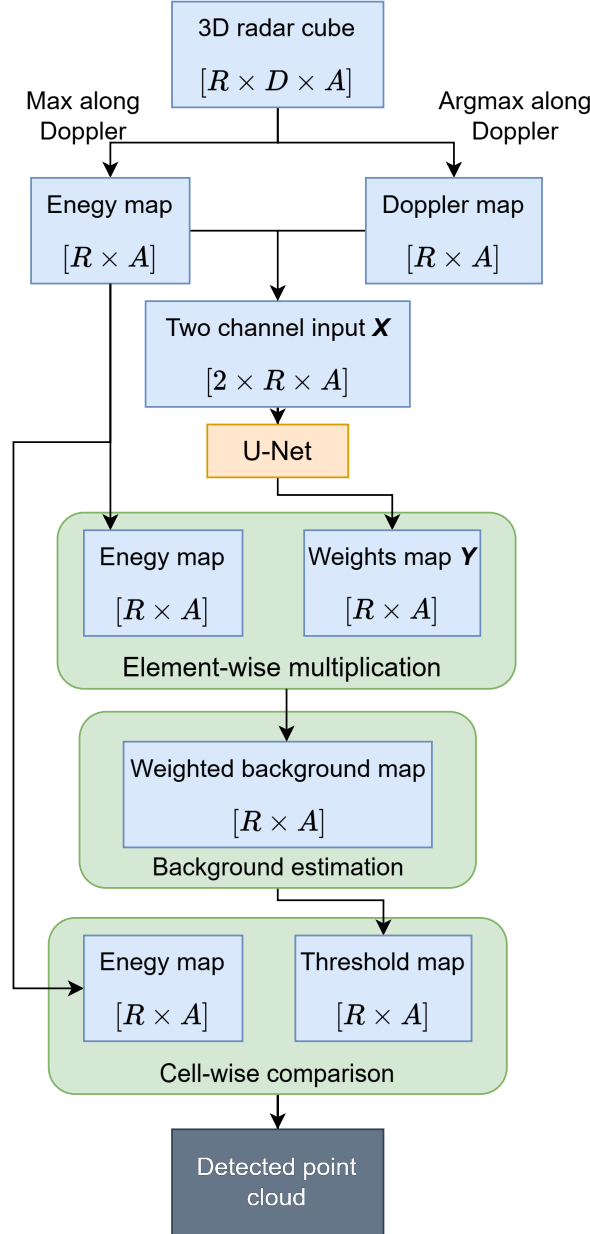


Figure 3.6: Pipeline of the proposed hybrid detector. Max and argmax operations compress the Doppler dimension into energy and Doppler-index maps. A U-Net predicts a background-reliability map, which suppresses probable target cells before local background estimation and cell-wise thresholding.

Let the normalised radar cube(divided by the global maximum power value) as the input be denoted by:

$$\mathbf{R} \in \mathbb{R}^{N_r \times N_a \times N_d}, \quad (3.31)$$

where $r \in \{0, \dots, N_r - 1\}$, $a \in \{0, \dots, N_a - 1\}$, and $d \in \{0, \dots, N_d - 1\}$ index the range, azimuth, and Doppler dimensions, respectively. In the radar configuration used in this study, $N_d = 128$. Max and argmax operations along the Doppler dimension produce an energy map and a normalised Doppler-index map as:

$$E_{r,a} = \max_{0 \leq d < N_d} R_{r,a,d}, \quad (3.32)$$

$$D_{r,a} = \frac{1}{N_d - 1} \arg \max_{0 \leq d < N_d} R_{r,a,d}. \quad (3.33)$$

The index values are normalised before this operation, by dividing the Doppler-bin index by $N_d - 1 = 127$ maps $D_{r,a}$ to $[0, 1]$. The two maps are stacked to construct the Unet network input

$$\mathbf{X} = \text{stack}(\mathbf{E}, \mathbf{D}) \in \mathbb{R}^{2 \times N_r \times N_a}. \quad (3.34)$$

This Doppler compression from 128 bins to two channels assumes that each range–azimuth cell contains at most one dominant target response. Under this assumption, $E_{r,a}$ retains the strength of the dominant return, while $D_{r,a}$ retains the Doppler bin associated with that return. The representation aligns the network input with the two-dimensional range–azimuth detection problem and avoids the cost of processing the full cube with heavier three-dimensional convolutions. Its limitation is that secondary Doppler peaks are discarded. If multiple objects with different, spread radial velocities contribute to the same range–azimuth cell, only the strongest response is represented.

The tensor $\mathbf{X}$ is then processed by a U-Net, denoted by f_θ , which predicts

$$\mathbf{Y} = f_\theta(\mathbf{X}) \in [0, 1]^{N_r \times N_a}. \quad (3.35)$$

The network architecture is shown in Fig. 3.7. The encoder contains three feature scales. Each scale uses convolution, batch normalisation, and rectified linear unit activation to extract spatial features, while strided convolution and max pooling reduce the spatial resolution and increase the number of channels. The decoder restores the resolution by bilinear interpolation. At each decoder scale, the upsampled features are concatenated with the corresponding encoder features through a skip connection and refined by convolutional blocks. A final 3×3 convolution with one output channel is followed by a sigmoid activation. Consequently, a large value of Y indicates that the cell is likely to represent background, whereas a small value indicates a probable target response.

Retaining this continuous output is central to the proposed method. Direct thresholding would assign the same binary label to all scores above the selected operating point, despite differences in their predicted confidence. It would also make point generation depend on a network-output threshold without an explicit connection to the local radar background. The proposed method instead uses the continuous output to control how strongly each cell contributes to background estimation. The final decision remains an explicit comparison between the measured radar energy and a locally derived threshold.

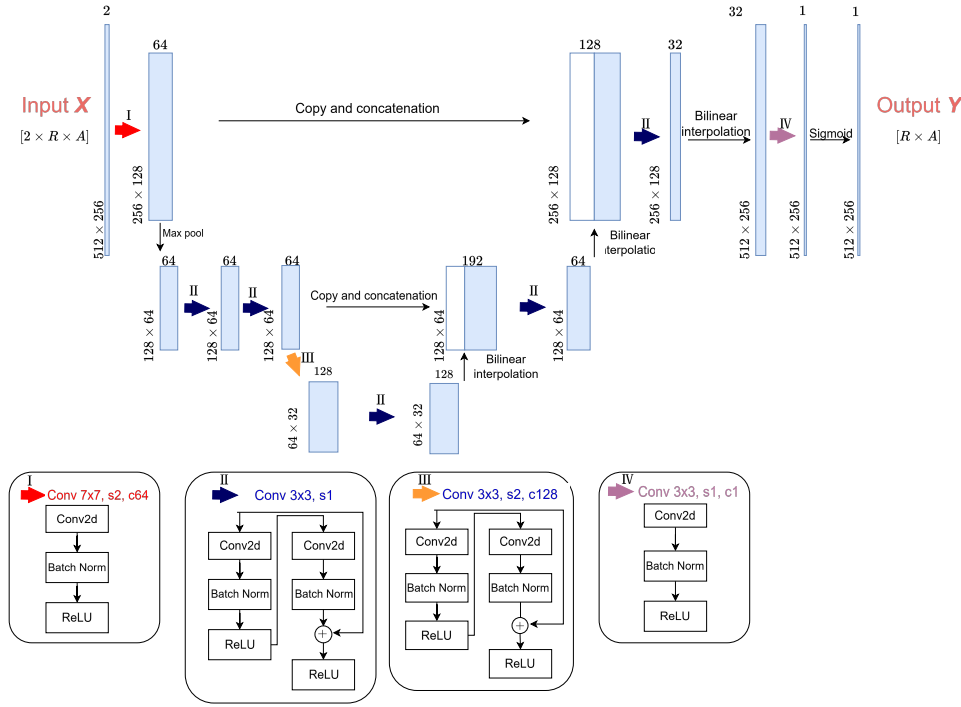


Figure 3.7: U-Net used to estimate the background-reliability map. The encoder extracts features at three spatial scales, and the decoder uses bilinear interpolation and skip connections to recover the input resolution. A one-channel convolution and sigmoid activation produce the final map. The four blocks at the bottom of the figure show the operations represented by the respective coloured arrow.

The network is supervised using an occupancy map $\mathbf{H} \in \{0, 1\}^{N_r \times N_a}$ obtained by voxelising the synchronised LiDAR point cloud on the range–azimuth grid. To reduce sensitivity to small spatial discrepancies between the radar measurements and LiDAR-derived labels, the hard occupancy map is expanded along the azimuth dimension. In accordance with the implementation, soft labels are used, which means instead of the hard label given by LiDAR of 0 and 1, the label spreads to nearby with a Gaussian blur with kernel size 1×5 and standard deviations $(\sigma_r, \sigma_a) = (0.1, 2.0)$ is applied to $\mathbf{H}$. Here use $\sigma = 0.1$ as zero to meet PyTorch requirements. Given a vertical kernel size of 1, no vertical blurring is applied regardless of σ . The blurred values are divided by the scaling constant W_c , clipped to $[0, 1]$, and combined with the original hard labels by element-wise maximisation:

$$\tilde{G}_{r,a} = \text{clip}_{[0,1]} \left(\frac{\mathcal{G}_{1 \times 5, (0.1, 2.0)}(\mathbf{H})_{r,a}}{W_c} \right), \quad (3.36)$$

$$G_{r,a} = \max(H_{r,a}, \tilde{G}_{r,a}), \quad (3.37)$$

where $\mathcal{G}_{1 \times 5, (0.1, 2.0)}$ denotes the Gaussian-blur operator. The final target $G_{r,a} \in [0, 1]$ preserves every original occupied cell while assigning smaller values to neighbouring azimuth cells. This softening accommodates limited spatial misalignment without removing the original LiDAR occupancy labels.

Because $G_{r,a}$ represents target occupancy whereas $Y_{r,a}$ represents background reliability, the predicted target probability is defined as $p_{r,a} = 1 - Y_{r,a}$. Focal loss is used to address the

imbalance between occupied and background cells and to emphasise difficult samples [50]:

$$\mathcal{L}_{\text{focal}} = -\frac{1}{N_r N_a} \sum_{r=0}^{N_r-1} \sum_{a=0}^{N_a-1} [\alpha G_{r,a} (1 - p_{r,a})^\gamma \log(p_{r,a}) + (1 - \alpha) (1 - G_{r,a}) p_{r,a}^\gamma \log(1 - p_{r,a})], \quad (3.38)$$

where $\alpha \in [0, 1]$ balances the target and background terms, and $\gamma \geq 0$ down-weights well-classified cells.

The proposed network was trained for 50 epochs with a batch size of 16 on an NVIDIA Tesla V100 GPU. Training used focal-loss parameters $\alpha = 0.95$ and $\gamma = 2.0$. The model was optimised using AdamW with an initial learning rate of 10^{-4} and a weight decay of 10^{-4} . A cosine-annealing schedule reduced the learning rate over the 50 training epochs to a minimum of 10^{-6} . The random seed was fixed at 42, and the checkpoint with the lowest validation loss was selected for evaluation.

At inference, the background-reliability map is multiplied element-wise by the original energy map:

$$B_{r,a}^w = Y_{r,a} E_{r,a}. \quad (3.39)$$

Probable background cells remain approximately unchanged, while probable target cells are attenuated. The resulting map B^w is used directly for local background estimation. The operation is not a normalised weighted average: the subsequent mean is not divided by the sum of the reliability weights. A low value of $Y_{r,a}$ therefore reduces the contribution of a probable target cell to the estimated background, which limits threshold inflation caused by target contamination. The example of $Y_{r,a}$ and B^w is shown in Figure 3.8

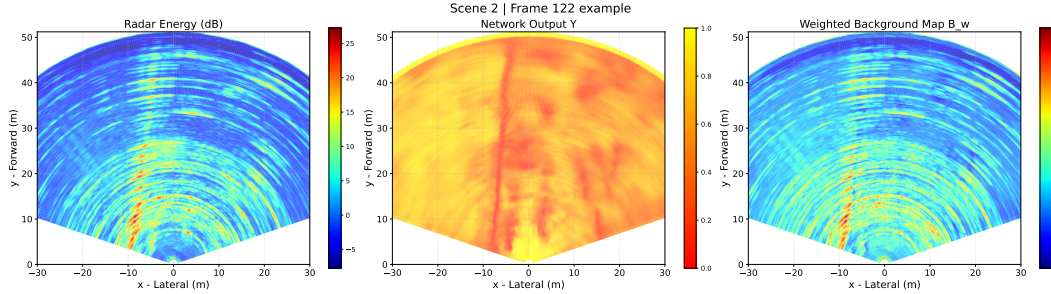


Figure 3.8: This figure from the left to right shows the example of original radar energy map, network output $Y_{r,a}$ and weighted background map B^w

For each cell under test (r, a) , the local background is estimated using a 5×5 window centred on that cell. The centre cell is excluded, no additional guard cells are used, and reflection padding is applied at the range and azimuth boundaries, which is to mirror the inner pixels on the edge. Let

$$\mathcal{N} = \{-2, -1, 0, 1, 2\}^2 \setminus \{(0, 0)\} \quad (3.40)$$

denote the 24 relative positions in the window. The local background estimate is:

$$\hat{\mu}_{r,a} = \frac{1}{24} \sum_{(u,v) \in \mathcal{N}} B_{\rho_r(r+u), \rho_a(a+v)}^w, \quad (3.41)$$

where $\rho_r(\cdot)$ and $\rho_a(\cdot)$ denote reflection mappings at the two boundaries. The cell-wise threshold and final detection are then defined as:

$$T_{r,a} = \eta \hat{\mu}_{r,a}, \quad (3.42)$$

$$Z_{r,a} = \begin{cases} 1, & E_{r,a} > T_{r,a}, \\ 0, & E_{r,a} \leq T_{r,a}, \end{cases} \quad (3.43)$$

where η is the threshold factor, and $\hat{\mu}_{r,a}$ is the estimated background level. The comparison uses the original energy map rather than the weighted map, so the network modifies only the background estimate and does not directly remove responses from the detection input. Varying η at inference adjusts the balance between detection probability and false-alarm probability without retraining the network. The procedure is described as CFAR-like because it retains local background estimation and an adjustable cell-wise threshold, but η is selected empirically rather than derived from an assumed distribution with a guaranteed design probability of false alarm. It is selected via a grid search over $[1.0, 4.0]$ to maximize the average performance.

