LiDAR-Supervised First-Return Occupancy Grid Mapping for Automotive Radar on the RaDelft Dataset

Version of August 16, 2026

Yiwei Pang

LiDAR-Supervised First-Return Occupancy Grid Mapping for Automotive Radar on the RaDelft Dataset

THESIS

submitted in partial fulfillment of the
requirements for the degree of

MASTER OF SCIENCE

in

ELECTRICAL ENGINEERING

by

Yiwei Pang

TU Delft

Microwave Sensing, Signals and Systems (MS3) Group
Department of Microelectronics
Faculty EEMCS, Delft University of Technology
Delft, the Netherlands
www.ewi.tudelft.nl

LiDAR-Supervised First-Return Occupancy Grid Mapping for Automotive Radar on the RaDelft Dataset

Author: Yiwei Pang
Student id: 6149421

Abstract

Automotive radar is robust to adverse weather and illumination conditions, but its sparse, noisy, and not always spatially accurate measurements make dense occupancy estimation challenging. This thesis investigates light detection and ranging (LiDAR)-supervised radar-based first-return occupancy grid mapping on a polar range–azimuth grid. Radar is used as the sole input modality during inference, while synchronized LiDAR point clouds are processed into radar-aligned supervision representing visible free space, the nearest occupied boundary, and the unknown region behind it.

The proposed framework formulates occupancy estimation as an azimuth-wise first-return prediction problem rather than independent cell-wise classification. The complete range–Doppler–azimuth (RDA) radar cube is processed using a Range–Azimuth-Preserving Doppler Spectrum Encoder with parallel multi-scale Doppler-only branches, followed by a U-Net 3+ prediction backbone. The predicted first-return distribution is rendered into a visibility-consistent three-state occupancy grid. Training combines a Lovasz-based occupancy objective with a smoothed cumulative distribution function (CDF)-L1 boundary loss. Radar-aware azimuth weighting and causal temporal feature fusion are further investigated as controlled extensions.

Experiments are conducted on the RaDelft dataset using a scene-level train, validation, and test split. The proposed single-frame baseline achieves a free-space intersection over union (IoU) of 0.79, an unfree-space IoU of 0.82, a mean IoU (mIoU) of 0.80, and a first-return mean absolute error (MAE) of 5.38 m. It outperforms an odometry-compensated multi-frame inverse sensor model and a RaDelft-adapted PolarNet reference taken from the literature. Additional experiments show that retaining the complete Doppler spectrum

is more effective than fixed mean or maximum aggregation. The proposed Doppler encoder provides accuracy comparable to a reference encoder from the literature, but at lower computational cost.

Overall, the results demonstrate that structured first-return prediction from full Doppler-resolved radar measurements provides an effective approach to LiDAR-supervised polar occupancy mapping, with the single-frame full-RDA model offering the most favourable accuracy–efficiency balance among the evaluated full-RDA configurations.

Thesis Committee:

Chair:	dr. F. Fioranelli, Faculty EEMCS, TU Delft
University supervisor:	dr. F. Fioranelli, Faculty EEMCS, TU Delft
Committee Member:	dr. J.F.P. (Julian) Kooij , Faculty Mechanical Engineering, TU Delft
Advisor:	dr. S. Yuan, Faculty EEMCS, TU Delft

Preface

Completing this thesis has been a long and rewarding journey. Studying at TU Delft was the first time I had travelled so far from home on my own. This experience has taught me not only academically, but also how to adapt to a new environment and continue moving forward through uncertainty and challenges.

First and foremost, I would like to thank my family. Their unconditional support, encouragement, and trust have accompanied me throughout my studies, despite the distance between us. I feel especially fortunate that, during this period, I was able to spend a month travelling through Europe with my parents. The time we shared together has become one of my most cherished memories from my years in Europe. I am deeply grateful for everything they have done for me and for always giving me the confidence to pursue my own path.

I would also like to express my sincere gratitude to my supervisors, Dr Francesco Fioranelli and Dr Sen Yuan, for their guidance and support throughout this thesis. Their feedback, discussions, and encouragement helped me navigate the technical challenges and develop a clearer understanding of how to approach research independently. I greatly appreciate the time and patience they dedicated to this work.

Finally, I would like to thank all the friends who have been part of my life during this journey. You have been an important source of emotional support throughout both my studies and everyday life. Cooking and sharing good food together, spending time having fun, and studying side by side gave me many moments of happiness and relief along the way. These seemingly ordinary moments became an important reason why I was able to keep going. I am deeply grateful for your companionship and for helping me maintain a positive state of mind throughout my studies.

Finishing this thesis marks the end of an important chapter of my life. I am grateful not only for what I have learned academically, but also for the people I met and the experiences I gained along the way.

Yiwei Pang
Delft, the Netherlands
August 16, 2026

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background and Motivation

Reliable perception of the surrounding environment is a fundamental requirement for autonomous driving. An autonomous vehicle must continuously reason about the spatial layout around itself in order to support safe navigation, motion planning, and collision avoidance. This requires not only detecting individual objects, but also understanding which spatial regions are visible, free, occupied, or not directly observable.

Occupancy grid mapping (OGM) provides a structured representation for this task by discretizing the environment into spatial cells and assigning an occupancy state to each cell. Classical occupancy grid mapping has been widely used in mobile robotics to represent uncertain spatial information from range-sensor measurements [21, 8]. In autonomous driving, occupancy grids remain useful because they provide information about both obstacles and available free space, which is essential for scene understanding and planning [23]. Compared with object-level representations such as bounding boxes or sparse detections, an occupancy grid can describe irregular obstacle shapes, partially visible structures, and regions that do not belong to predefined object categories.

Automotive perception systems commonly rely on multiple sensing modalities, including cameras, LiDAR, and radar [36, 3]. Cameras provide dense appearance information, but their performance can be affected by illumination changes, glare, and adverse weather conditions. LiDAR provides accurate three-dimensional geometric measurements of visible scene surfaces and is therefore highly useful for constructing spatial maps, but its measurements can also be affected by weather-related scattering and occlusion [7]. In addition, LiDAR deployment is often associated with higher hardware cost and integration complexity compared with radar-based sensing [44]. Radar, in contrast, directly measures range and Doppler responses and can provide long-range sensing and velocity-related information with greater robustness to adverse weather and poor lighting conditions [33, 36]. These properties make radar an attractive modality for automotive occupancy estimation, especially in scenarios

where optical sensors may be degraded or visually ambiguous.

Despite these advantages, radar-based occupancy estimation remains challenging because radar measurements are sparse and noisy and are affected by clutter, sidelobes, multipath, limited angular resolution, and measurement uncertainty [33]. Conventional radar mapping typically relies on detection-based preprocessing and hand-crafted inverse sensor models, whereas learning-based approaches seek to infer spatial representations more directly from radar measurements. A detailed review of these approaches is provided in Chapter 2.

Learning-based radar occupancy estimation further requires suitable geometric supervision. LiDAR provides accurate measurements of visible surfaces and obstacle boundaries and can therefore be used to construct radar-aligned reference labels. However, radar and LiDAR observe the environment through different sensing mechanisms, leading to discrepancies in visibility, spatial response, and the locations of observed object surfaces [39, 17]. Against this background, this thesis adopts a first-return formulation centred on the nearest observable boundary along each azimuth direction. A formal definition of this representation is provided in Section 1.2.

In addition to range and azimuth information, automotive radar measurements provide a Doppler-resolved response at each spatial location. However, radar-based perception pipelines often compress the Doppler dimension before spatial prediction, for example through mean or maximum aggregation, thereby discarding spectral information before it can be exploited by the learning model. This motivates investigating whether retaining and learning from the complete Doppler-resolved measurement can benefit radar-based first-return occupancy estimation.

The RaDelft dataset provides a suitable basis for this study because it contains synchronized high-resolution automotive radar, LiDAR, camera, and ego-motion measurements collected in real-world driving scenarios [29]. It has been used in different research [34, 43, 42, 11] where access to preprocessed radar data, i.e., before point cloud generation, is desired. Its synchronized sensor measurements enable LiDAR-derived geometric supervision to be aligned with Doppler-resolved radar observations, while retaining radar as the only input modality during inference.

Accordingly, this thesis investigates LiDAR-supervised radar-based first-return occupancy estimation from full range–Doppler–azimuth measurements on a radar-aligned polar grid using RaDelft dataset.

1.2 Problem Definition and Terminology

This thesis investigates LiDAR-supervised radar-based first-return occupancy grid mapping on a radar-aligned polar range–azimuth grid. Radar is the only input modality used during inference, while synchronized LiDAR point clouds are processed to generate reference occupancy labels for training and evaluation. Formally, the core problem is to learn a mapping between the radar input X_{RAD} and the corresponding LiDAR-derived reference occupancy map Y_{LiDAR}.

For consistency with the subsequent methodology and experimental chapters, this sensor-derived reference is referred to as the *ground-truth map* throughout the thesis. This terminology does not imply a perfect sensor-independent description of the scene, since the labels remain subject to LiDAR visibility, sampling, and cross-sensor alignment limitations. The learned mapping is thus expressed as

$$\hat{Y} = f_{\Phi}(X_{\mathrm{RAD}}), \tag{1.1}$$

where f_{Φ} denotes the radar-to-OGM model with learnable parameters Φ, and $\hat{Y}$ denotes the rendered polar occupancy prediction. The detailed output parameterization and occupancy-rendering procedure are presented in Chapter 3.

To fully contextualize this mapping, it is essential to define the structures of the target output space, its evaluation criteria, and the input modality. The target output space—referred to as the *first-return occupancy grid map*—provides a structured spatial representation of the environment. For each azimuth direction, range cells before the nearest visible obstacle boundary are labelled as free, the boundary cell is labelled as occupied, and cells behind the boundary are labelled as unknown because they are not directly observable from the current measurement. If no first return is present within the valid sensing range, the corresponding azimuth direction, i.e. a column in the chosen 2D data format, is represented as a no-hit column.