3.4. Summary

This chapter described the signal-processing foundations, comparison methods, and proposed hybrid detector used in this thesis. CA-CFAR and OS-CFAR provide model-based baselines with explicit local thresholding, while the reproduced deep learning detector predicts occupancy directly from three consecutive radar cubes. The proposed method instead compresses each radar cube into energy and dominant-Doppler maps and uses a U-Net to estimate a continuous background-reliability map. This map modifies the local background estimate, while the final decision compares the original radar energy with an explicit cell-wise threshold. Consequently, the threshold factor can be adjusted at inference time without retraining the network. The following chapter evaluates these methods on the RaDelft dataset and examines the contribution of the principal design choices through ablation experiments.

Results and Analysis

This chapter presents the dataset used for the study, together with evaluation metrics, baseline results, proposed method results, and ablation studies. The analysis first introduces the RaDelft dataset and the metrics used in the experiments, then compares CFAR baselines, the deep learning baseline identified from the literature, and the proposed method.

4.1. RaDelft Dataset

Publicly available automotive radar datasets are relatively scarce when considering datasets designed for target detection, and therefore containing experimental data before point-cloud generation [6], [12]. Compared to simulated datasets, real-world data retain the complex clutter/interference and dense-target characteristics of real driving environments, which are difficult to model precisely. This is crucial for validating the practical performance of detection algorithms. Among existing options, this report uses the *RaDelft* dataset, a publicly available real-world multisensor automotive dataset that was collected at TU Delft [6], [14].

The *RaDelft* dataset contains approximately 16,975 frames of real-world driving-scene data collected in Delft, Netherlands [6], [14]. The sensor suite is jointly calibrated and time-synchronized, consisting of radar, LiDAR, camera, and odometry sensors; its cross-sensor extrinsic calibration follows the joint calibration framework described by Domhof et al. [51]. The radar data are provided at multiple processing stages, including raw Analog-to-Digital Converter (ADC) data, radar cubes after range, Doppler and angle estimation, and point clouds, allowing detailed analysis of the signal processing pipeline.

The radar system is a Frequency-Modulated Continuous-Wave (FMCW) radar operating in a MIMO configuration. It uses 12 transmitters (Tx) and 16 receivers (Rx), forming a virtual array of 192 elements in total. The topology of the virtual array is shown in Figure 4.1. With 86 elements in the azimuth direction, this large aperture provides a theoretical angular resolution of 1.33° at boresight. However, the array is sparse in the elevation direction, which limits elevation resolution and can create artefacts in its estimation. Therefore, this thesis focuses on 2D detection without considering elevation. The waveform parameters are configured with a start frequency of 76 GHz and an effective bandwidth of 750 MHz, yielding a theoretical range resolution of 0.2 m. The system operates in a Time-Division Multiple Access (TDMA) mode [6]. The final processed radar cubes have dimensions of $512 \times 128 \times 256 \times 2$ (range

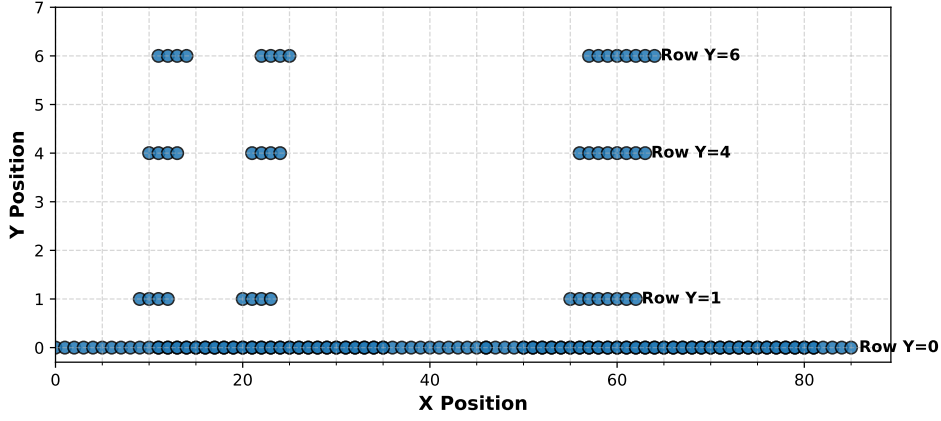


Figure 4.1: Antenna array topology for the radar used to collect the *RaDelft* dataset.

$\times$ Doppler $\times$ azimuth $\times$ [elevation index, power]). In the experiments, the elevation cube is discarded because of the sparsity of the elevation dimension, and the power cubes ($512 \times 128 \times 256$) are used as the input.

Before presenting the experimental results, two limitations of the dataset should be mentioned. First, the maximum unambiguous Doppler velocity is very low, at only 2.48 m/s, due to the TDMA scheme with many transmitters [34], [35]. Although the dataset authors mention a method to extend this range, it was not applied in the provided dataset. This is because the experiments in this thesis do not require precise Doppler information, hence this approach was not implemented here. Second, a residual mismatch remains between the sensors, although they have been calibrated. The radar captures signals and echoes in its field of views almost instantaneously, while the LiDAR scans its field of view by rotating. This causes a position mismatch for moving objects, because they move while the LiDAR is still scanning. This error depends on the LiDAR rotation direction and the object's speed. Due to the lack of a simple and suitable correction, this issue is left for future work. The Figure 3.4 mentioned in chapter 3 shows the overlapped data of target and non-target. This overlap could potentially caused by the misalignment of LiDAR and radar data, the presence of extended targets, or the presence of dense scenes with multiple targets.

4.2. Performance Metrics

To evaluate the experimental results, this thesis uses both accuracy-oriented and efficiency-oriented metrics. For detection accuracy, the main metrics are probability of detection (P_d), probability of false alarm (P_{fa}), and Chamfer Distance (CD). For target-level detection, P_d and P_{fa} are used. P_d measures the ratio of ground-truth targets correctly identified by the radar, while P_{fa} indicates the false-positive rate. They are calculated here as follows:

$$P_d = \frac{TP}{TP + FN}, \quad P_{fa} = \frac{FP}{TN + FP},$$

where TP , FP , FN , and TN represent true positives, false positives, false negatives, and true negatives, respectively.

However, these binary metrics have a significant limitation in the context of our dataset. As noted in the previous section, the time-synchronization mismatch between the radar and the

scanning LiDAR can cause slight position shifts. In the P_d calculations, if a detection shifts even slightly due to sensor mismatch, it is counted as a miss on the occupancy grid, causing the P_d score to drop sharply despite the detection being physically close.

To address this, CD is used as a complementary performance metric [6], [52]. Unlike the binary nature of P_d , CD measures the spatial similarity between the predicted radar point set (S_{pred}) and the ground-truth set from LiDAR (S_{gt}). Differ from the definition that calculates the average squared distance from each point in one set to its nearest neighbor in the other set. Here used in this thesis it calculates the average distance from each point in one set to its nearest neighbor in the other set without square, the formula is as follows:

$$CD(S_{pred}, S_{gt}) = \frac{1}{|S_{pred}|} \sum_{x \in S_{pred}} \min_{y \in S_{gt}} \|x - y\|_2 + \frac{1}{|S_{gt}|} \sum_{y \in S_{gt}} \min_{x \in S_{pred}} \|y - x\|_2.$$

Here, S_{pred} represents the predicted point cloud, S_{gt} represents the ground-truth point cloud, and x and y represent points in the predicted and ground-truth point clouds, respectively.

This “soft” metric is more robust to the slight spatiotemporal errors inherent in the dataset. It focuses on the overall structural alignment rather than a strict hit-or-miss binary threshold, providing a fairer evaluation of the spatial performance of radar-detected point clouds. Nevertheless, CD is not without flaws. Since it relies on squared Euclidean distance, it is highly sensitive to outliers; a single noise point far from the target can disproportionately inflate the error. Furthermore, it focuses solely on the nearest neighbor distance, meaning it does not account for the density distribution of the points. For now, this report uses CD in conjunction with P_d and P_{fa} to provide a broader view of performance. However, neither set of metrics is entirely ideal for this specific sensor setup. While P_d and P_{fa} are too rigid to handle sensor synchronization errors, CD is too vulnerable to noise and outliers. This trade-off highlights a limitation in current evaluation methods for asynchronous radar-LiDAR systems. Consequently, future research should aim to develop specialized metrics that tolerate slight temporal misalignments while remaining robust to noise and point-density variation.

In addition to accuracy-oriented metrics, computational cost is also evaluated. Real-time processing is essential in automotive scenarios, because Advanced Driver Assistance Systems (ADAS) must collect information and make decisions with low latency. Moreover, modern ADAS algorithms often run on edge devices, whose computational capacity is limited compared with desktop CPUs or GPUs. Therefore, the number of parameters and Floating-Point Operations (FLOPs) are introduced as model-efficiency metrics. The number of parameters approximates the model size that must be loaded, while FLOPs estimate the number of arithmetic operations required for one forward pass at inference. Compared with runtime per frame, which can vary substantially across hardware platforms, these metrics provide a more device-independent estimate of computational cost. The FLOPs and parameter counts were estimated in this work using the thop package [53].

4.3. Baseline Results

The overall results are shown in Table 4.1 using the metrics introduced in the previous section. The table compares four methods under the same training and testing configuration. For threshold-based methods, the reported values use the threshold factor that achieved the lowest CD for the corresponding configuration tuned based on the testing data. The proposed method

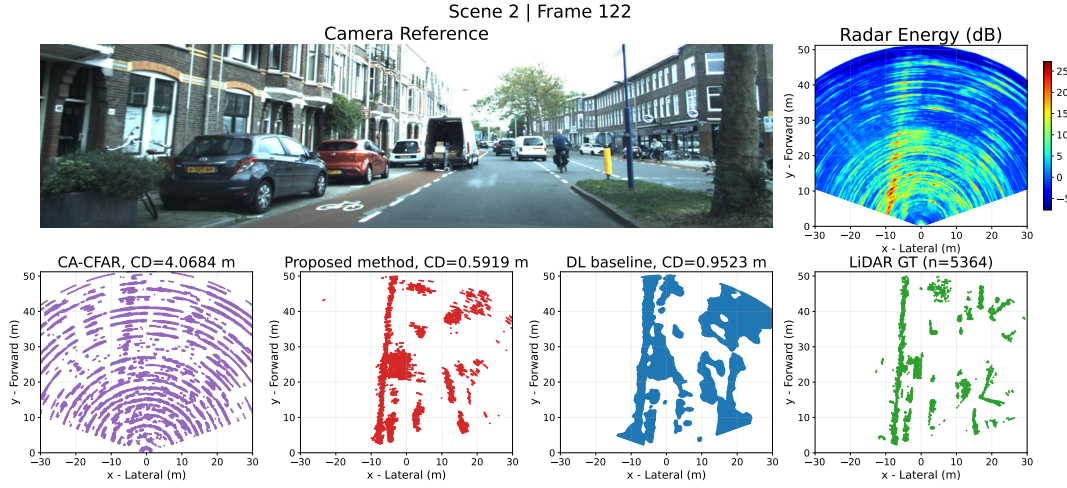


Figure 4.2: Overall point-cloud comparison. In the top left corner is the camera image, and in the top right corner is the radar energy map in XY view. The second row from the left to the right shows the point clouds generated by CA-CFAR, the proposed method, and Roldan et al. [6] respectively, as well as the LiDAR ground truth.

achieves the lowest CD, indicating the best spatial alignment with the LiDAR ground truth. Compared with the CFAR baselines, the proposed method improves P_d , P_{fa} , and CD. Compared with the deep learning baseline, it has a lower P_d but also a substantially lower P_{fa} and computational cost.

Table 4.1: Overall results for the baselines and the proposed method. The values are mean values over all the samples of the test dataset. The CA-CFAR and OS-CFAR and proposed method results here are shown with the best configurations, threshold factor and number of training cells and guardcells.

Method	P_d (%)	P_{fa} (%)	CD (m)	Parameters	FLOPs
CA-CFAR	22.84	10.93	3.19	/	30.1M
OS-CFAR	16.47	8.04	3.29	/	114.8M
Roldan et al. [6]	83.47	21.70	1.60	13.442M	131.66G
Proposed	42.95	5.00	1.36	0.87M	10.58G

In addition to the quantitative metrics, an overall visualization is shown in Figure 4.2. A single frame from Scene 2 is shown as an example, with the camera image and radar energy map in the XY view in the upper row, and the point clouds generated by CA-CFAR, the proposed method, Roldan et al. [6] respectively, as well as the LiDAR ground truth in the second row. This visualization illustrates representative differences between the methods. CA-CFAR generates many scattered points across the field of view, without forming clear object shapes. In contrast, the deep learning baseline generates denser point clouds, but the detected shapes are spatially inflated and essentially overestimated.

The following subsections discuss the CFAR baselines, the deep learning baseline, and the proposed method in more detail.