When evaluating these occupancy predictions, the occupied and unknown states are grouped into an *unfree* class representing regions that cannot be interpreted as confidently traversable by the ego-vehicle. Evaluation considers both occupancy overlap and the geometric accuracy of the estimated first-return boundary. The corresponding metrics are defined in Chapter 3.

Finally, regarding the input space X_{RAD}, the primary modality considered in this thesis is the complete range–Doppler–azimuth radar cube, which retains the Doppler-resolved measurement at each range–azimuth location. Fixed Doppler-compressed range–azimuth representations are considered as comparison configurations. The radar preprocessing, feature encoding, prediction architecture, and training formulation are described in Chapter 3.

1.3 Research Questions

The objective of this thesis is to investigate LiDAR-supervised radar-only first-return occupancy estimation, with particular attention to the formulation of the prediction task, the representation and encoding of Doppler-resolved radar measurements, and the design of the learning objective used to supervise the proposed model architectures. Accordingly, the following research questions are addressed:

RQ1 How can radar-only occupancy estimation be formulated and supervised, through an appropriate learning objective, so that the predicted polar occupancy grid inherently follows first-return visibility constraints?

RQ2 How can the Doppler-resolved information in the complete range–Doppler–azimuth radar cube be encoded while preserving the range–azimuth geometry required for first-return occupancy prediction, and how does this compare with fixed Doppler-compressed representations?

RQ3 How can the loss function formulation jointly supervise occupancy structure and first-return boundary localization, and to what extent can radar-aware supervision weighting and causal temporal context further improve the intended single-frame framework?

1.4 Thesis Contributions

To address these research questions, this thesis makes the following contributions:

- **Visibility-consistent first-return occupancy formulation and supervision.** Radar occupancy estimation is formulated as an azimuth-wise first-return or no-hit prediction problem rather than as independent classification of individual range–azimuth cells. A corresponding LiDAR-based supervision pipeline is developed to construct radar-aligned polar targets, in which the nearest observable boundary determines the visible free region, occupied first-return cell, and unknown region behind it. Steps for ground removal, local ground-surface estimation, height filtering, radar-grid projection, and validity masking are incorporated to obtain geometrically consistent supervision.
- **Range–azimuth-preserving Doppler-spectrum encoding.** A Doppler Spectrum Encoder is developed to process the complete range–Doppler–azimuth radar cube without collapsing the Doppler dimension through fixed aggregation. Parallel Doppler-only convolutional branches extract spectral features at multiple scales while preserving the range–azimuth geometry, leaving spatial reasoning to the subsequent two-dimensional prediction backbone. The resulting representation is evaluated against fixed Doppler-compressed inputs and alternative Doppler encoding strategies.
- **Task-aligned learning for structured first-return prediction.** A loss function formulation is developed that combines occupancy-oriented supervision of the rendered grid with direct supervision of the first-return boundary distribution. This jointly targets occupancy overlap and nearest-boundary localization. Building on the resulting single-frame framework, Radar-Aware Weighting and Causal Temporal Feature Fusion are further investigated as controlled extensions to assess the effects of azimuth-dependent supervision reliability and short-term radar context.

The principal contribution of this thesis therefore lies in the integrated formulation of LiDAR-supervised radar occupancy estimation as a structured first-return prediction problem, together with Doppler-resolved radar encoding and task-aligned

supervision. The individual use of Doppler information, multi-scale convolution, LiDAR supervision, or temporal fusion is not claimed as novel in isolation. The main results of this thesis are being prepared for submission as a journal paper to the *IEEE Transactions on Radar Systems*.

1.5 Thesis Outline

Chapter 2 reviews prior work relevant to radar-based occupancy and free-space estimation. It first discusses traditional inverse sensor models and learning-based radar occupancy approaches, with particular attention to their prediction formulations and the use of LiDAR-derived supervision. It then examines radar input representations and Doppler encoding strategies, focusing on the trade-off between preserving Doppler-resolved measurement information and computational efficiency. The chapter concludes by summarizing the methodological gaps relevant to structured first-return prediction and range–azimuth-preserving Doppler encoding.

Chapter 3 presents the methodology used in this thesis. It describes the RaDelft dataset and sensor alignment, the LiDAR-based first-return label-generation procedure, the radar input representations, the Range–Azimuth-Preserving Doppler Spectrum Encoder, and the structured first-return prediction and occupancy-rendering framework. It further introduces the radar-aware weighting and causal temporal-fusion extensions, defines the odometry-compensated multi-frame inverse sensor model and the RaDelft-adapted PolarNet model used as reference methods, and specifies the training and evaluation protocols.

Chapter 4 reports the experimental results. The proposed single-frame baseline is compared with approaches from the literature, i.e., an odometry-compensated multi-frame inverse sensor model and a RaDelft-adapted PolarNet reference. The chapter then evaluates alternative radar input representations and Doppler encoder designs, followed by the radar-aware weighting and causal temporal-fusion extensions. Quantitative results are complemented by qualitative comparisons, per-frame error analysis, representative failure cases, and an evaluation of computational efficiency.

Chapter 5 summarizes and interprets the main findings in relation to the research objectives and the reviewed literature. It discusses the role of Doppler-resolved encoding, the behaviour of the structured first-return formulation, the implications of LiDAR-derived supervision, and the benefits and limitations of the proposed extensions. The chapter concludes with the principal limitations of the study and directions for future research.

Chapter 2

Related Work

This chapter reviews prior work relevant to radar-based first-return occupancy estimation. It first examines traditional inverse sensor models and learning-based approaches for radar occupancy and free-space estimation, with particular attention to their input representations, supervision strategies, and output formulations. It then reviews radar input representations and Doppler encoding methods, focusing on how existing approaches balance measurement preservation, spatial structure, and computational efficiency. These two lines of work establish the context for the methodological choices investigated in this thesis.

2.1 Radar-Based Occupancy and Free-Space Estimation

Object-centric methods represent known traffic participants as discrete instances, whereas occupancy grids provide a scene-centric description of free, occupied, and unobserved space [40, 8]. Occupancy representations therefore complement object detection and provide information directly relevant to navigation and motion planning for autonomous vehicles [31].

2.1.1 Traditional Inverse Sensor Models

Traditional radar occupancy mapping generally combines a hand-crafted inverse sensor model with recursive probabilistic updating. The inverse sensor model converts radar measurements into free- and occupied-space evidence, which is accumulated across observations using Bayesian or evidential update rules. Existing approaches differ mainly in how they represent radar uncertainty, interpret missing detections, and incorporate scene dynamics.

Early radar-specific methods focused on converting detection-level measurements into static occupancy evidence. Werber et al. [38] defined spatial occupancy likelihoods around automotive radar measurements of range, azimuth, and amplitude. Degerman et al. [5] extended statistical radar occupancy mapping to three dimen-

sions and related the measured signal-to-noise ratio to detection probability through an explicit radar observation model. Slutsky and Dobkin [32] addressed the interpretation of non-detected regions through a dual inverse sensor model. Its positive component assigns occupied-space evidence around radar detections, whereas its negative component uses the absence of detections along a viewing direction to infer free-space evidence. Although these approaches account for important radar characteristics, the conversion from measurements to grid evidence remains manually specified.

Subsequent work increased the flexibility and state complexity of model-based occupancy mapping. Porębski [27] proposed a customizable inverse sensor model that separates grid-cell selection from the assignment of free- and occupied-space evidence and supports both Bayesian and Dempster–Shafer representations. For dynamic traffic scenes, Coué et al. [4] introduced Bayesian Occupancy Filtering, in which occupancy and motion-related states are recursively predicted and updated. Haag et al. [10] later proposed the UNIFY framework, which represents static occupancy and velocity in separate Bayesian grid layers while using radar-specific observation models to handle measurement ambiguities. These extensions provide more flexible and dynamic representations, but require increasingly complex observation, transition, and uncertainty models.

Despite their interpretability and probabilistic consistency, traditional inverse sensor models remain dependent on hand-designed likelihood functions, detection thresholds, uncertainty distributions, and sensor-specific assumptions. Such components may require substantial retuning across radar configurations and operating environments, while sparse detections, limited angular resolution, multipath, and ambiguous scattering can produce fragmented occupancy evidence.

Classical ray-based mapping already represents a visibility relationship between free space, a measured surface, and the unobserved region behind it [8]. The limitation of radar-specific inverse sensor models is therefore more specific. Radar measurements may contain sparse, ambiguous, or multiple responses along a similar viewing direction, while model-based approaches generally convert these measurements into cell-wise occupancy evidence through predefined observation rules. Explicitly representing the nearest observable boundary or a no-hit state as the primary prediction variable has received comparatively less attention in radar occupancy estimation.

2.1.2 Learning-Based Radar Occupancy Estimation

Learning-based radar occupancy methods replace hand-crafted inverse sensor models with data-driven mappings from radar observations to dense spatial representations. Existing approaches differ mainly in their radar input format, output formulation, and strategy for transferring supervision from higher-resolution sensors.

Early methods primarily used radar detections or two-dimensional radar representations. Sless et al. [31] rasterized temporally aggregated radar clusters into a bird's-eye-view (BEV) grid and formulated free-, occupied-, and unobserved-space

estimation as semantic segmentation using LiDAR-derived labels. Although the learned model outperformed classical Bayesian mapping, the use of clustered detections and multi-frame aggregation limits the measurement information available to the network and introduces dependence on temporal alignment. Weston et al. [39] instead learned an inverse sensor model from raw polar radar power returns using partial LiDAR occupancy labels and heteroscedastic uncertainty modeling[1]. However, the supervision remained spatially incomplete, and the output was still predicted through cell-wise segmentation.

Other work preserved the native polar geometry or addressed discrepancies between radar and LiDAR supervision. Nowruzi et al. [22] proposed PolarNet for open-space segmentation directly in the polar domain. Kung et al. [17] used polar sliding-window inference to extend prediction beyond the supervised range. Wei et al. [37] learned a single-frame polar inverse sensor model from sparse radar detections and subsequently combined the predicted measurement grid with radar Doppler velocities to construct a dynamic occupancy grid map. While these approaches improve radar alignment, polar-domain processing, or temporal efficiency, they primarily rely on two-dimensional inputs or introduce Doppler information only after the initial occupancy prediction.

Notably, cross-sensor supervision is itself a source of modelling uncertainty. Weston et al. [39] used partial LiDAR occupancy labels rather than treating all LiDAR observations as uniformly valid supervision. Kung et al. [17] further showed that the physical sensing discrepancy between radar and LiDAR can produce inconsistent targets: objects visible to LiDAR may be weakly observable or absent in radar, while radar may retain responses from regions that are occluded in LiDAR measurements. Their label preprocessing was therefore designed to reduce the influence of LiDAR observations that are not reliably supported by radar. These studies indicate that LiDAR-derived supervision should be selected and transformed with respect to radar visibility and sensing characteristics rather than transferred directly as an error-free occupancy target.

More recent methods have considered richer radar measurements and generative completion. Jin et al. [15] processed range–Doppler matrices using a transformer-based segmentation network. Their polar ground-truth maps combine camera-derived semantic free-space information with LiDAR-based vehicle localization, thereby retaining more pre-detection radar information than detection-level approaches. Nevertheless, their input does not jointly preserve the complete range–Doppler–azimuth structure, and occupied and unknown regions are represented by a common non-free class. Park et al. [26] formulated single-frame occupancy estimation as conditional image-to-image translation from a radar detection grid to a pixel-wise free-space map. This avoids temporal accumulation but remains dependent on the information retained after radar detection and thresholding.

[1] Heteroscedastic uncertainty refers to input-dependent predictive uncertainty, where the noise or uncertainty level is allowed to vary across observations or spatial locations rather than being assumed constant.

Table 2.1: Representative learning-based radar occupancy methods.

	Input and Supervision		Prediction	
Method	**Radar input**	**Supervision**	**Output representation**	**Prediction formulation**
Sless et al. [31]	Temporally aggregated radar clusters	LiDAR-derived occupancy labels	Cartesian three-state BEV	Cell-wise segmentation
Weston et al. [39]	Raw polar radar power returns	Partial LiDAR occupancy labels	Probabilistic occupancy grid	Probabilistic cell-wise prediction
PolarNet [22]	Polar radar representation	Open-space labels	Polar free-space mask	Pixel-wise segmentation
Kung et al. [17]	Polar radar image	Preprocessed LiDAR labels	Polar occupancy grid	Cell-wise prediction with sliding-window inference
Wei et al. [37]	Single-frame radar detections	LiDAR-derived polar grids	Polar grid and dynamic OGM	Cell-wise grid prediction; post-hoc Doppler fusion
Jin et al. [15]	Range–Doppler matrix	Camera- and LiDAR-derived labels	Binary polar free/non-free grid	Cell-wise segmentation
Park et al. [26]	Single-frame radar detection grid	Paired occupancy maps	Pixel-wise free-space map	Conditional image translation

Table 2.1 compares the reviewed representative learning-based radar occupancy methods in terms of their radar input, supervision source, output representation, and prediction formulation.

As summarized in Table 2.1, learning-based approaches reduce dependence on manually designed sensor models and support a range of occupancy and free-space representations. Several methods also exploit the native polar geometry of radar measurements. For example, Jin et al. [15] construct polar targets in which the first transition from free to non-free space denotes an obstacle and cells behind this transition are treated as unknown. However, occupied and unknown regions are represented by a common non-free class, while the prediction itself remains formulated at the cell level.

Among the reviewed learning-based methods, explicit parameterization of a single first-return or no-hit distribution for each azimuth remains uncommon. Many approaches also rely on detection-level inputs or two-dimensional radar representations that preserve only part of the original measurement structure. At the same time, LiDAR-derived supervision introduces discrepancies in sensing range, visibility, occlusion, spatial resolution, and material response. These observations identify two closely related modeling questions yet to answer: how occupancy supervision should be structured around the observable boundary, and how sufficient information from the multidimensional radar measurement should be retained for learning.

2.2 Radar Input Representations and Doppler Encoding

The choice of radar input representation determines which measurement information remains available to a learning-based perception model. Detection-level point clouds are compact and convenient for geometric processing, but are produced after thresholding and peak selection, which may suppress weak or spatially extended responses. Denser representations obtained after the fast Fourier transform (FFT) retain measurements over range, azimuth (and/or elevation), and Doppler, but require greater memory and computation [45, 6]. Existing methods therefore adopt different compromises between information preservation, computational efficiency, and compatibility with established network architectures.

A common strategy is to reduce the multidimensional radar spectrum to two-dimensional image-like representations. Range–azimuth maps retain the polar spatial layout used for detection, segmentation, or occupancy prediction, whereas range–Doppler maps preserve range and radial-velocity information without directly resolving azimuth. Fixed operations such as summation, averaging, or maximum selection can be used to collapse the remaining dimension. These operations enable the use of standard two-dimensional convolutional networks, but reduce an entire Doppler spectrum to one scalar at each range–azimuth location. Spectra with different peak widths, secondary components, or energy distributions may therefore produce similar aggregated responses.

Multi-view approaches retain additional information by processing several two-dimensional projections of the same radar tensor. Ouaknine et al. [24] separately encoded range–azimuth, range–Doppler, and azimuth–Doppler views before combining them in a shared latent representation. This strategy exploits complementary information from different coordinate pairs while avoiding the cost of processing the complete three-dimensional tensor at once. However, each branch still operates on a projected view and therefore does not preserve the joint range–azimuth–Doppler relationship within a single representation.

Other methods process local or complete radar tensors more directly. Major et al. [20] demonstrated vehicle detection from dense range–azimuth–Doppler measurements, showing the value of retaining the Doppler dimension for learning-based radar perception. Palffy et al. [25] extracted local three-dimensional radar blocks around detected targets and classified them using a convolutional network. Although these local cubes preserve the Doppler distribution around a road user, the approach depends on a preceding detection stage. RADDet [45] instead processes a complete range–azimuth–Doppler tensor using a ResNet-based backbone. Such full-tensor approaches retain the Doppler bins at the input, but their computational cost increases with tensor size and their generic backbones do not explicitly separate spectral encoding from spatial feature extraction.

Learned representations have also been introduced at earlier stages of the radar-processing pipeline. FFT-RadNet [28] avoids explicitly constructing a complete range–azimuth–Doppler tensor and instead learns to recover angular information

from a range–Doppler spectrum. This reduces the memory and computational cost associated with conventional angle estimation while supporting both vehicle detection and free-space segmentation.

Within the context of occupancy mapping, Doppler information has also been introduced after the spatial prediction stage. As discussed in Section 2.1.2, Wei et al. [37] combine a predicted polar measurement grid with radar Doppler velocities to construct a dynamic occupancy grid map. Doppler is therefore used as a post-hoc motion attribute rather than as a pre-detection spectrum jointly encoded for the initial occupancy prediction.

Roldan et al. [29] provide a particularly relevant architectural reference for Doppler encoding. Their DopplerEncoder applies two sequential three-dimensional convolutional layers followed by three-dimensional max pooling, transforming a two-channel range–azimuth–Doppler tensor into 64 range–azimuth feature channels. The resulting representation is then processed by a two-dimensional spatial backbone for dense radar detection. Their ablation study further showed that replacing the Doppler-resolved input with mean aggregation degraded detection performance, indicating that the Doppler distribution contains useful information beyond a fixed spatial projection.

This architecture is particularly relevant because it illustrates learned Doppler compression before two-dimensional spatial reasoning without first reducing the spectrum through a fixed aggregation operation. However, the encoder was developed for dense radar detection rather than occupancy estimation, and its three-dimensional convolutions jointly operate across Doppler and neighbouring spatial locations. This leaves open the question of how Doppler-resolved measurements can be encoded while more explicitly preserving the range–azimuth geometry required by a downstream occupancy model.

A different compromise between spectral preservation and data volume is adopted by RadarOcc [6]. The method processes a range–azimuth–elevation–Doppler tensor, but compresses the Doppler spectrum at each spatial location into a compact descriptor containing the three strongest power values and their indices, together with the spectral mean and standard deviation. This representation retains selected Doppler statistics at a fraction of the original data volume and was shown to outperform average pooling in most occupancy metrics. However, the retained statistics and number of spectral peaks are predefined rather than learned jointly with the downstream task.

Taken together, prior work retains Doppler information in different ways through multi-view projection, full-tensor processing, learned dimensional reduction, handcrafted spectral descriptors, or post-hoc velocity integration. These approaches illustrate a recurring trade-off between retaining the Doppler-resolved measurement and limiting the computational cost of the subsequent spatial network. Learned Doppler compression before two-dimensional spatial processing provides a particularly relevant compromise, while preserving range–azimuth geometry during this compression remains an open design choice for downstream occupancy estimation.

2.3 Summary

This chapter reviewed two lines of work relevant to radar-based first-return occupancy estimation. Traditional radar occupancy mapping provides interpretable and probabilistically grounded estimates, but commonly relies on hand-crafted inverse sensor models, detection thresholds, and predefined uncertainty assumptions. Learning-based methods reduce this dependence and support direct prediction of occupancy or free-space representations from radar measurements. However, among the reviewed approaches, occupancy is still predominantly formulated through cell-wise prediction, while explicit estimation of a first-return or no-hit variable for each azimuth remains uncommon. Cross-sensor supervision introduces an additional challenge because radar and LiDAR differ in visibility, spatial resolution, scattering behaviour, and the physical locations of their observed responses, thus making not straightforward to use LiDAR as a ground truth for radar.