4.3.1. CFAR Baselines

This subsection reports the CFAR baseline results under different training-cell and guard-cell configurations. Table 4.2 shows the results of CA-CFAR and OS-CFAR under different configurations. Here, T denotes the number of training cells and G denotes the number of guard cells. For example, $T = 5$ means that five training cells are used around the cell under test, and $G = 5$ means that five guard cells are excluded around it. The experiments focus on CA-CFAR and OS-CFAR because CA-CFAR is used as part of the proposed method, while OS-CFAR is a widely used CFAR variant [16], [21]. Therefore, these two methods provide representative classical baselines without reporting every CFAR variant. The results show that changing the CFAR configuration affects performance, but does not resolve the main limitation. All CFAR configurations produced low P_d , relatively high P_{fa} , and high CD. CA-CFAR with $T = 5$ and $G = 5$ achieved a higher P_d than CA-CFAR with $T = 5$ and $G = 0$, but this came with a higher P_{fa} . Similarly, OS-CFAR varied across configurations, but none of the tested settings reduced the CD below 3 m.

Table 4.2: CFAR results under different configurations. T denotes the number of training cells and G denotes the number of guard cells. Each result is reported using the threshold factor that achieved the best CD.

Method	P_d (%)	P_{fa} (%)	CD (m)	FLOPs
CA-CFAR ($T = 5, G = 5$)	22.84	10.93	3.19	30.1M
CA-CFAR ($T = 5, G = 0$)	13.89	8.26	3.24	54.1M
CA-CFAR ($T = 9, G = 0$)	18.95	9.73	3.22	58.9M
OS-CFAR ($T = 5, G = 5$)	16.47	8.04	3.29	114.8M
OS-CFAR ($T = 5, G = 0$)	5.19	3.02	3.58	334.9M
OS-CFAR ($T = 9, G = 0$)	11.30	5.83	3.42	382.2M

After the quantitative comparison, Figure 4.3 shows the point clouds generated by four CFAR configurations. The first row shows from left to right the camera image, radar energy map and LiDAR point cloud in XY view, while the second row shows the CFAR point clouds. The visualization further shows that parameter tuning alone does not make the CFAR outputs geometrically reliable. Across the tested configurations, the point clouds remain sparse or scattered. This is consistent with the high CD values in Table 4.2, which indicate poor spatial alignment with the LiDAR ground truth.

Figure 4.4 and Figure 4.5 show good and bad frame examples for the best CA-CFAR and OS-CFAR configurations, with $T = 5$, $G = 5$, and $\alpha = 1.5$. The good/bad frames were selected according to their CD values, with the best frame having the lowest CD and the worst frame having the highest CD. Even in the best examples, the generated point clouds are sparse and do not accurately represent target shapes. In the worst examples, the point clouds are scattered and fail to capture meaningful structure. The visual and quantitative results consistently indicate that the CFAR baselines are limited in automotive scenarios. They are interpretable and tunable, but they struggle to generate dense and geometrically accurate point clouds, as reflected by their low P_d , relatively high P_{fa} , and high CD values. This behaviour is consistent with the limitations discussed for dense and non-homogeneous automotive scenes: targets are often extended, their apparent size changes with range and viewing direction, and target and clutter responses may not be well separated by a single power threshold. These limitations

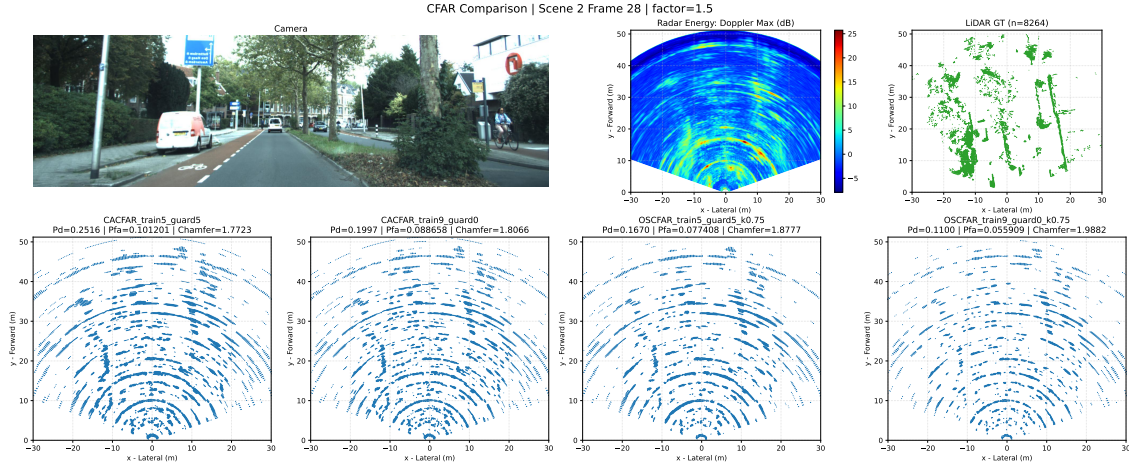


Figure 4.3: Point-cloud comparison of four CFAR configurations for an example of data frame. CFAR point clouds are shown in the second row.

motivate methods that retain explicit threshold control while using learned features to improve the target representation.

Figure 4.6 shows a bar plot with the frame-level CD values for CA-CFAR across the test scenes. This plot makes the variation across frames visible, rather than only reporting one single mean value as in Table 4.2.

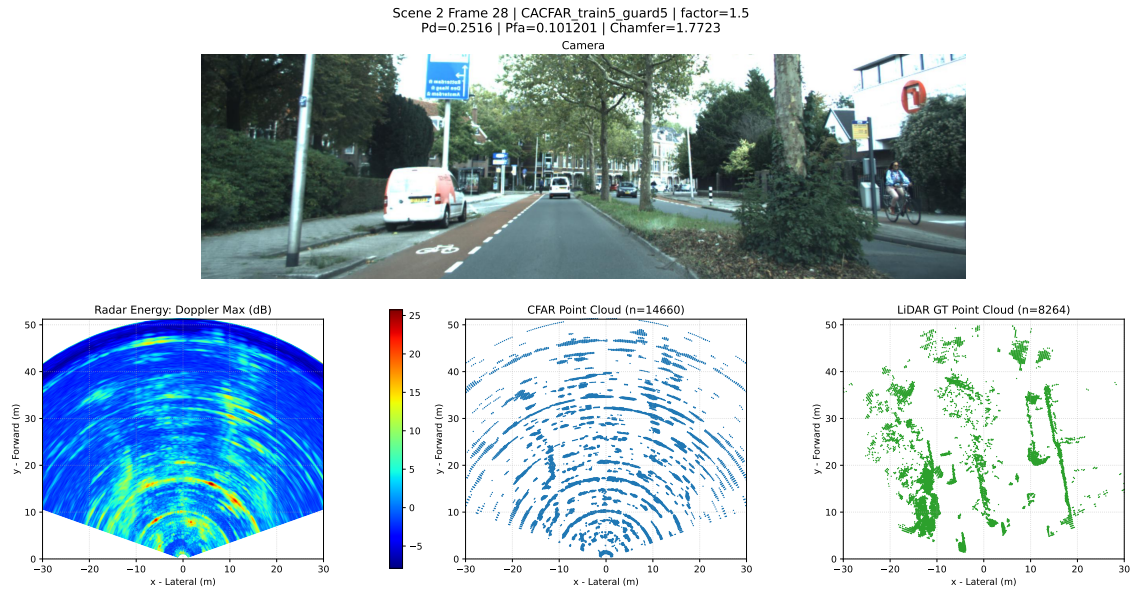
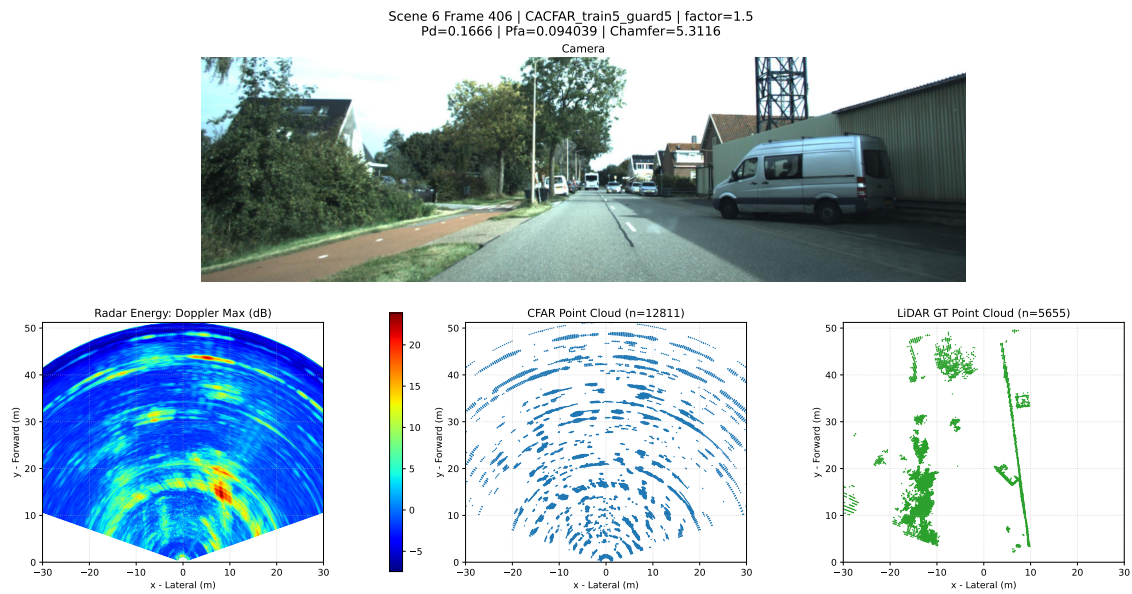
(a) Example point cloud of a *good* frame generated by CA-CFAR(b) Example point cloud of a *bad* frame generated by CA-CFAR

Figure 4.4: Example point clouds generated by CA-CFAR under the configuration $T = 5$, $G = 5$, and $\alpha = 1.5$. The frames are selected according to their CD values. Good frames are selected based on the lowest CD values, while bad frames are selected based on the highest CD values.

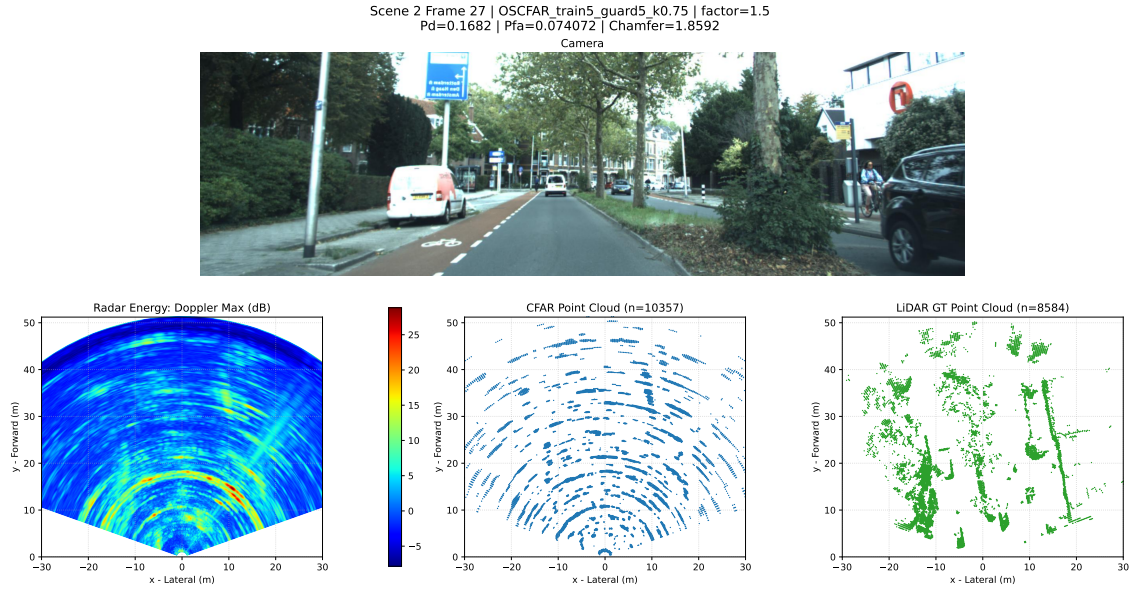
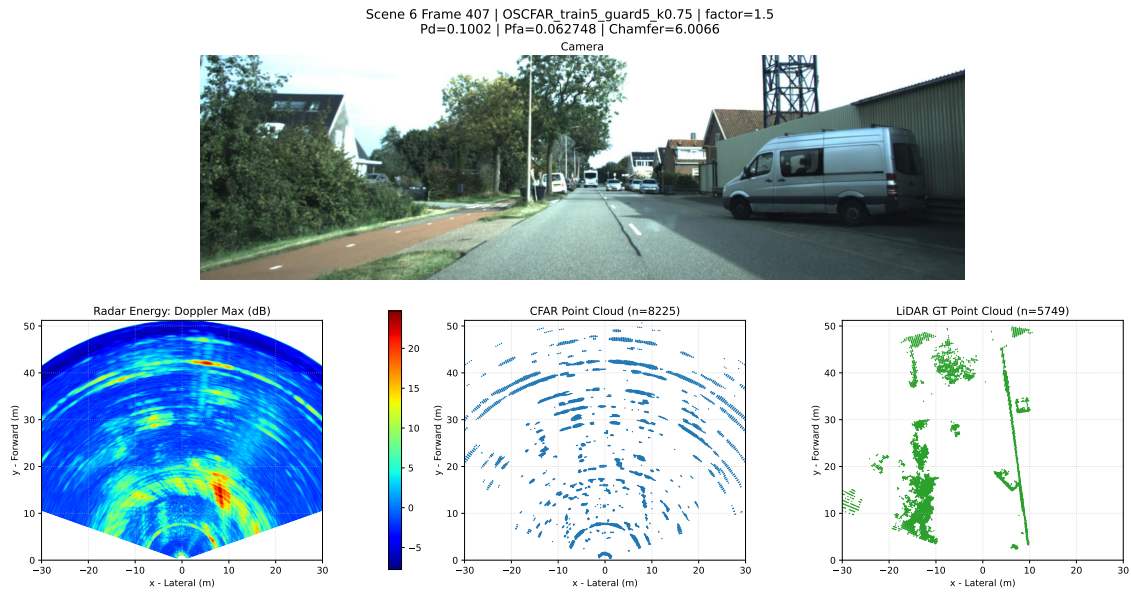
(a) Example point cloud of a *good* frame generated by OS-CFAR(b) Example point cloud of a *bad* frame generated by OS-CFAR

Figure 4.5: Example point clouds generated by OS-CFAR under the configuration $T = 5$, $G = 5$, and $\alpha = 1.5$. The frames are selected according to their CD values.