The reviewed literature also shows that the choice of radar input representation determines how much Doppler-resolved information remains available to the model. Detection-level inputs and fixed two-dimensional projections reduce computational cost but discard part of the measured spectrum, whereas full-tensor processing preserves more information at substantially greater computational expense. Learned Doppler compression provides an intermediate solution by transforming the Doppler spectrum into range–azimuth feature channels before spatial processing, although the way in which spatial geometry is preserved during this operation varies across architectures.

In summary, these observations motivate two methodological directions examined in this thesis: structuring occupancy prediction around the nearest observable boundary along each azimuth, and retaining Doppler-resolved radar information while preserving the range–azimuth geometry required for spatial occupancy estimation. The following chapters describe the dataset, target generation procedure, network architecture, training objective, and evaluation methodology used to investigate these design choices.

Chapter 3

Methodology

This chapter presents the methodology for LiDAR-supervised radar-based first-return occupancy grid mapping. The chapter is organized into two main parts. Section 3.1 describes the proposed method, including LiDAR-based first-return target generation, radar representation and Doppler encoding, structured first-return prediction, the loss function, and the controlled extensions. Section 3.2 then defines the experimental setup, including the RaDelft dataset and sensor processing, reference methods, training configuration, and evaluation protocol.

3.1 Proposed Method

This section presents the proposed radar-based first-return occupancy mapping framework. Section 3.1.1 first provides an overview of the complete processing pipeline. Section 3.1.2 then describes the construction of radar-aligned first-return supervision from synchronized LiDAR measurements. Sections 3.1.3 and 3.1.4 present the radar input representations and Doppler encoding strategies, including the proposed Range–Azimuth-Preserving Doppler Spectrum Encoder. Section 3.1.5 defines the structured azimuth-wise first-return prediction and occupancy-rendering formulation. Section 3.1.6 presents the learning objective and the Radar-Aware Weighting extension. Section 3.1.7 then introduces the Causal Temporal Feature Fusion architecture for incorporating short-term radar context.

3.1.1 Framework Overview

A LiDAR-supervised radar-to-occupancy-grid mapping framework is developed to estimate first-return occupancy from automotive radar measurements. As shown in Figure 3.1, the framework contains a radar prediction branch and a LiDAR-based supervision branch. The radar branch receives Doppler-resolved radar measurements[1]

[1]Here, Doppler-resolved refers to the radar representation after Doppler FFT, where the responses across the Doppler bins are retained at each range–azimuth location rather than being collapsed into a Doppler-compressed range–azimuth map.

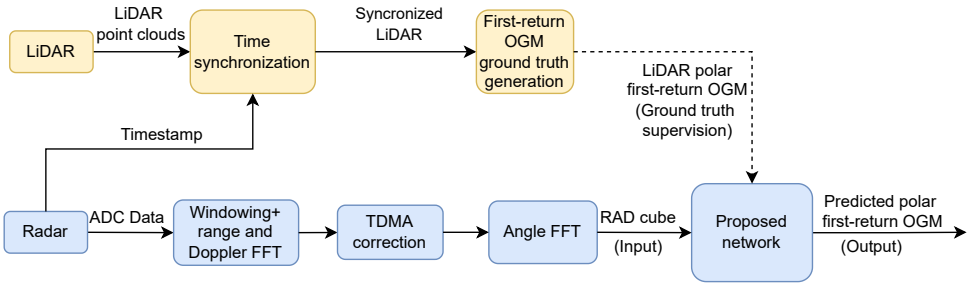

Figure 3.1: Overview of the proposed LiDAR-supervised radar-to-first-return OGM framework. Doppler-resolved radar measurements form the network input, while synchronized LiDAR measurements are processed into radar-aligned first-return occupancy targets for supervision. The predicted first-return representation is subsequently rendered into free, occupied, and unknown regions.

and predicts an azimuth-wise first-return or no-hit representation. The LiDAR branch processes synchronized point clouds into radar-aligned first-return targets. The predicted first-return representation is finally rendered into a three-state polar occupancy grid containing free space, the occupied first-return boundary, and the unknown region behind that boundary.

The proposed method follows four principal stages. *First*, synchronized LiDAR measurements are transformed into the radar coordinate system and processed to construct radar-aligned first-return occupancy targets. *Second*, the radar measurement is represented either by the complete range–Doppler–azimuth cube or, for comparison, by a fixed Doppler-compressed range–azimuth representation. For the full-RDA configuration, the proposed Range–Azimuth-Preserving Doppler Spectrum Encoder transforms the Doppler-resolved measurement into a range–azimuth feature representation while preserving the spatial layout of the target grid. *Third*, a U-Net 3+ based spatial backbone predicts an azimuth-wise first-return or no-hit distribution, from which the three-state occupancy grid is rendered. *Finally*, at the training stage the model is optimized using a task-aligned loss function that combines occupancy-oriented and first-return boundary supervision.

The proposed framework is centred on four methodological design choices. The first is the construction of LiDAR-derived first-return occupancy supervision, which converts sparse LiDAR obstacle evidence into a structured radar-aligned target containing visible free space, the nearest occupied boundary, and the unknown region behind it. The second is the formulation of occupancy estimation as an azimuth-wise first-return prediction problem rather than independent cell-wise classification. The third is the use of a range–azimuth-preserving Doppler-spectrum encoding strategy, which retains Doppler-resolved radar information while maintaining the spatial correspondence required by the target grid. The fourth is a task-aligned loss function that combines Lovasz-based supervision of the rendered occupancy structure with a

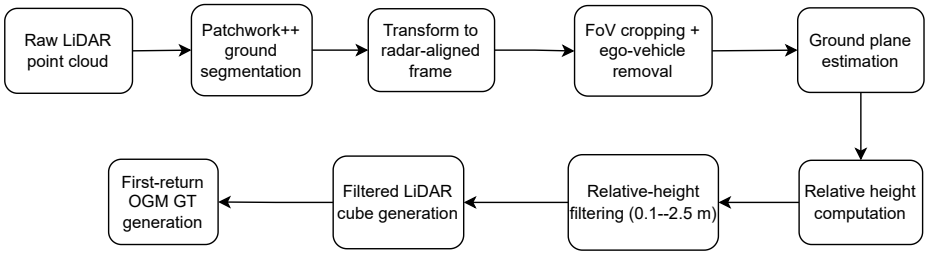

Figure 3.2: Pipeline for constructing radar-aligned first-return occupancy targets from synchronized LiDAR measurements.

smoothed CDF-L1 term that directly supervises the first-return boundary.

3.1.2 LiDAR-Based First-Return Target Generation

This section describes the construction of LiDAR-derived first-return supervision on the radar-aligned polar grid. Synchronized LiDAR measurements are used as the geometric reference because they provide substantially more accurate spatial information than the radar measurements used as network input. Rather than treating individual LiDAR points as independent occupied labels, the point cloud is processed to identify obstacle evidence relevant to the radar field of view and then converted into an azimuth-wise first-return target. The resulting representation distinguishes visible free space, the nearest observed obstacle boundary, and the region that is not directly observable behind that boundary.

The target-generation pipeline is illustrated in Figure 3.2. Raw LiDAR measurements are first separated into ground and non-ground components using Patchwork++ [19]. The resulting point sets are transformed into the radar-aligned coordinate frame, restricted to the radar field of view, and filtered to remove LiDAR returns originating from the ego-vehicle body and chassis. Local ground information is then used to evaluate obstacle height relative to the surrounding road surface, after which the retained obstacle points are projected onto the radar polar grid. The nearest valid obstacle evidence along each azimuth direction defines the first return from which the final occupancy target is constructed. Regions that do not provide reliable LiDAR supervision are excluded through the validity mask described later in this section.

Figure 3.3 shows a representative frame at several stages of the target-generation process for illustrative purposes. Panel (a) provides the corresponding camera view, while panels (b)–(d) illustrate the raw LiDAR measurements, their transformation into the radar-aligned frame, and the subsequent restriction to the radar field of view after ego-vehicle removal, respectively. The later stages show how ground segmentation and relative-height filtering progressively retain the obstacle evidence used for first-return supervision.