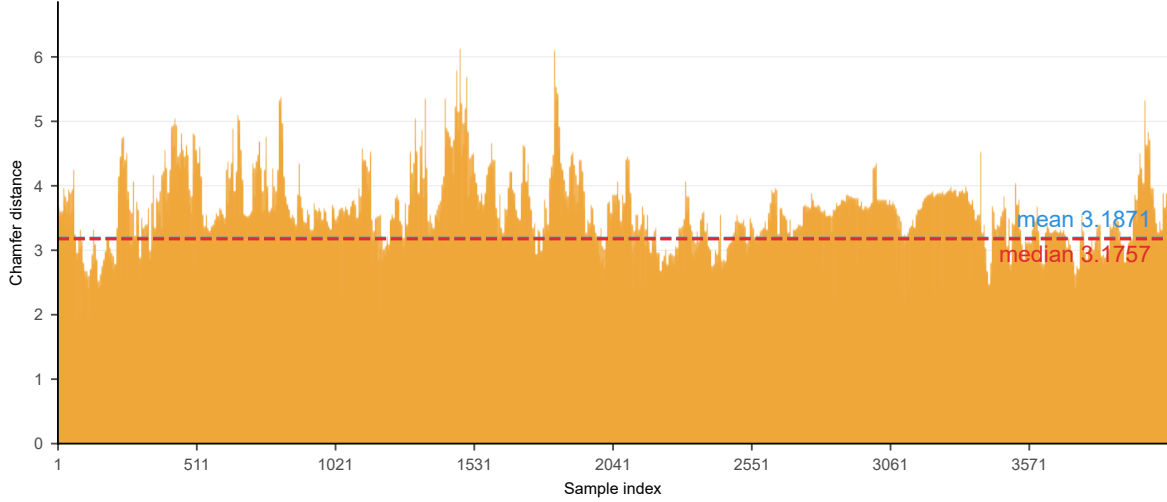


Figure 4.6: Bar plot of CD for all test frames using CA-CFAR with $T = G = 5$ and $\alpha = 1.5$.

4.3.2. Deep Learning Baseline

This subsection reports the results of the deep automotive radar detector proposed by Roldan et al. [6] and re-implemented as part of this thesis work. Figure 4.7 shows the CD distribution of the deep learning baseline across all test frames. The bar plot indicates that most frames have lower CD values than the CFAR baselines, showing the advantage of data-driven detection for spatial point-cloud generation.

The qualitative examples in Figure 4.8 show the main limitation of this baseline. Although it achieved the highest P_d in Table 4.1, its point clouds are often spatially inflated and overestimated in terms of extension, which is consistent with its higher P_{fa} . Some generated points also appear in regions without clear radar-energy support, making the detections difficult to interpret from the radar measurement alone. This suggests that the deep learning baseline improves detection density and spatial alignment, but it does so with more false positives, limited post-training control over the operating point, and substantially higher computational cost.

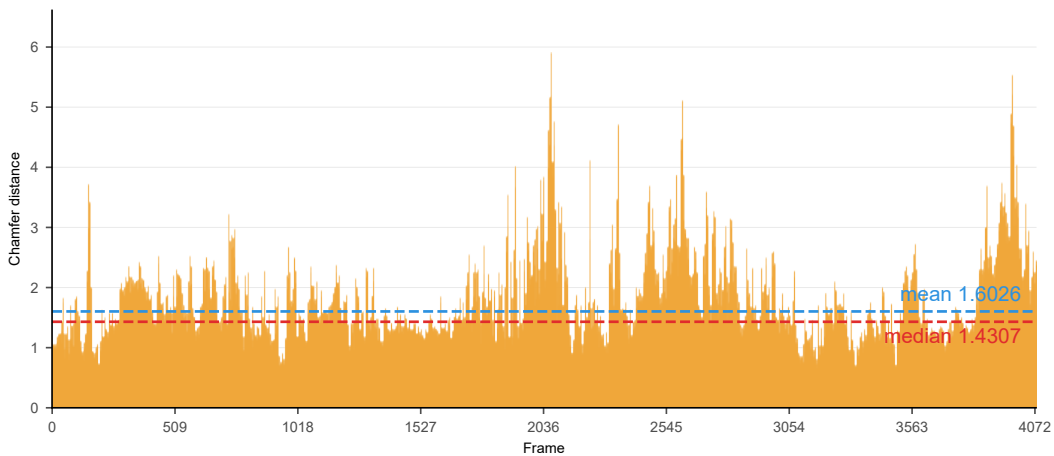
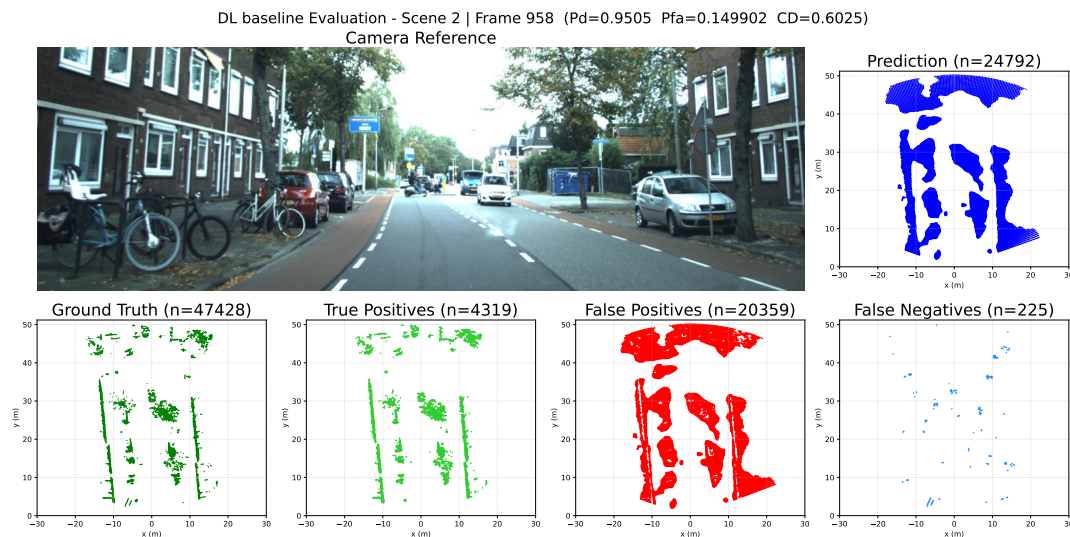
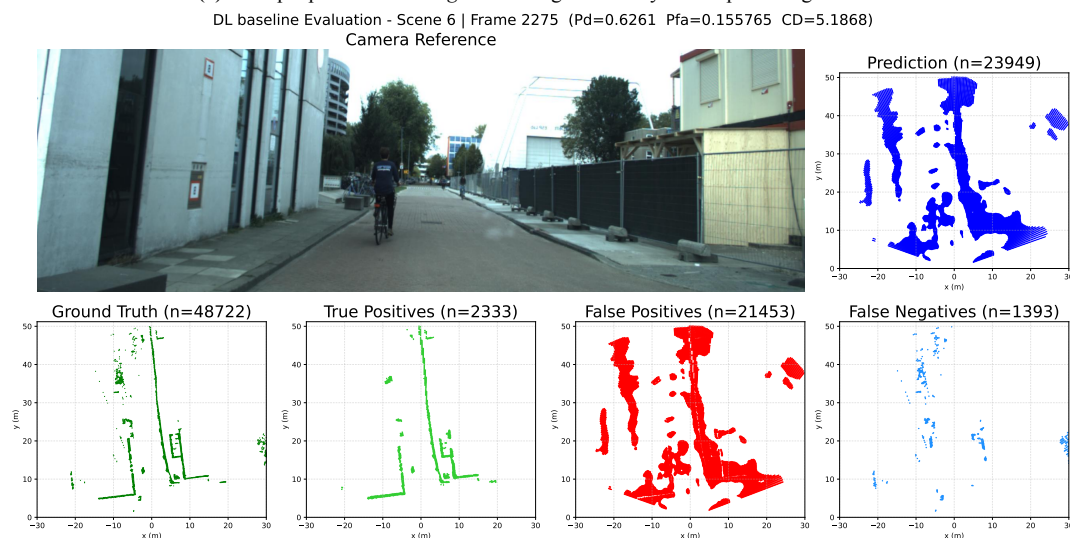


Figure 4.7: CD bar plot for all test frames using the deep learning baseline proposed by Roldan et al. [6].



(a) Example point cloud of a *good* frame generated by the deep learning baseline



(b) Example point cloud of a *bad* frame generated by the deep learning baseline

Figure 4.8: Point clouds of good and bad frames generated by the deep learning baseline. The first row is the camera and point cloud generated by network in XY view. The second row from left to right is the LiDAR point cloud and the True Positives, False Positives and False Negatives points of network.

4.3.3. Summary of Baseline Results

The baseline results reveal two complementary limitations. The CFAR baselines are interpretable and adjustable, but their point clouds remain sparse and geometrically inaccurate in dense automotive scenes. The deep learning baseline generates more complete point clouds and achieves the highest P_d , but this comes with spatially inflated detections, higher P_{fa} , limited interpretability, and much higher FLOPs. These observations motivate a detector that combines learned radar features with an explicit threshold factor α , so that the final operating point can still be adjusted after training.

4.4. Proposed Method Results

This section presents the results of the proposed method using the same metrics as the baselines. For each threshold-based configuration, the reported operating point uses the α value that achieved the lowest CD for that experiment. The goal of the proposed method is not to maximize P_d alone, but to produce cleaner point clouds with low CD, controlled P_{fa} , lower FLOPs, and an adjustable threshold factor.

Figure 4.9 shows two example frames with relatively low and high CD values. For the low-CD frame, the point clouds generated by the proposed method are more compact and better aligned with the LiDAR ground truth than those generated by CFAR, while avoiding the spatially inflated appearance of the deep learning baseline. For the high-CD frame, the generated points differ substantially from the LiDAR points. This difference may arise partly from physical differences between the sensors, as some responses may be detected only by radar. Moreover, radar can exhibit strong sidelobes in the presence of strong targets. For targets at high elevation, these sidelobes may be detected as false-positive points, consistent with the ring-like detections reported for CFAR processing on RaDelft [6]. Furthermore, the proposed method was trained on a limited dataset and may not generalise well to all scenarios, especially those that are underrepresented in the training data. This limitation could lead to poor performance on certain frames and higher CD values.

The ROC curve comparison in Figure 4.10 shows how changing α affects the trade-off between P_d and P_{fa} , and demonstrates that the proposed method provides a better operating curve than CA-CFAR when considering an average of P_d and P_{fa} for all frames of the test Scenes 2 and 6.

Figure 4.11 shows in a bar plot that most frames have relatively low CD. The violin plots in Figure 4.12 and Figure 4.13 show that the proposed method has a lower median CD and a narrower interquartile range than both CA-CFAR and the deep learning baseline, indicating more consistent performance of spatial accuracy across frames.

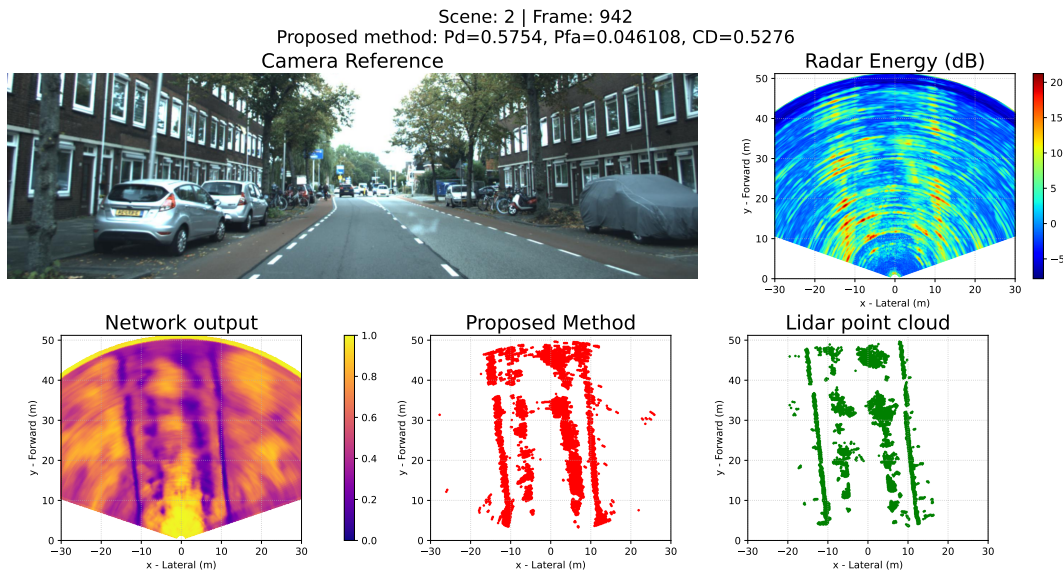
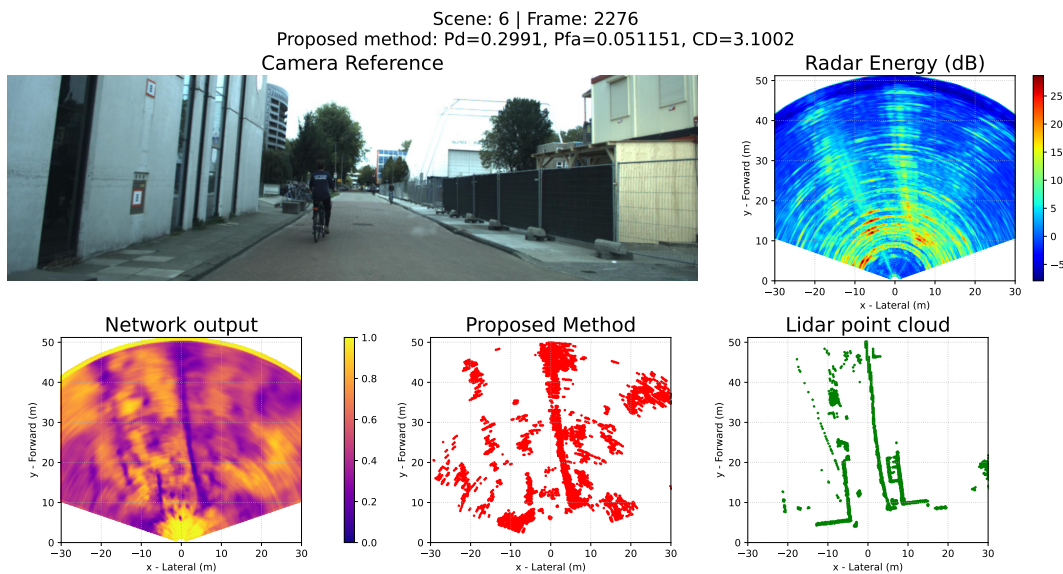
(a) Example point cloud of a *good* frame generated by the proposed method(b) Example point cloud of a *bad* frame generated by the proposed method

Figure 4.9: Point clouds of good and bad frames generated by the proposed method. In the first row is the camera and radar energy map in XY view. In the bottom row from left to right are network output, final generated point cloud and LiDAR point cloud.

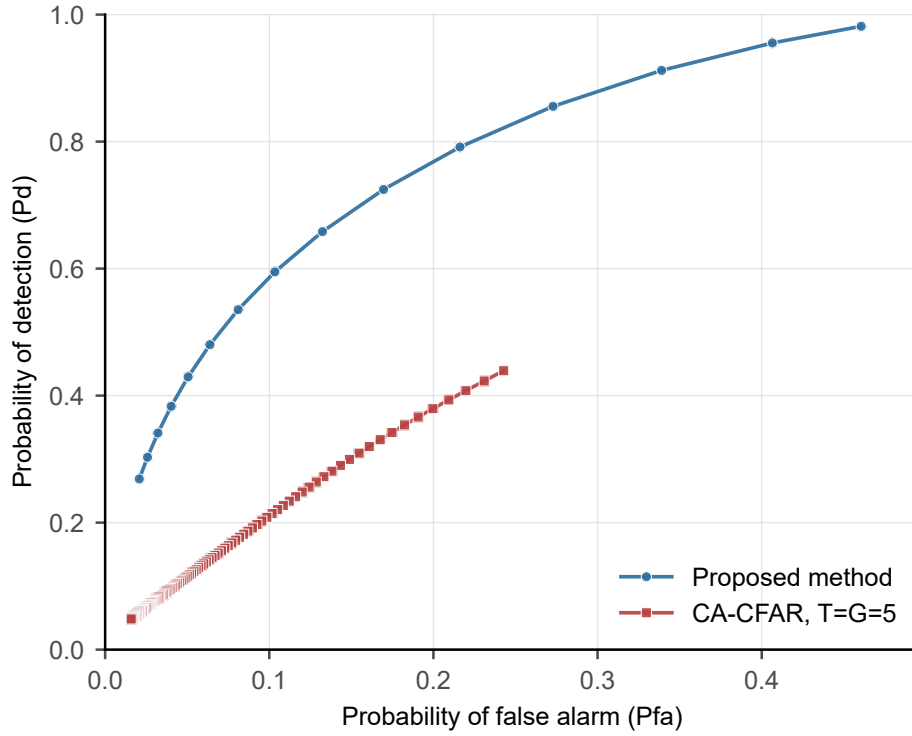


Figure 4.10: ROC curve comparison between CA-CFAR and the proposed method on test Scenes 2 and 6. The curve shows method performance in terms of P_d and P_{fa} under different threshold factors. The factors are swept across a range of values from 1.0 to 4.0. Curves closer to the upper-left corner indicate better performance.

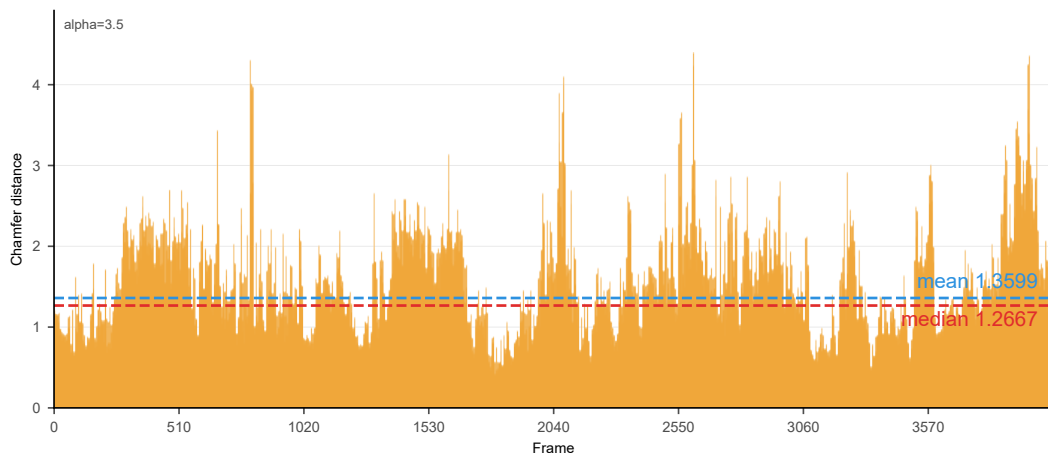


Figure 4.11: Bar plot of CD for the proposed method on the test scenes, with 4080 frames in total.

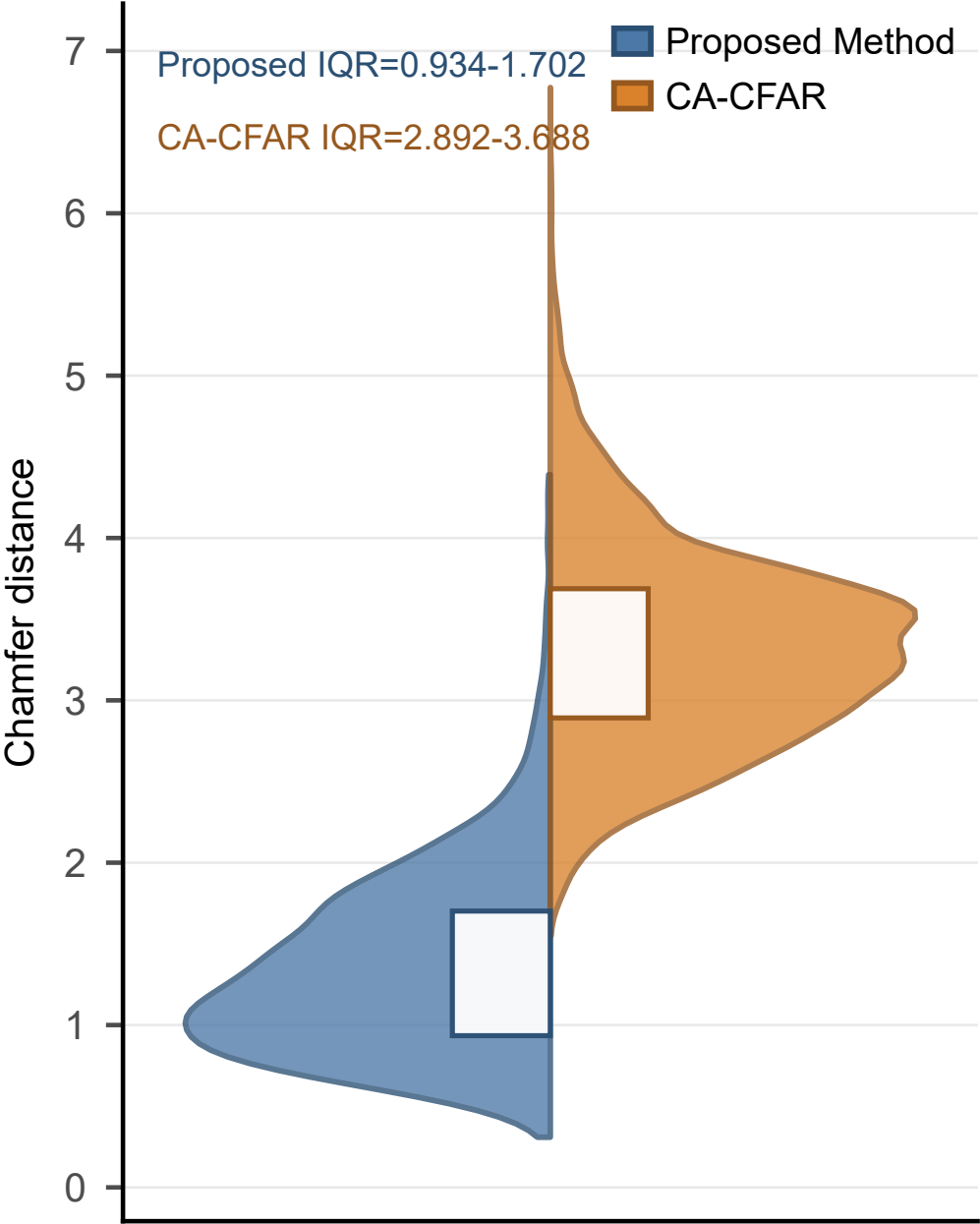


Figure 4.12: Violin plot of the proposed method versus CA-CFAR. The IQR is the CD range from the 25th to the 75th percentile.

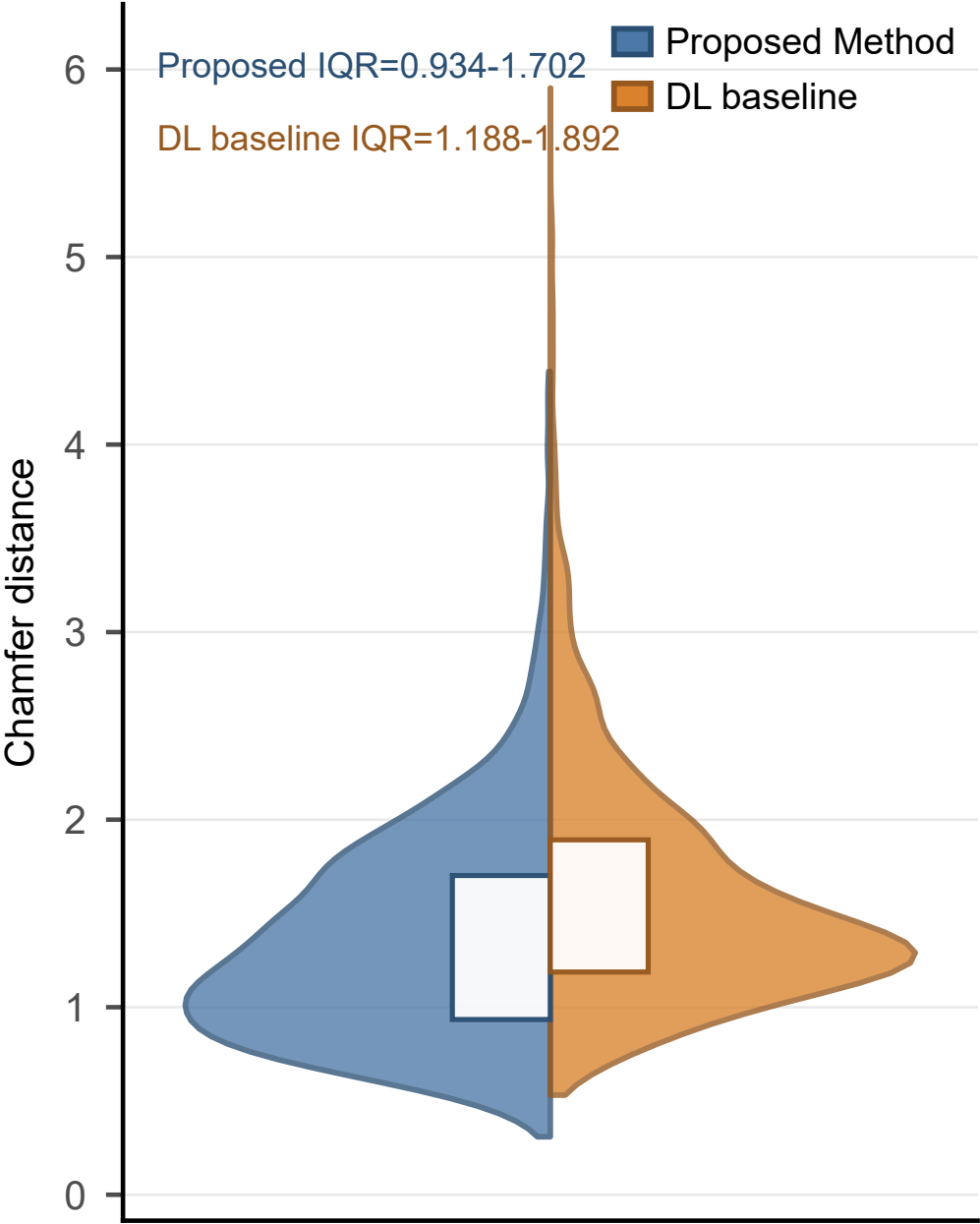


Figure 4.13: Violin plot of the proposed method versus the deep learning baseline. The IQR is the CD range from the 25th to the 75th percentile.