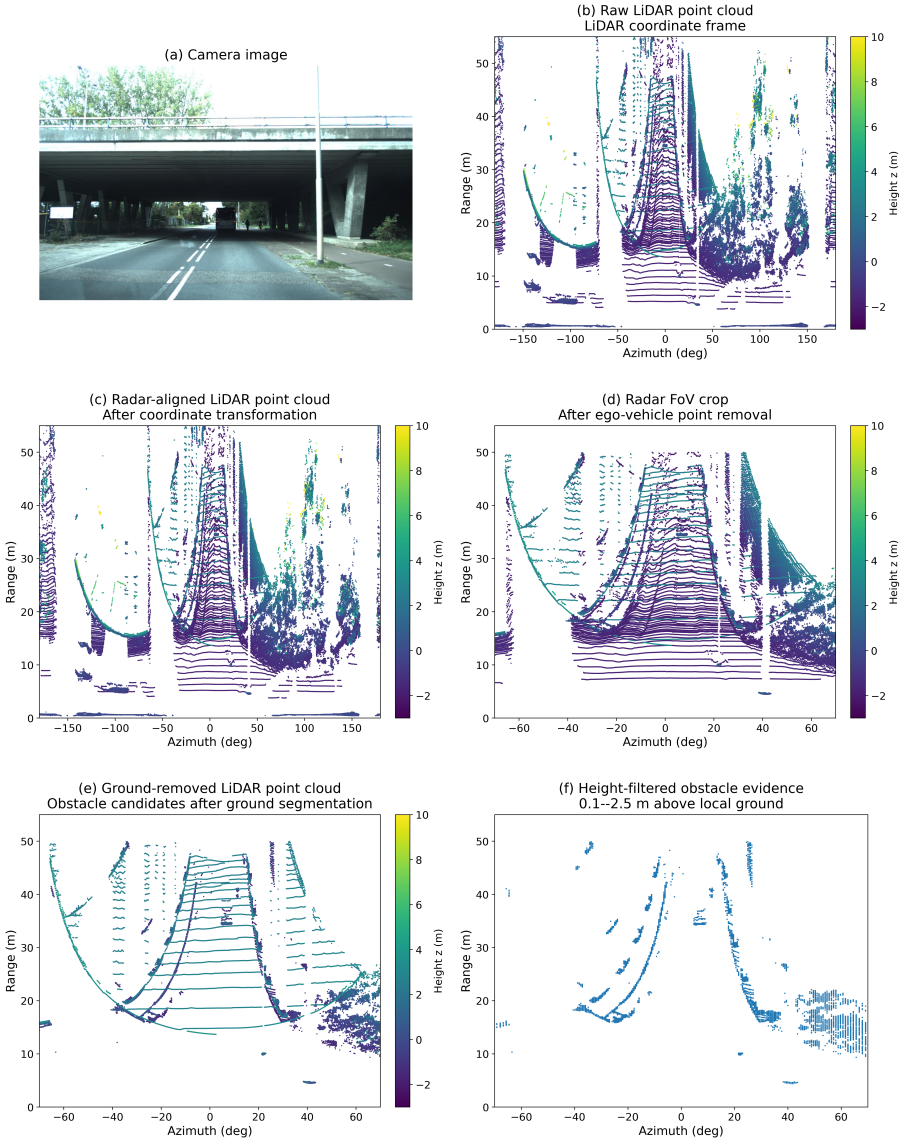

Figure 3.3: Representative LiDAR target-generation stages shown left to right, top to bottom: (a) camera reference, (b) raw LiDAR point cloud, (c) radar-aligned LiDAR point cloud, (d) radar-FoV crop after ego-vehicle removal, (e) ground-removed obstacle candidates, and (f) height-filtered obstacle evidence used for first-return supervision.

The following subsections describe the individual processing stages more in detail and the construction of the final first-return occupancy target.

Ground Removal Using Patchwork++

The raw LiDAR point cloud contains measurements from both the road surface and above-ground objects. Since road-surface returns should not be interpreted as occupied obstacle evidence, the point cloud is first separated into ground and non-ground components. As illustrated in Figure 3.3(e), this step distinguishes the road surface from the above-ground measurements used as candidate obstacle evidence.

Patchwork++ is used for ground segmentation [19]. The method partitions each LiDAR point cloud into ground and non-ground points using spatially adaptive ground estimation, making it suitable for driving scenes containing sloped, uneven, or locally varying road surfaces.

The two resulting point sets are retained for different stages of the subsequent processing. Ground points provide the geometric reference required for local ground-surface estimation, whereas non-ground points form the initial set of obstacle candidates. This separation prevents road-surface measurements from being directly interpreted as obstacles while retaining the ground information required for relative-height filtering.

Patchwork++ is applied in the original LiDAR coordinate frame. The segmented ground and non-ground point sets are subsequently transformed and spatially filtered as described in the following section.

Radar-Frame Alignment and Spatial Filtering

After ground segmentation, both the ground and non-ground point sets are transformed from the LiDAR coordinate frame into the radar-aligned coordinate frame. Performing the transformation after Patchwork++ preserves the separation between the ground reference points and the non-ground obstacle candidates while placing both sets in the common coordinate system used for subsequent occupancy-label generation.

The transformed point clouds are then restricted to the spatial region relevant to the radar measurement. Essentially, points outside the radar field of view are removed, and returns associated with the ego vehicle are excluded. These operations prevent LiDAR measurements outside the effective radar sensing region or originating from the vehicle itself from contributing to the generated supervision.

The same spatial filtering is applied to both the ground and non-ground point sets. The retained ground points are subsequently used to estimate the local road surface, while the retained non-ground points are treated as candidate obstacle measurements for relative-height filtering. This ensures that both point sets are defined in the same radar-aligned spatial region before the following processing stages.

Local Ground-Surface Estimation and Height Filtering

After radar-frame alignment and spatial filtering, the retained ground points are used to estimate a local ground surface for each LiDAR frame. A local reference is required because the road surface cannot be assumed to remain globally horizontal across real-world driving scenes. Road slope, local elevation changes, and changes in vehicle orientation can shift the vertical position of the road surface in the radar-aligned coordinate frame. Consequently, obstacle filtering based only on a fixed absolute height threshold could produce inconsistent labels across frames.

The local ground surface is approximated by a plane fitted to the segmented ground points within the radar-aligned region of interest, as:

$$z = ax + by + c, \tag{3.1}$$

where x, y, and z denote the coordinates of a LiDAR point in the radar-aligned frame, and a, b, and c are the estimated plane parameters. To reduce the influence of incorrectly segmented points and local surface irregularities, the plane estimate is refined using residual-based filtering. An initial plane is fitted to the available ground points, after which points with an absolute normal residual greater than 0.08 m are discarded. If at least 20 inlier points remain, the plane is re-estimated from these points. The resulting estimate is accepted only if the residual root-mean-square error does not exceed 0.08 m and the median absolute residual does not exceed 0.05 m.

If insufficient reliable ground points are available to obtain an accepted estimate in the current frame, or if the fitted plane fails these quality criteria, the most recent valid ground-plane estimate from an earlier frame is reused. If no earlier valid estimate is available, a constant ground height of $z = -1.8223$ m is used as a fallback. This ensures that a ground reference remains available for subsequent relative-height filtering, in which non-ground points between 0.1 m and 2.5 m above the estimated local ground surface are retained as obstacle candidates.

For each retained non-ground point, the height relative to the estimated local ground surface is computed as:

$$h = z - (ax + by + c). \tag{3.2}$$

Using relative rather than absolute height reduces the sensitivity of the obstacle-selection process to road slope and frame-dependent changes in the ground position. A non-ground point is retained as obstacle evidence only if

$$h_{\min} \leq h \leq h_{\max}, \tag{3.3}$$

with

$$0.1 \text{ m} \leq h \leq 2.5 \text{ m}. \tag{3.4}$$

The lower threshold suppresses residual ground points, road-surface noise, and other very low structures that are not treated as relevant obstacle evidence. The

upper threshold removes elevated measurements such as tree canopies, overhead structures, and high object parts that are not directly relevant to the near-ground occupancy representation considered in this work.

The resulting height-filtered point set provides the obstacle evidence used for polar-grid projection. Figure 3.3(f) illustrates the retained points after relative-height filtering.

Polar Grid Projection

The height-filtered LiDAR obstacle points are projected onto the polar range–azimuth grid defined by the radar representation. This converts the three-dimensional LiDAR obstacle evidence into a two-dimensional spatial support that is directly aligned with the radar range–azimuth coordinates.

For each retained point (x, y, z) in the radar-aligned coordinate frame, where x denotes the forward direction, y the lateral direction, and z the vertical direction, the range and azimuth coordinates are computed as:

$$r = \sqrt{x^2 + y^2 + z^2}, \tag{3.5}$$

and

$$\theta = \arctan 2(y, x). \tag{3.6}$$

Let $\{r_i\}_{i=1}^{R}$ and $\{\theta_j\}_{j=1}^{A}$ denote the centres of the radar range and azimuth bins, respectively. Each LiDAR point is assigned independently to the nearest radar range and azimuth bin as:

$$i_r = \arg\min_i |r - r_i|, \qquad i_\theta = \arg\min_j |\theta - \theta_j|. \tag{3.7}$$

Nearest-bin assignment is used because the physical radar azimuth angles are not uniformly spaced. The azimuth support is obtained from the angle-FFT spatial-frequency grid and mapped to physical angles through the corresponding arcsine relation. Consequently, a constant angular spacing $\Delta\theta$ is not assumed.

Each retained LiDAR point contributes obstacle evidence to the corresponding polar cell (i_r, i_θ). Multiple points assigned to the same cell do not change its binary obstacle-evidence state. The resulting sparse polar obstacle map is subsequently used to determine the nearest LiDAR-supported obstacle boundary along each azimuth direction.

First-Return Occupancy Grid Construction

After projection onto the radar-aligned polar grid, the retained LiDAR obstacle evidence is converted into a structured first-return occupancy target. Rather than treating all projected obstacle cells as occupied, only the nearest obstacle evidence along each azimuth direction is retained as the observable boundary. Figure 3.4

illustrates the resulting labelling rule for each azimuth column with and without observed obstacle evidence.

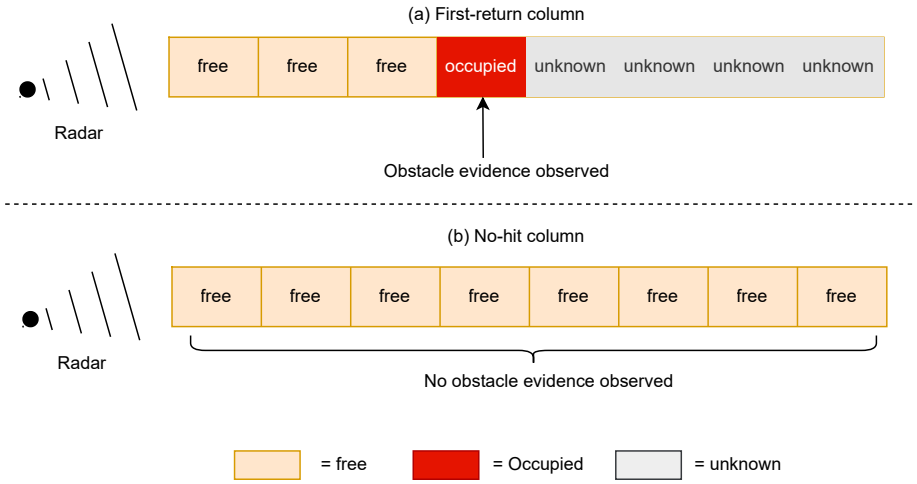

Figure 3.4: First-return occupancy labelling rule for azimuth columns with (top) and without (bottom) observed obstacle evidence.