4.4.1. Backbone analysis

Different U-Net variants were evaluated for the proposed method to assess their behaviour in this task. Table 4.3 shows the results. As network depth and the corresponding level of extracted features vary between architectures, the default number of layers was selected for each network to enable a fair comparison. U-Net++ and U-Net 3+ achieve slightly higher P_d and lower CD than the original U-Net, but they also have higher P_{fa} and significantly more parameters and FLOPs. This indicates that while these variants can improve detection performance, they do so at the cost of increased computational complexity. To balance performance and efficiency, the original U-Net is selected as the preferred backbone for the proposed method in this study.

Table 4.3: Results for different U-Net variants using the default number of layers reported in the original papers [54], [55], [56]. A different layer configuration is used in the final results.

Method	P_d (%)	P_{fa} (%)	CD (m)	Parameters	FLOPs
U-Net[54]	47.89	5.72	1.472	14.318M	19.722G
U-Net++[55]	49.74	6.09	1.432	15.656M	42.14G
U-Net 3+[56]	51.63	6.39	1.437	21.393M	245G

4.4.2. Ablation Study

To verify the effectiveness of the proposed method, an ablation study is conducted. The study evaluates different U-Net variants, U-Net depths, and soft target mechanism.

Table 4.4 compares different U-Net depths. Here, the level denotes the number of feature scales in the U-Net. For example, a 2-level U-Net contains one downsampling and one upsampling operation. Increasing the depth from 2 to 4 levels improves P_d , but also increases P_{fa} and computational cost. In terms of CD, the 3-level U-Net achieves the best result, indicating that this depth provides the best balance between spatial accuracy and efficiency for this detection task. Increasing the network depth further does not improve the overall behavior, possibly because the more complex networks become more sensitive to the training data. This is consistent with Figure 4.14: the training loss decreases as the depth increases, while the validation loss first decreases and then increases for deeper networks. Therefore, a 3-level U-Net is selected as the backbone for the proposed method.

Table 4.4: Results for different U-Net depths. Level denotes the number of feature scales in the U-Net. Here the 3-level version is used for the final implementation as it achieves the best CD.

Method	P_d (%)	P_{fa} (%)	CD (m)	Parameters	FLOPs
2-level	40.99	6.08	1.453	0.2M	5.99G
3-level	42.95	5.05	1.359	0.87M	10.58G
4-level	49.69	6.61	1.437	3.56M	15.15G
5-level	47.89	5.72	1.472	15.656M	19.722G

After the network architecture ablation, Table 4.5 evaluates the effect of soft targets mentioned in 3.3. Compared with hard targets, soft targets improve CD from 1.397 to 1.359 and increase P_d from 36.95% to 42.95%. The increase in P_{fa} from 4.07% to 5.05% indicates that the softer

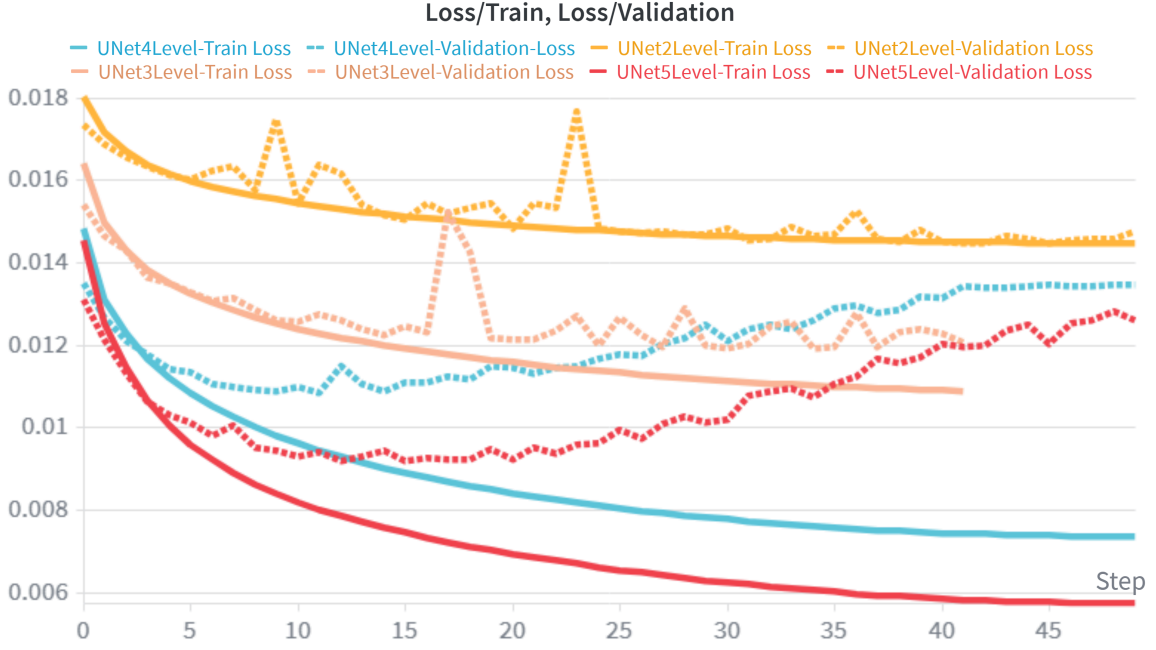


Figure 4.14: Validation loss and training loss for different U-Net depths, whose performance is analyzed in Table 4.4

supervision activates a broader spatial region around each target. This trade-off is consistent with the intended role of soft targets: they provide more tolerant spatial supervision under radar-LiDAR mismatch, improving spatial alignment while slightly increasing false positives.

Table 4.5: Comparison of soft and hard targets with focal loss.

Method	P_d (%)	P_{fa} (%)	CD (m)
Soft target + focal loss	42.95	5.05	1.359
Hard target + focal loss	36.95	4.07	1.397

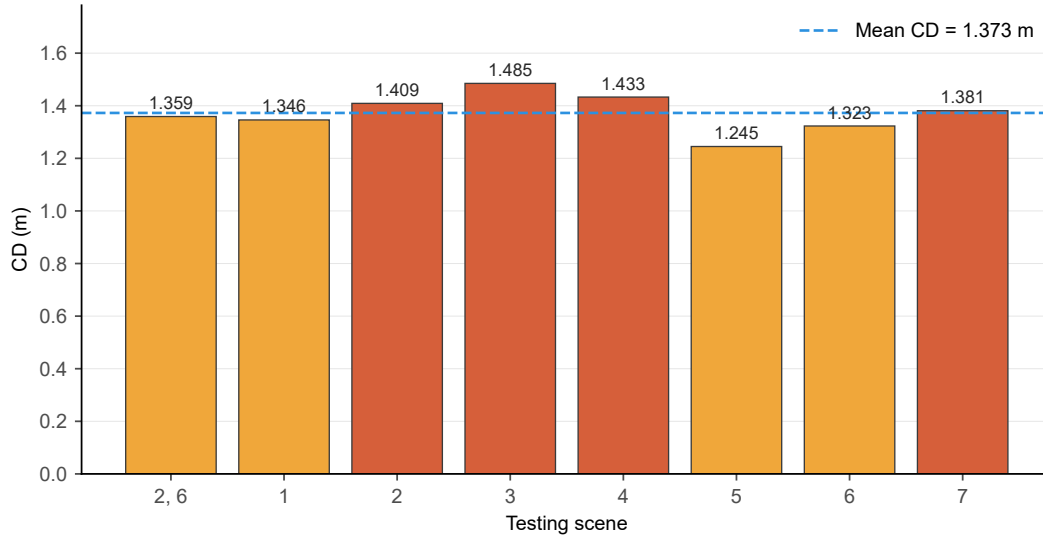
4.4.3. Leave One Scene Out experiment

Finally, a leave-one-scene-out experiment is conducted to evaluate how the proposed method generalizes across scenes. The results are shown in Table 4.6. Each row reports the best α for that split, selected according to the lowest CD under the corresponding experiment. Therefore, this table evaluates best-case spatial accuracy performance for each unseen scene, rather than comparing all scenes at one fixed threshold.

The CD values range from 1.245 to 1.485, showing that the proposed method does not suffer catastrophic spatial degradation when tested on unseen scenes. However, the balance between P_d and P_{fa} varies substantially across scenes. Scene 3 has the highest CD and low P_d with low P_{fa} , which indicates a conservative or under-detection case. Scene 1 has the highest P_d and highest P_{fa} , while still maintaining a relatively low CD, indicating a more aggressive detection case. Scene 5 achieves the lowest CD, but with a moderately high P_{fa} , suggesting good spatial alignment with more activated detections. Fig 4.15 shows a clearer CD comparison in different

Table 4.6: Best- α leave-one-scene-out results. Each row reports the operating point with the lowest CD for that split.

Training scenes	Testing scenes	P_d (%)	P_{fa} (%)	CD (m)
1, 3, 4, 5, 7	2, 6	42.95	5.05	1.359
2, 3, 4, 5, 6, 7	1	61.13	12.50	1.346
1, 3, 4, 5, 6, 7	2	39.30	3.72	1.409
1, 2, 4, 5, 6, 7	3	32.54	3.77	1.485
1, 2, 3, 5, 6, 7	4	42.78	6.08	1.433
1, 2, 3, 4, 6, 7	5	55.83	8.12	1.245
1, 2, 3, 4, 5, 7	6	53.75	8.68	1.323
1, 2, 3, 4, 5, 6	7	44.47	6.74	1.381

**Figure 4.15:** Leave-one-scene-out CD bar plot. The bars show the best- α CD values for each leave-one-scene-out configuration, with the corresponding testing scenes indicated. Each value is the average across the corresponding test scenes; red indicates a value above the overall average, whereas yellow indicates a value below it.

scenes by bar plot, where the red bars are scenes above the average CD and yellow bars are the scenes below the average CD. These results show that the best operating point is scene-dependent. Future work could therefore use scene-adaptive α selection or validation-based threshold tuning instead of using a single fixed threshold across all scenes.

4.5. Summary

This chapter evaluated the proposed hybrid detector against CA-CFAR, OS-CFAR, and the deep learning detector of Roldan et al. [6] on the RaDelft dataset. The results showed that the CFAR baselines retained explicit threshold control but produced sparse and spatially inaccurate point clouds across the tested configurations. The deep learning baseline achieved the highest detection probability, with $P_d = 83.47\%$, but also produced the highest false-alarm probability, with $P_{fa} = 21.70\%$, and required substantially greater computational resources.

The proposed method achieved the lowest Chamfer Distance of the evaluated methods, reducing CD to 1.36 m compared with 3.19 m for CA-CFAR, 3.29 m for OS-CFAR, and 1.60 m for the deep learning baseline. It also reduced the measured P_{fa} to approximately 5.0%, although its P_d of 42.95% remained below that of the deep learning baseline. The ablation experiments identified the three-level U-Net as the most favourable balance between spatial accuracy and computational cost, while soft targets improved spatial alignment at the expense of a moderate increase in false alarms. Finally, the leave-one-scene-out experiments produced CD values between 1.245 and 1.485 m, indicating broadly similar best-case spatial performance across unseen scenes, although the preferred balance between P_d and P_{fa} remained scene-dependent. Overall, the results support learned background-reliability estimation as a means of improving local threshold-based radar detection while retaining post-training control over the operating point.

Conclusion and Future Work

5.1. Conclusion

This thesis investigated whether learned, data-dependent background weighting can mitigate the limitations of fixed-window CFAR detection while retaining an explicit and adjustable local thresholding stage. Classical CFAR detectors provide interpretable decisions and low computational cost, but their fixed reference-cell rules are poorly suited to dense automotive scenes in which targets are extended and the background is non-homogeneous. Fully learned detectors can model more complex radar features, but may require substantially greater computation and provide less control over the final operating point after training. To bridge these approaches, this thesis proposed a hybrid detector in which a U-Net predicts a continuous background-reliability map and a CFAR-like procedure uses this map to construct cell-wise detection thresholds.

The principal contribution of the proposed method is therefore not the direct replacement of threshold-based detection with a neural network. Instead, the network learns which range–azimuth cells should contribute to the local background estimate. Probable target cells are attenuated before the background is averaged, thereby limiting threshold inflation caused by target contamination, while the original radar energy remains the input to the final decision. This separation preserves an explicit relationship between the measured radar response, the estimated local background, and the resulting detection. It also allows the threshold factor to be changed at inference time without retraining the network.