Let $O(r,\theta) \in \{0,1\}$ denote the projected LiDAR obstacle evidence at range bin r and azimuth bin θ. If obstacle evidence is present in an azimuth column, the nearest occupied range bin is selected as the first-return index:

$$t_\theta = \min\{r \mid O(r,\theta) = 1\}. \tag{3.8}$$

For a valid first return t_θ, the corresponding occupancy label is defined as:

$$y(r,\theta) = \begin{cases} \text{free}, & r < t_\theta, \\ \text{occupied}, & r = t_\theta, \\ \text{unknown}, & r > t_\theta. \end{cases} \tag{3.9}$$

Cells before the first return are therefore interpreted as visible free space, the first-return cell represents the nearest observed obstacle boundary, and cells behind that boundary are labelled as unknown because they are not directly observable from the current LiDAR measurement. This follows the visibility structure underlying ray-based occupancy mapping [8].

If no obstacle evidence is present in an azimuth column, the entire column is assigned to the no-hit state. In this case, no occupied boundary is defined within the valid sensing range and all valid range cells in the column are labelled as free. The no-hit state is retained explicitly in the structured target representation and is later treated as an additional outcome alongside the physical first-return range bins.

The resulting target is therefore defined independently for each azimuth direction. If a valid first return is present, the corresponding azimuth column consists of free space before the first-return bin, an occupied state at the first-return bin, and unknown space behind it. If no valid first return is present, the valid portion of the column is represented entirely as free space. This representation preserves the visible free-space structure of the LiDAR observation while reducing multiple projected obstacle responses along the same azimuth to the nearest observable boundary.

Valid Supervision Region

Notably, not all cells in the implemented occupancy tensor correspond to physically meaningful or sufficiently reliable LiDAR supervision. The valid supervision region is therefore restricted to the physical radar-aligned support for which the LiDAR-derived first-return target is considered reliable.

The first exclusion concerns padded bins that do not correspond to the physical radar sensing region. Following the polar-grid projection described in Section 3.1.2, the radar-aligned physical region used for supervision contains 500 range bins and 240 azimuth bins. For network processing, the corresponding tensors have a spatial size of 512×256, providing dimensions compatible with the successive downsampling operations of the U-Net 3+ backbone. The additional bins therefore contain no physical supervision: the last 12 range bins are excluded, together with 8 padded azimuth bins on each side. After these padded regions are removed, the valid radar-aligned support used for training and evaluation is 500×240.

A second exclusion is required because the front radar installation partially occludes low-height LiDAR measurements within a limited angular sector. This can result in incomplete obstacle evidence and consequently unreliable first-return targets, particularly when missing LiDAR returns would otherwise be interpreted as free space or as a no-hit observation.

Figure 3.5 illustrates the affected angular region using height-filtered LiDAR obstacle evidence projected onto the radar range–azimuth grid.

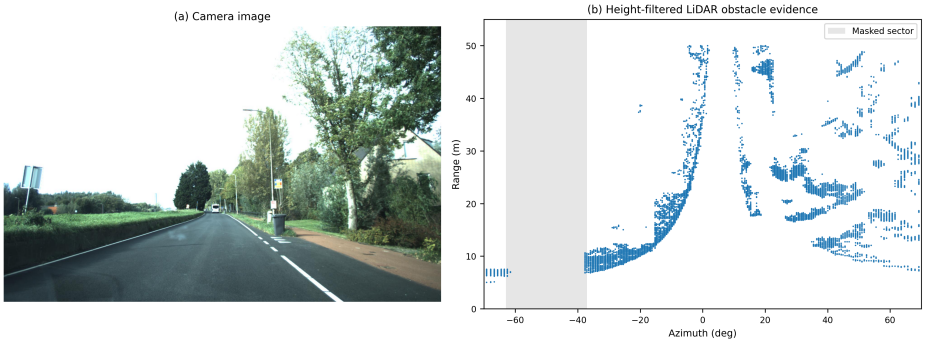

Figure 3.5: Example of the LiDAR supervision sector affected by partial occlusion from the front radar installation. (a) Synchronized front-camera image. (b) Height-filtered LiDAR obstacle evidence in polar range–azimuth coordinates, with the affected sector shown by the gray shaded region.

Because of this, the azimuth sector from approximately $-63°$ to $-37°$ is therefore excluded from the valid supervision region. Although LiDAR measurements may still be present within this sector, their completeness is not considered sufficiently reliable for consistent first-return target generation.

To summarize, the resulting valid supervision region excludes both non-physical padded cells and the angular sector affected by LiDAR visibility degradation. The corresponding validity mask is applied consistently during model training and quantitative evaluation, as described later in Sections **??** and 3.2.4.

3.1.3 Radar Input Representations

This subsection describes the radar input representations considered for first-return occupancy prediction. The primary representation is the complete RDA radar cube, which retains the Doppler-resolved measurement at every range–azimuth location. Two fixed Doppler-compressed range–azimuth representations are additionally considered to assess the effect of removing the Doppler dimension before network inference.

The full-RDA representation preserves the complete Doppler spectrum and is subsequently processed by a learned Doppler encoder, whereas the compressed variants reduce the Doppler dimension to a single scalar at each range–azimuth location before spatial prediction. The following subsections first define the full-RDA representation and its alignment with the LiDAR-derived supervision, and then introduce the Doppler-mean and Doppler-maximum range–azimuth variants.

Range–Doppler–Azimuth Radar Representation

For each radar frame, the preprocessed radar measurement is represented as a range–Doppler–azimuth tensor

$$X_{\mathrm{RAD}} \in \mathbb{R}^{R \times D \times A}, \tag{3.10}$$

where (R), (D), and (A) denote the numbers of range, Doppler, and azimuth bins, respectively. Although the radar processing chain can also provide elevation estimates, elevation information is not used as an input to the models considered in this thesis. The prediction task is defined on a two-dimensional range–azimuth occupancy grid, and the radar input therefore retains the power measurements only in range, Doppler, and azimuth dimensions.[2]

For network processing, the radar cube is provided in batched form:

$$X_{\mathrm{RAD}} \in \mathbb{R}^{B \times R \times D \times A}, \tag{3.11}$$

where B denotes the batch size. The radar-aligned physical support contains 500 range bins, 128 Doppler bins, and 240 azimuth bins. Before being passed to the network, the spatial dimensions are zero-padded from 500×240 to 512×256 to obtain dimensions compatible with the successive downsampling operations of the U-Net 3+ backbone. Specifically, 12 range bins are appended at the far-range end and 8 azimuth bins are added on each side. These padded bins do not correspond to physical radar measurements and are therefore excluded from supervision and evaluation using the validity mask defined in Section 3.1.2.

For Doppler encoding, the tensor is rearranged such that the Doppler dimension precedes the spatial dimensions:

$$X_{\mathrm{RAD}}^{\mathrm{BDRA}} \in \mathbb{R}^{B \times D \times R \times A}. \tag{3.12}$$

This rearrangement changes only the tensor layout and does not modify the radar measurements.

Let $X_{\mathrm{RAD}}(d, r, \theta)$ denote the radar response at Doppler bin d, range bin r, and azimuth bin θ. The corresponding target $Y(r, \theta)$ is defined at the same physical range–azimuth location after LiDAR projection onto the radar grid.

The azimuth dimension follows the physical radar angular support rather than a uniformly spaced angular grid. As described in Section 3.1.2, LiDAR measurements are assigned to the nearest radar range and azimuth bins so that the target grid follows the same spatial sampling as the radar representation.

Unlike a fixed Doppler-compressed representation, the full radar cube retains the complete Doppler response at each range–azimuth location. This preserves information about spectral peak location, spectral width, and multi-component structure in the Doppler information that would be lost when the Doppler dimension is reduced to a single scalar, e.g. by a summing operation.

Figure 3.6 illustrates this distinction. Panel (a) shows a Doppler-sum range–azimuth map for spatial reference only, whereas the model input remains the full

[2] Elevation estimates are available separately in the RaDelft radar processing output, but they are not included as an input channel in the proposed framework. Extending the formulation to elevation-aware or three-dimensional occupancy prediction is left for future work.

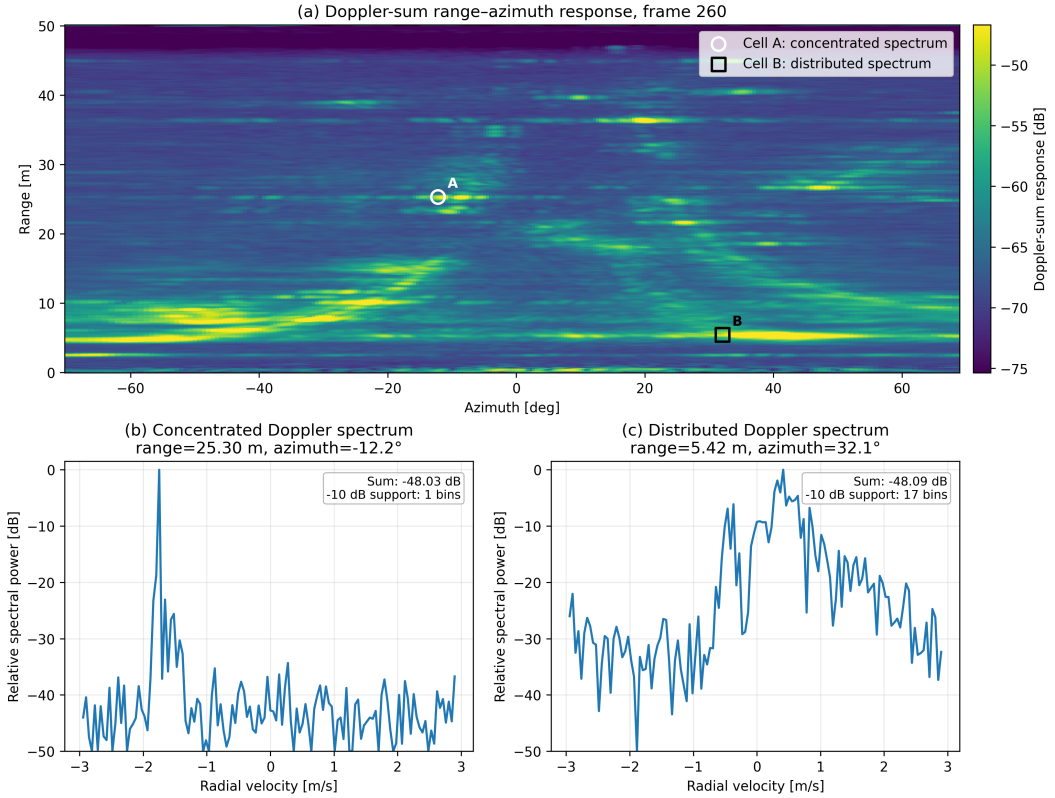

Figure 3.6: Representative Doppler spectra from two range–azimuth cells with nearly identical Doppler-sum responses but different spectral distributions shown in (b) and (c). Panel (a) is shown for spatial reference only; the full Doppler-resolved cube is used as the primary network input.

Doppler-resolved radar cube. The two selected cells have nearly identical Doppler-sum responses of -48.03 dB and -48.09 dB, but their underlying spectra differ substantially. Cell A exhibits a concentrated response with a -10 dB spectral support of one Doppler bin, whereas Cell B exhibits a distributed response spanning 17 bins. Fixed scalar aggregation therefore removes spectral characteristics that remain instead available in the full-RDA representation.

Doppler-Compressed Range–Azimuth Variants

Two fixed Doppler-compressed range–azimuth representations are also considered to evaluate the effect of removing the Doppler dimension before network inference. Given the preprocessed radar cube $X_{\mathrm{RAD}}(d, r, \theta)$, Doppler-mean aggregation is defined as

$$X_{\text{mean}}(r,\theta) = \frac{1}{D}\sum_{d=1}^{D} X_{\text{RAD}}(d,r,\theta), \tag{3.13}$$

whereas Doppler-maximum aggregation is defined as

$$X_{\text{max}}(r,\theta) = \max_{d} X_{\text{RAD}}(d,r,\theta). \tag{3.14}$$

Both operations collapse the Doppler dimension and produce a single-channel range–azimuth representation. For an individual frame, the resulting network input has dimensions

$$1 \times R \times A, \tag{3.15}$$

while a training batch is represented as

$$B \times 1 \times R \times A. \tag{3.16}$$

Doppler-mean aggregation summarizes the complete spectrum through uniform averaging, whereas Doppler-maximum aggregation retains only the strongest response at each range–azimuth location. Consequently, both representations discard the internal spectral structure available in the full range–Doppler–azimuth measurement.

A Doppler-sum representation was also considered during preliminary experiments. Since the number of Doppler bins D is fixed, Doppler summation and Doppler averaging differ only by the constant scale factor D,

$$X_{\text{sum}}(r,\theta) = D\, X_{\text{mean}}(r,\theta). \tag{3.17}$$

The final comparison therefore retains Doppler mean and Doppler maximum as the two fixed compression strategies, avoiding a largely redundant representation while preserving the distinction between averaging the complete spectrum and retaining its strongest response.

The resulting range–azimuth maps are then passed directly to the same U-Net 3+ spatial prediction backbone used by the full-RDA framework and therefore do not employ a learned Doppler encoder. Compared with the full-RDA input, these representations reduce the model-side input dimensionality and remove Doppler-resolved information before learning. They are therefore used to assess the effect of fixed Doppler compression relative to retaining and learning from the complete Doppler-resolved measurement.

Since the compressed variants modify both the radar representation and the associated front-end processing, they are treated as complete input-processing configurations rather than encoder-only ablations. The learnable Doppler encoding strategies used with the full-RDA representation are described separately in Section 3.1.4.

3.1.4 Doppler Encoding

For configurations using the full RDA radar cube, the Doppler dimension must be transformed into a compact range–azimuth feature representation before spatial prediction. Three encoding strategies are considered for this purpose. A single-layer projection provides a simple reference that treats the Doppler bins directly as input channels. A "Roldan-style encoder" inspired by [29] provides a structured three-dimensional convolutional reference adapted from prior radar processing work. Finally, the proposed Range–Azimuth-Preserving Doppler Spectrum Encoder applies multi-scale Doppler-only processing while explicitly preserving the range–azimuth geometry.

All three encoders receive the full Doppler-resolved radar measurement and produce range–azimuth feature maps for the subsequent U-Net 3+ backbone. Their main difference therefore lies in how the Doppler spectrum is processed before spatial reasoning. This is explained in the following.

Single-Layer Doppler Projection Encoder

As a simple full-RDA reference, the Doppler bins are treated as input channels and projected directly into 16 range–azimuth feature channels using a single 1×1 convolution:

$$F_{\text{proj}} = \text{Conv}_{1\times1}\left(X_{\text{RAD}}^{\text{BDRA}}\right), \qquad F_{\text{proj}} \in \mathbb{R}^{B\times16\times R\times A}. \tag{3.18}$$

This projection preserves the range–azimuth resolution but does not explicitly model local relationships between neighbouring Doppler bins. The resulting features are passed to the same U-Net 3+ prediction backbone as the other full-RDA configurations. Because this model was trained using an earlier objective setting, it is included as a reference configuration rather than as a strictly controlled encoder-only ablation.

"Roldan-Style" Doppler Encoder

A second full-RDA reference is adapted from the DopplerEncoder of Roldan et al. [29]. The power-only radar tensor is processed by two three-dimensional convolutional layers with kernel sizes $(5,3,3)$ and $(4,3,3)$, both using a Doppler stride of 4, followed by three-dimensional max pooling over the remaining Doppler dimension. The feature width is adapted to 16 channels so that the encoder interfaces with the same U-Net 3+ prediction backbone used by the proposed configuration. For $D = 128$, the Doppler dimension is reduced as $128 \to 32 \to 8 \to 1$, producing

$$F_{\text{Roldan}} \in \mathbb{R}^{B\times16\times R\times A}. \tag{3.19}$$

Unlike the proposed Doppler Spectrum Encoder, the 3×3 spatial support of the three-dimensional convolutions also mixes neighbouring range–azimuth locations during Doppler encoding.

Range–Azimuth-Preserving Doppler Spectrum Encoder

The proposed Range–Azimuth-Preserving Doppler Spectrum Encoder is designed to compress the Doppler dimension while explicitly preserving the range–azimuth geometry of the radar measurement. In contrast to the Roldan-style encoder, the three-dimensional convolutional kernels have unit support in the range and azimuth dimensions. Doppler-spectrum encoding is therefore performed independently at each range–azimuth location, while spatial reasoning is deferred to the subsequent two-dimensional U-Net 3+ backbone.

For an input radar cube

$$X_{\mathrm{RAD}}^{\mathrm{BDRA}} \in \mathbb{R}^{B\times D\times R\times A}, \tag{3.20}$$

with $D = 128$, the tensor is first expanded to

$$X_{\mathrm{enc}} \in \mathbb{R}^{B\times 1\times D\times R\times A}. \tag{3.21}$$

The encoder then processes the Doppler spectrum using three parallel branches with Doppler kernel lengths

$$k_{\mathrm{D}} \in 3, 7, 15. \tag{3.22}$$

These kernel lengths were selected as a practical multi-scale design to provide progressively larger receptive fields along the Doppler dimension. The three branches therefore capture relatively narrow, intermediate, and broader spectral patterns, respectively, while the use of odd kernel sizes allows symmetric padding and preserves alignment along the Doppler axis. The selected values are not intended to correspond to predefined physical velocity intervals, but rather to provide complementary local contexts over the Doppler spectrum.

Each branch contains two three-dimensional convolutions with kernel size $(k_{\mathrm{D}}, 1, 1)$, stride $(4, 1, 1)$, and symmetric padding along the Doppler dimension. The first convolution maps the input from 1 to 16 channels, and the second retains 16 channels. Each convolution is followed by batch normalization and a rectified linear unit (ReLU) activation. For every branch, the Doppler dimension is consequently reduced as

$$128 \to 32 \to 8, \tag{3.23}$$

while the range and azimuth dimensions remain unchanged.

The three kernel lengths provide different receptive fields along the Doppler axis and allow the encoder to learn spectral patterns over different scales without mixing neighbouring range–azimuth cells. Each branch produces

$$F^{(k_{\mathrm{D}})} \in \mathbb{R}^{B\times 16\times 8\times R\times A}. \tag{3.24}$$

The three branch outputs are concatenated along the feature-channel dimension, producing 48 channels, and are subsequently fused using a $1 \times 1 \times 1$ convolution:

$$F_{\text{fused}} \in \mathbb{R}^{B\times 16\times 8\times R\times A}. \tag{3.25}$$

The fusion layer is followed by batch normalization and ReLU. A final Doppler-only max-pooling operation with kernel size and stride $(8, 1, 1)$ reduces the remaining Doppler dimension from 8 to 1. After removing the singleton dimension, the final encoder output is

$$F_{\text{D}} \in \mathbb{R}^{B\times 16\times R\times A}. \tag{3.26}$$

Figure 3.7 summarizes the encoder architecture.