The three objectives defined in Chapter 1 were addressed under a common experimental setting on the RaDelft dataset. First, CA-CFAR, OS-CFAR, and a deep learning detector were established as representative baselines. The CFAR experiments showed that changing the reference and guard-cell configurations altered the balance between detections and false alarms but did not overcome the poor geometric correspondence of their generated point clouds. The deep learning baseline generated denser point clouds and achieved the highest probability of detection, but its outputs were often spatially inflated and were associated with a higher false-alarm probability and substantially greater computational cost. These results exposed the complementary limitations that motivated the hybrid design.

Second, the proposed detector combined a compact two-channel representation of radar en-

ergy and dominant Doppler index with learned background-reliability estimation and local thresholding. Third, the overall comparison and ablation experiments evaluated both the resulting performance and the main architectural choices. As reported in Table 4.1, the proposed method achieved a CD of 1.36 m, compared with 3.19 m for CA-CFAR, 3.29 m for OS-CFAR, and 1.60 m for the deep learning baseline. It also achieved a measured P_{fa} of approximately 5.00%, while the corresponding values were 10.93%, 8.04%, and 21.70%, respectively. The proposed method did not maximise every metric: its P_d of 42.95% remained below the 83.47% achieved by the deep learning baseline. Its advantage instead lay in the combined operating characteristics of lower geometric error, fewer false alarms, post-training threshold adjustability, and lower model complexity. The final three-level U-Net contained 0.87 million parameters and required 10.58 GFLOPs, compared with 13.442 million parameters and 131.66 GFLOPs for the deep learning baseline.

While ablation study validate the efficacy of network structure choices, the cross-scene evaluation highlights the robustness of the final framework in realistic settings. Leave-one-scene-out validation firmly confirms the generalization capability of the model across unseen environments, achieving resilient point cloud quality across diverse test scenarios. At the same time, the observed variability in detection and false alarm trade-offs reveals that the method’s ideal operating point is sensitive to environmental context, suggesting that adaptive thresholding is a key consideration for deployment.

Taken together, these findings suggest that the reliability of the cells used for background estimation is a central design consideration for automotive radar detection. Fixed CFAR windows determine this reliability in advance, whereas the proposed method makes it dependent on the observed radar features. The results further indicate that learning need not replace the complete detection rule: it can instead support the part of a structured detector that is most difficult to specify with a fixed model. Within the evaluated setting, this hybrid organisation provides a useful balance between data-dependent feature extraction and an explicit, adjustable decision procedure.

Several limitations should be considered when interpreting these conclusions. The experiments used only the RaDelft dataset and focused on two-dimensional range–azimuth detection, so the results do not establish generalisation to other radar configurations, locations, or environmental conditions. The LiDAR-derived labels are an imperfect proxy for radar target visibility because the two sensors have different physical responses and acquisition processes. Residual temporal and spatial misalignment can therefore affect both training and evaluation. In addition, the max–argmax compression retains only the dominant Doppler component in each range–azimuth cell, and the fixed 5×5 neighbourhood still imposes a predefined spatial scale on background estimation. The reported procedure is consequently CFAR-like rather than a statistically calibrated CFAR detector: its threshold factor is selected empirically and does not guarantee a specified design probability of false alarm. Finally, neither strict occupancy-grid metrics nor CD fully resolves the evaluation difficulties caused by cross-sensor mismatch, outliers, and differences in point density.

Overall, this thesis shows that learned background-reliability estimation is a promising alternative to either fixing the reference set entirely in advance or replacing the detection rule with an end-to-end network. On RaDelft, the resulting detector improved geometric point-cloud quality and false-alarm performance over the evaluated CFAR baselines, while requiring sub-

stantially less computation than the evaluated deep learning baseline. Its remaining limitations identify concrete directions for improving supervision, adaptivity, statistical calibration, and external validity.

5.2. Future Work

There are few potential directions that can be considered as future work.

1. The most immediate direction is to improve the ground truth used for supervision and evaluation. LiDAR provides detailed spatial measurements, but it is only a proxy for radar target visibility. The rotating LiDAR acquires a scene over time, whereas the radar measurements used in this thesis represent a much shorter acquisition interval. Moving objects can consequently appear at different positions in the two sensor representations even after extrinsic calibration. Moreover, differences in penetration, reflectivity, angular response, and sensitivity mean that an object or surface visible to one sensor is not necessarily represented in the same way by the other. Future work could compensate for the temporal offset using LiDAR scan timing, vehicle odometry, and estimated object motion, and could represent the residual alignment uncertainty explicitly in the training targets. Sensor-aware labels that distinguish geometric occupancy from expected radar visibility may provide a more appropriate supervisory signal than treating every LiDAR point as an equally valid radar target. second direction is to make the remaining fixed components of the proposed detector adaptive. In the current evaluation, the threshold factor is selected using aggregate validation performance and the local background is computed with a fixed 5×5 window. The leave-one-scene-out results indicate that the favourable balance between P_d and P_{fa} changes across scenes, suggesting that one global threshold factor may not be optimal for every frame or operating condition. A future detector could predict a frame-level or region-level threshold factor from summary features using a small auxiliary network, or calibrate the factor against an explicit false-alarm objective. The reference geometry could similarly be extended through multi-scale windows, range-dependent windows, or a learned neighbourhood selection mechanism. Such extensions should retain the explicit detection stage and should be evaluated not only for accuracy but also for calibration, stability, and computational cost.
2. The statistical model underlying automotive radar detection also warrants further study. Classical target and background models were developed for settings that do not necessarily reflect the long-tailed interference, dense clutter, extended targets, sidelobes, and rapidly changing local statistics encountered in road scenes. The results of this thesis show that fixed power-based models and fixed reference rules are insufficient in the evaluated data, but they do not identify a complete replacement distribution. Future work could therefore characterise background and target responses conditionally on range, Doppler, scene type, and local spatial context. A model that combines such conditional statistics with learned reliability estimates may make it possible to derive thresholds with better calibrated false-alarm behaviour, rather than selecting the threshold factor solely from empirical performance.
3. Evaluation should be extended beyond the present dataset and metric set.
 - Cross-dataset experiments are required to further verify the generalization of the

method, which is to determine whether the learned background-reliability representation transfers across different dataset with different radar hardware, antenna geometries, resolutions, locations, and weather conditions.

- The current suite of evaluation metrics could be further expanded. While P_d and P_{fa} measure cell-wise detection capability and CD evaluates the spatial accuracy of the point clouds, they do not adequately capture point cloud density and inflation under minor spatio-temporal misalignments. Potential directions include exploring target-level matching criteria, robust point-set distance metrics, and uncertainty-aware occupancy representations. Finally, the efficiency indicated by parameter counts and FLOPs should be verified through deployment on representative automotive edge hardware, with measurements of runtime, memory use, latency, and energy consumption.

5.2.1. Alternative Directions Explored During This Study

Two preliminary directions considered during this study further clarify the importance of accurate supervision. Because targets and background cannot be separated using power alone, the first research activity attempted to learn a latent representation in which the background response of each cell followed a prescribed distribution. Under this formulation, each radar cell was mapped to a multidimensional vector and distributional constraints, such as Kullback–Leibler divergence or Wasserstein distance, were considered as a means of regularising background samples. A threshold derived in this latent space could, in principle, connect learned feature separation with an explicitly controlled false-alarm probability. In practice, this direction did not produce a sufficiently stable separation to be developed into the final method. A likely obstacle was that inaccurate or sensor-inconsistent labels assigned some radar responses to the wrong class, making it difficult to impose a coherent background distribution. This direction may become more viable with improved radar-specific labels, uncertainty-aware supervision, and a systematic study of whether the assumed latent distribution is appropriate.

The second preliminary direction attempted to filter LiDAR-derived labels that were unlikely to correspond to radar-visible targets. The filter combined object height with the peak-to-average ratio across the Doppler dimension. Height was used conservatively so that low obstacles relevant to vehicle safety were retained, while the Doppler peak-to-average ratio was intended to identify cells without a distinct radar response. This procedure removed some vegetation-related LiDAR points while preserving much of the principal vehicle and infrastructure structure, but it did not yield a consistent improvement across the evaluation metrics. The experiment illustrates that label cleaning based on hand-designed rules can exchange one form of mismatch for another: removing sensor-inconsistent labels may also discard weak but valid radar targets or alter the density of the reference point cloud. A more rigorous continuation would compare the filtered and unfiltered labels under controlled ablations, analyse the removed points by object type and range, and replace hard filtering with probabilistic radar-visibility weights where possible.

These exploratory directions converge on the same broader requirement. Further progress depends not only on a more expressive detector, but also on a clearer definition of what constitutes a radar target and how that definition should be represented by another sensing modality. Better-aligned and radar-aware supervision would support adaptive threshold selection, learned reference geometries, latent statistical modelling, and more reliable cross-dataset eval-

uation. These developments could move the current hybrid design from empirically adjustable CFAR-like detection towards a detector whose adaptability and false-alarm behaviour are both measurable and statistically interpretable.

References

- [1] F. Sezgin, D. Vriesman, D. Steinhauser, R. Lugner, and T. Brandmeier, “Safe autonomous driving in adverse weather: Sensor evaluation and performance monitoring,” in *2023 IEEE Intelligent Vehicles Symposium (IV)*, 2023, pp. 1–6. DOI: [10.1109/IV55152.2023.10186596](https://doi.org/10.1109/IV55152.2023.10186596).
- [2] I. Bilik, O. Longman, S. Villeval, and J. Tabrikian, “The rise of radar for autonomous vehicles: Signal processing solutions and future research directions,” *IEEE Signal Processing Magazine*, vol. 36, no. 5, pp. 20–31, 2019. DOI: [10.1109/MSP.2019.2926573](https://doi.org/10.1109/MSP.2019.2926573).
- [3] M. A. Richards and W. A. Scheer James A. and Holm, *Principles of Modern Radar, Volume I - Basic Principles*. SciTech Publishing, 2010, ISBN: 978-1-891121-52-4. [Online]. Available: <https://app.knovel.com/hotlink/toc/id:kpPMRVIBP8/principles-modern-radar/principles-modern-radar>.
- [4] A. De Maio and M. Greco, *Modern Radar Detection Theory* (Electromagnetics and Radar). Institution of Engineering and Technology, 2015, ISBN: 9781613531990. [Online]. Available: <https://books.google.nl/books?id=ZKGiCwAAQBAJ>.
- [5] J. Yang, J. Yi, T. Sakamoto, and X. Wan, “An extended target detector using image-processing techniques exploiting energy-spillover phenomenon in radar echoes,” *IEEE Sensors Journal*, vol. 23, no. 19, pp. 22 919–22 929, 2023. DOI: [10.1109/JSEN.2023.3305494](https://doi.org/10.1109/JSEN.2023.3305494). [Online]. Available: <https://doi.org/10.1109/JSEN.2023.3305494>.
- [6] I. Roldan, A. Palffy, J. F. P. Kooij, D. M. Gavrilă, F. Fioranelli, and A. Yarovoy, “A deep automotive radar detector using the radelft dataset,” *IEEE Transactions on Radar Systems*, vol. 2, pp. 1062–1075, 2024. DOI: [10.1109/TRS.2024.3485578](https://doi.org/10.1109/TRS.2024.3485578).
- [7] A. Srivastav and S. Mandal, “Radars for autonomous driving: A review of deep learning methods and challenges,” *IEEE Access*, vol. 11, pp. 97 147–97 168, 2023.
- [8] D. Gusland, S. Rolfsjord, and B. Torvik, “Deep temporal detection - a machine learning approach to multiple-dwell target detection,” in *2020 IEEE International Radar Conference (RADAR)*, 2020, pp. 203–207. DOI: [10.1109/RADAR42522.2020.9114828](https://doi.org/10.1109/RADAR42522.2020.9114828).
- [9] Y. Lin, X. Wei, Z. Zou, and W. Yi, “Deep learning based target detection method for the range-azimuth-doppler cube of automotive radar,” in *2023 IEEE 26th International Conference on Intelligent Transportation Systems (ITSC)*, 2023, pp. 2868–2873. DOI: [10.1109/ITSC57777.2023.10421960](https://doi.org/10.1109/ITSC57777.2023.10421960).
- [10] W. Jiang, Y. Ren, Y. Liu, and J. Leng, “Artificial neural networks and deep learning techniques applied to radar target detection: A review,” *Electronics*, vol. 11, no. 1, 2022, ISSN: 2079-9292. DOI: [10.3390/electronics11010156](https://doi.org/10.3390/electronics11010156). [Online]. Available: <http://www.mdpi.com/2079-9292/11/1/156>.
- [11] S. Abrecht, A. Hirsch, S. Raafatnia, and M. Woehrle, *Deep learning safety concerns in automated driving perception*, 2024. arXiv: [2309.03774](https://arxiv.org/abs/2309.03774) [cs.LG]. [Online]. Available: <https://arxiv.org/abs/2309.03774>.

- [12] Y. Zhou, L. Liu, H. Zhao, M. López-Benítez, L. Yu, and Y. Yue, “Towards deep radar perception for autonomous driving: Datasets, methods, and challenges,” *Sensors*, vol. 22, no. 11, p. 4208, 2022. DOI: [10.3390/s22114208](https://doi.org/10.3390/s22114208). [Online]. Available: <https://doi.org/10.3390/s22114208>.
- [13] F. J. Abdu, Y. Zhang, M. Fu, Y. Li, and Z. Deng, “Application of deep learning on millimeter-wave radar signals: A review,” *Sensors*, vol. 21, no. 6, 2021, ISSN: 1424-8220. DOI: [10.3390/s21061951](https://doi.org/10.3390/s21061951). [Online]. Available: <https://www.mdpi.com/1424-8220/21/6/1951>.
- [14] A. Pálffy et al., *RaDelft Dataset: A large-scale, real-life, and multi-sensor automotive dataset*, 4TU.ResearchData, 2024. DOI: [10.4121/4E277430-E562-4A7A-ADFE-30B58D9A5F0A](https://doi.org/10.4121/4E277430-E562-4A7A-ADFE-30B58D9A5F0A). [Online]. Available: <https://doi.org/10.4121/4E277430-E562-4A7A-ADFE-30B58D9A5F0A>.
- [15] S. M. Patole, M. Torlak, D. Wang, and M. Ali, “Automotive radars: A review of signal processing techniques,” *IEEE Signal Processing Magazine*, vol. 34, no. 2, pp. 22–35, 2017. DOI: [10.1109/MSP.2016.2628914](https://doi.org/10.1109/MSP.2016.2628914).
- [16] P. P. Gandhi and S. A. Kassam, “Analysis of cfar processors in nonhomogeneous background,” *IEEE Transactions on Aerospace and Electronic Systems*, vol. 24, no. 4, pp. 427–445, 1988. DOI: [10.1109/7.7185](https://doi.org/10.1109/7.7185).
- [17] H. M. Finn, “Adaptive detection mode with threshold control as a function of spatially sampled clutter-level estimates,” *Rca Rev.*, vol. 29, pp. 414–465, 1968.
- [18] V. G. Hansen and J. H. Sawyers, “Detectability loss due to “greatest of” selection in a cell-averaging cfar,” *IEEE Transactions on Aerospace and Electronic Systems*, vol. AES-16, no. 1, pp. 115–118, 1980. DOI: [10.1109/TAES.1980.308885](https://doi.org/10.1109/TAES.1980.308885).
- [19] G. Trunk, “Range resolution of targets using automatic detectors,” *IEEE Transactions on Aerospace and Electronic Systems*, vol. AES-14, no. 5, pp. 750–755, 1978. DOI: [10.1109/TAES.1978.308625](https://doi.org/10.1109/TAES.1978.308625).
- [20] H. Rohling, “New cfar-processor based on an ordered statistic,” in *International Radar Conference*, 1985, pp. 271–275.
- [21] H. Rohling, “Radar cfar thresholding in clutter and multiple target situations,” *IEEE Transactions on Aerospace and Electronic Systems*, vol. AES-19, no. 4, pp. 608–621, 1983. DOI: [10.1109/TAES.1983.309350](https://doi.org/10.1109/TAES.1983.309350).
- [22] J. T. Rickard and G. M. Dillard, “Adaptive detection algorithms for multiple-target situations,” *IEEE Transactions on Aerospace and Electronic Systems*, vol. AES-13, no. 4, pp. 338–343, 1977. DOI: [10.1109/TAES.1977.308466](https://doi.org/10.1109/TAES.1977.308466).
- [23] J. A. Ritcey, “Performance analysis of the censored mean-level detector,” *IEEE Transactions on Aerospace and Electronic Systems*, vol. AES-22, no. 4, pp. 443–454, 1986. DOI: [10.1109/TAES.1986.310780](https://doi.org/10.1109/TAES.1986.310780).
- [24] R. G. Zefreh, M. R. Taban, M. M. Naghsh, and S. Gazor, “Robust cfar detector based on censored harmonic averaging in heterogeneous clutter,” *IEEE Transactions on Aerospace and Electronic Systems*, vol. 57, no. 3, pp. 1956–1963, 2021. DOI: [10.1109/TAES.2020.3046050](https://doi.org/10.1109/TAES.2020.3046050).

- [25] A. Safa et al., “A low-complexity radar detector outperforming os-cfar for indoor drone obstacle avoidance,” *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 14, pp. 9162–9175, 2021. DOI: [10.1109/JSTARS.2021.3107686](https://doi.org/10.1109/JSTARS.2021.3107686).
- [26] J. Yoon, S. Lee, S. Lim, and S.-C. Kim, “High-density clutter recognition and suppression for automotive radar systems,” *IEEE Access*, vol. 7, pp. 58 368–58 380, 2019. DOI: [10.1109/ACCESS.2019.2914267](https://doi.org/10.1109/ACCESS.2019.2914267).
- [27] C.-H. Lin, Y.-C. Lin, Y. Bai, W.-H. Chung, T.-S. Lee, and H. Huttunen, “Dl-cfar: A novel cfar target detection method based on deep learning,” in *2019 IEEE 90th Vehicular Technology Conference (VTC2019-Fall)*, 2019, pp. 1–6. DOI: [10.1109/VTCFall1.2019.8891420](https://doi.org/10.1109/VTCFall1.2019.8891420).
- [28] M. M. Akhtar, Y. Li, W. Cheng, L. Dong, and Y. Tan, “OHDL: Radar target detection using optimized hybrid deep learning for automotive FMCW,” *Digital Signal Processing*, vol. 158, p. 104962, 2025. DOI: [10.1016/j.dsp.2024.104962](https://doi.org/10.1016/j.dsp.2024.104962). [Online]. Available: <https://doi.org/10.1016/j.dsp.2024.104962>.
- [29] C. Blackman, M. Uney, and Y. Zheng, “Detecting and measuring extended radar targets using deep learning,” in *2025 IEEE Radar Conference (RadarConf25)*, 2025, pp. 135–140. DOI: [10.1109/RadarConf2559087.2025.11204968](https://doi.org/10.1109/RadarConf2559087.2025.11204968).
- [30] Y. Cheng, J. Su, M. Jiang, and Y. Liu, “A novel radar point cloud generation method for robot environment perception,” *IEEE Transactions on Robotics*, vol. 38, no. 6, pp. 3754–3773, 2022. DOI: [10.1109/TR0.2022.3185831](https://doi.org/10.1109/TR0.2022.3185831). [Online]. Available: <https://doi.org/10.1109/TR0.2022.3185831>.
- [31] Z. Han et al., “DenserRadar: A 4d millimeter-wave radar point cloud detector based on dense LiDAR point clouds,” in *2024 IEEE 27th International Conference on Intelligent Transportation Systems (ITSC)*, 2024, pp. 930–936. DOI: [10.1109/ITSC58415.2024.10919654](https://doi.org/10.1109/ITSC58415.2024.10919654). [Online]. Available: <https://doi.org/10.1109/ITSC58415.2024.10919654>.
- [32] M. Schwarz, N. Kern, and B. Loesch, “Improving a deep learning automotive radar detector using perception task loss,” in *2025 IEEE Radar Conference (RadarConf25)*, 2025, pp. 1653–1658. DOI: [10.1109/RadarConf2559087.2025.11204994](https://doi.org/10.1109/RadarConf2559087.2025.11204994).
- [33] S. Sun, A. P. Petropulu, and H. V. Poor, “MIMO radar for advanced driver-assistance systems and autonomous driving: Advantages and challenges,” *IEEE Signal Processing Magazine*, vol. 37, no. 4, pp. 98–117, 2020. DOI: [10.1109/MSP.2020.2978507](https://doi.org/10.1109/MSP.2020.2978507). [Online]. Available: <https://doi.org/10.1109/MSP.2020.2978507>.
- [34] D. Zoeke and A. Ziroff, “Phase migration effects in moving target localization using switched mimo arrays,” in *2015 European Radar Conference (EuRAD)*, 2015, pp. 85–88. DOI: [10.1109/EuRAD.2015.7346243](https://doi.org/10.1109/EuRAD.2015.7346243).
- [35] C. M. Schmid, R. Feger, C. Pfeffer, and A. Stelzer, “Motion compensation and efficient array design for tdma fmcw mimo radar systems,” in *2012 6th European Conference on Antennas and Propagation (EUCAP)*, 2012, pp. 1746–1750. DOI: [10.1109/EuCAP.2012.6206605](https://doi.org/10.1109/EuCAP.2012.6206605).

- [36] X. Li, X. Wang, Q. Yang, and S. Fu, “Signal processing for TDM MIMO FMCW millimeter-wave radar sensors,” *IEEE Access*, vol. 9, pp. 167 959–167 971, 2021. DOI: [10.1109/ACCESS.2021.3137387](https://doi.org/10.1109/ACCESS.2021.3137387).
- [37] D. Brodeski, I. Bilik, and R. Giryas, “Deep radar detector,” in *2019 IEEE Radar Conference (RadarConf)*, 2019, pp. 1–6. DOI: [10.1109/RADAR.2019.8835792](https://doi.org/10.1109/RADAR.2019.8835792). [Online]. Available: <https://doi.org/10.1109/RADAR.2019.8835792>.
- [38] D.-H. Paek, S.-H. Kong, and K. T. Wijaya, *K-radar: 4d radar object detection for autonomous driving in various weather conditions*, 2023. arXiv: [2206.08171](https://arxiv.org/abs/2206.08171) [cs.CV]. [Online]. Available: <https://arxiv.org/abs/2206.08171>.
- [39] F. E. Nowruzi et al., *Deep open space segmentation using automotive radar*, 2020. DOI: [10.48550/arXiv.2004.03449](https://doi.org/10.48550/arXiv.2004.03449). arXiv: [2004.03449](https://arxiv.org/abs/2004.03449) [cs.CV]. [Online]. Available: <https://arxiv.org/abs/2004.03449>.
- [40] O. Schumann, J. Lombacher, M. Hahn, C. Wöhler, and J. Dickmann, “Scene understanding with automotive radar,” *IEEE Transactions on Intelligent Vehicles*, vol. 5, no. 2, pp. 188–203, 2020. DOI: [10.1109/TIV.2019.2955853](https://doi.org/10.1109/TIV.2019.2955853). [Online]. Available: <https://doi.org/10.1109/TIV.2019.2955853>.
- [41] Y. Wang, Z. Jiang, Y. Li, J.-N. Hwang, G. Xing, and H. Liu, “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*, vol. 15, no. 4, pp. 954–967, 2021. DOI: [10.1109/JSTSP.2021.3058895](https://doi.org/10.1109/JSTSP.2021.3058895). [Online]. Available: <https://doi.org/10.1109/JSTSP.2021.3058895>.
- [42] D. Kellner, M. Barjenbruch, J. Klappstein, J. Dickmann, and K. Dietmayer, “Tracking of extended objects with high-resolution doppler radar,” *IEEE Transactions on Intelligent Transportation Systems*, vol. 17, no. 5, pp. 1341–1353, 2016. DOI: [10.1109/TITS.2015.2501759](https://doi.org/10.1109/TITS.2015.2501759). [Online]. Available: <https://doi.org/10.1109/TITS.2015.2501759>.
- [43] M. Hassan, F. Fioranelli, A. Yarovoy, and S. Ravindran, “Radar multi object tracking using DNN features,” in *2023 IEEE International Radar Conference (RADAR)*, 2023, pp. 1–6. DOI: [10.1109/RADAR54928.2023.10371032](https://doi.org/10.1109/RADAR54928.2023.10371032). [Online]. Available: <https://doi.org/10.1109/RADAR54928.2023.10371032>.
- [44] Z. Yang, Z. Zhang, and C. Guo, “Robust improved variance index cfar algorithm for application in various complex environments,” *Signal, Image and Video Processing*, vol. 19, no. 11, p. 895, 2025.
- [45] D. Varciu, V.-D. Davidovici, A. G. Ham, and C. Nafornta, “Performance optimization of trimmed-mean cfar detectors in multi-target automotive scenarios,” in *2026 16th International Conference on Communications (COMM)*, IEEE, 2026, pp. 230–235.
- [46] H. Urkowitz, “Energy detection of unknown deterministic signals,” *Proceedings of the IEEE*, vol. 55, no. 4, pp. 523–531, 1967. DOI: [10.1109/PROC.1967.5573](https://doi.org/10.1109/PROC.1967.5573).
- [47] I. Roldan, A. Palffy, J. F. Kooij, D. M. Gavrila, F. Fioranelli, and A. Yarovoy, *RaDelft-Dataset: A large-scale, multi-sensory automotive radar dataset*, <https://github.com/RaDelft/RaDelft-Dataset>, 2024.

- [48] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, “Feature pyramid networks for object detection,” in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 936–944. DOI: [10.1109/CVPR.2017.106](https://doi.org/10.1109/CVPR.2017.106). [Online]. Available: <https://doi.org/10.1109/CVPR.2017.106>.
- [49] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778. DOI: [10.1109/CVPR.2016.90](https://doi.org/10.1109/CVPR.2016.90). [Online]. Available: <https://doi.org/10.1109/CVPR.2016.90>.
- [50] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, “Focal loss for dense object detection,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 2980–2988. DOI: [10.1109/ICCV.2017.324](https://doi.org/10.1109/ICCV.2017.324).
- [51] J. Domhof, J. F. P. Kooij, and D. M. Gavrila, “A joint extrinsic calibration tool for radar, camera and LiDAR,” *IEEE Transactions on Intelligent Vehicles*, vol. 6, no. 3, pp. 571–582, 2021. DOI: [10.1109/TIV.2021.3065208](https://doi.org/10.1109/TIV.2021.3065208). [Online]. Available: <https://doi.org/10.1109/TIV.2021.3065208>.
- [52] H. Fan, H. Su, and L. J. Guibas, “A point set generation network for 3d object reconstruction from a single image,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 605–613.
- [53] Lyken, *Pytorch-opcounter: Count the flops / macs of your pytorch model*, <https://github.com/Lyken17/pytorch-OpCounter>, 2019.
- [54] O. Ronneberger, P. Fischer, and T. Brox, “U-net: Convolutional networks for biomedical image segmentation,” in *International Conference on Medical image computing and computer-assisted intervention*, Springer, 2015, pp. 234–241.
- [55] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, “Unet++: A nested u-net architecture for medical image segmentation,” *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, pp. 3–11, 2018.
- [56] H. Huang et al., “Unet 3+: A full-scale connected unet for medical image segmentation,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2020, pp. 1055–1059.