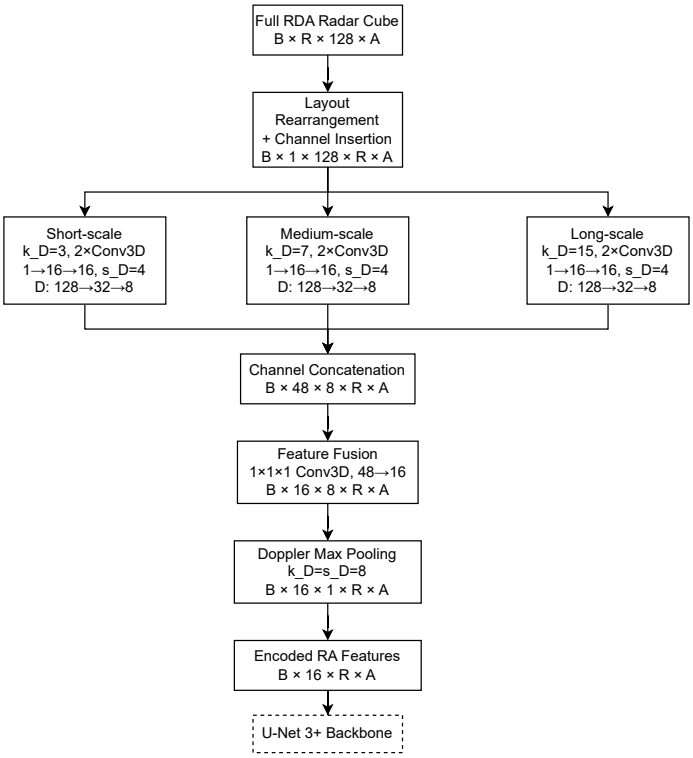

Figure 3.7: Architecture of the proposed Range–Azimuth-Preserving Doppler Spectrum Encoder. Here, k_D and s_D denote the Doppler-axis kernel size and stride, respectively, while D indicates the Doppler-bin resolution through the encoder. Three parallel branches with $k_D = \{3, 7, 15\}$ reduce D from 128 to 8 before feature fusion and final Doppler pooling, while preserving the range–azimuth dimensions.

The final max-pooling operation is applied to learned multi-channel spectral features rather than directly to the raw Doppler spectrum. The encoder therefore retains learnable Doppler processing before dimensional reduction while avoid-

ing three-dimensional spatial convolution across neighbouring range–azimuth cells. No additional fixed Doppler-sum branch is used. The resulting 16-channel range–azimuth representation is passed directly to the structured first-return prediction backbone described in Section 3.1.5.

3.1.5 Structured First-Return Prediction

The encoded range–azimuth representation is processed by a U-Net 3+ spatial backbone to estimate the nearest observable boundary along each azimuth direction. Rather than predicting the free, occupied, and unknown states independently for every grid cell, the network produces a structured first-return representation from which the final three-state occupancy grid is rendered.

Figure 3.8 shows the complete proposed single-frame architecture, consisting of the Range–Azimuth-Preserving Doppler Spectrum Encoder described in Section 3.1.4, the U-Net 3+ spatial prediction backbone, and the structured first-return output head.

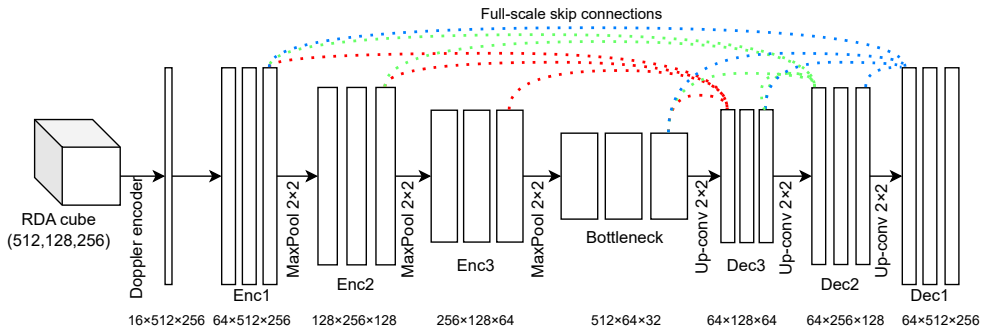

Figure 3.8: Architecture of the proposed single-frame first-return occupancy model. The Doppler-resolved radar cube is encoded into range–azimuth features, processed by the U-Net 3+ spatial backbone, and mapped to an azimuth-wise first-return or no-hit representation.

The spatial dimensions shown in Figure 3.8 correspond to the padded network tensor with a spatial size of 512×256, obtained by zero-padding the physical radar-aligned support from 500×240. Specifically, 12 range bins are appended at the far-range end and 8 azimuth bins are added on each side. These padded bins are used only for network compatibility and are excluded from supervision and quantitative evaluation, which are restricted to the valid physical region defined in Section 3.1.2.

Architecture Design Rationale

The prediction architecture is designed around the visibility structure of the LiDAR-derived supervision. For each azimuth direction, the target is determined by the

nearest observed obstacle boundary or by a no-hit outcome. The network therefore estimates this structured variable directly, rather than treating the occupancy state of every range–azimuth cell as an independent prediction problem.

After Doppler encoding, spatial reasoning is performed on the range–azimuth plane using an encoder–decoder backbone. Multi-scale spatial context is useful for combining local radar evidence with broader scene structure, suppressing isolated responses, and producing a spatially coherent estimate of the first-return boundary. The resulting prediction is subsequently converted into free, occupied, and unknown regions according to the same first-return visibility rule used for LiDAR target generation.

U-Net 3+ Based Backbone

A U-Net 3+ encoder–decoder is used as the spatial prediction backbone on the polar range–azimuth grid [12]. Its full-scale skip connections aggregate features from multiple spatial resolutions, allowing local radar evidence to be combined with broader range–azimuth context during dense prediction.

For the proposed full-RDA configuration, the backbone receives the 16-channel range–azimuth feature representation produced by the Range–Azimuth-Preserving Doppler Spectrum Encoder. The Doppler-mean and Doppler-maximum variants bypass the Doppler encoder and provide their single-channel range–azimuth representations directly to the same spatial backbone. The alternative full-RDA encoder configurations likewise interface with the backbone through a 16-channel feature representation.

The U-Net 3+ backbone therefore provides a common spatial prediction stage across the evaluated radar representations and Doppler encoders. U-Net 3+ was selected because its full-scale skip connections combine features from multiple encoder and decoder resolutions, allowing the network to preserve fine range–azimuth localization while simultaneously exploiting broader spatial context. This combination is well suited to dense first-return occupancy prediction, where accurate boundary localization must be inferred together with larger-scale free-space structure. Its role is to perform multi-scale range–azimuth reasoning and produce the features required for the azimuth-wise first-return and no-hit prediction described in Section 3.1.5. The backbone itself is used as an established dense prediction architecture rather than as a methodological contribution of this thesis.

First-Return and No-Hit Prediction Formulation

The prediction target follows the same azimuth-wise structure as the LiDAR-derived supervision defined in Section 3.1.2. For each azimuth direction, the network estimates either the range bin corresponding to the first observable obstacle boundary or a no-hit outcome.

Let F denote the range–azimuth feature tensor provided to the prediction network. For the full-RDA configurations, F corresponds to the output of one of

the three Doppler encoding strategies introduced above: the single-layer projection encoder, the Roldan-style encoder, or the proposed Range–Azimuth-Preserving Doppler Spectrum Encoder. The prediction network then produces a first-return score tensor

$$S = f_\Phi(F), \qquad S \in \mathbb{R}^{B\times A\times(R+1)}, \tag{3.27}$$

where B, A, and R denote the batch size, number of azimuth bins, and number of valid physical range bins, respectively. For each azimuth bin, the first R entries correspond to candidate first-return positions, while the additional entry represents the no-hit class.

For hard prediction, the estimated state for each azimuth bin θ is obtained as

$$\hat{t}_\theta = \arg\max_{k\in\{0,\dots,R\}} S(\theta,k), \tag{3.28}$$

where $k = 0,\dots,R-1$ correspond to physical range bins and $k = R$ denotes the no-hit state.

The predicted first-return or no-hit state determines the complete occupancy structure of the corresponding azimuth column. A physical first-return prediction defines the free region before the boundary, the occupied boundary cell, and the unknown region behind it, whereas a no-hit prediction corresponds to an entirely free valid column. The hard and the differentiable soft rendering procedures are described in the rest of this Section 3.1.5.

Occupancy Grid Rendering

The azimuth-wise first-return representation is converted into a three-state occupancy grid using the same visibility structure adopted for LiDAR target generation. Two rendering modes are used. A differentiable soft rendering is employed during training to provide occupancy probabilities for the Lovasz component of the loss function, whereas a hard rendering is used at inference to obtain discrete occupancy predictions for evaluation and visualization.

Let

$$P_{\text{pred}}(\theta,k) = \text{softmax}\left(S(\theta,:)\right)_k, \qquad k\in\{0,\dots,R\}, \tag{3.29}$$

denote the predicted first-return distribution for azimuth bin θ, where $k = 0,\dots,R-1$ correspond to physical range bins and $k = R$ denotes the no-hit state. The physical-range probabilities are denoted by

$$P_{\text{range}}(\theta,k) = P_{\text{pred}}(\theta,k), \qquad k = 0,\dots,R-1. \tag{3.30}$$

For range bin r, the *soft* occupied probability is given directly by the probability that the first return occurs at that location:

$$P_{\text{occ}}(r,\theta) = P_{\text{range}}(\theta,r). \tag{3.31}$$

A cell is unknown when the first return occurs at a smaller range and therefore occludes the cell from direct observation. Its probability is consequently given by the exclusive cumulative probability, where "exclusive" means that the cumulative sum includes only first-return probabilities at ranges strictly smaller than the current bin (r), excluding the probability at (r) itself:

$$P_{\text{unknown}}(r,\theta) = \sum_{k=0}^{r-1} P_{\text{range}}(\theta,k). \tag{3.32}$$

Conversely, a cell is free when no first return has occurred up to and including its range position. The corresponding probability is:

$$P_{\text{free}}(r,\theta) = 1 - \sum_{k=0}^{r} P_{\text{range}}(\theta,k). \tag{3.33}$$

These three probabilities satisfy:

$$P_{\text{free}}(r,\theta) + P_{\text{occ}}(r,\theta) + P_{\text{unknown}}(r,\theta) = 1. \tag{3.34}$$

Because the no-hit probability is not included in the cumulative sum over the physical range bins, it contributes implicitly to the free-space probability. Consequently, a high no-hit probability increases the probability that cells throughout the valid range remain free. The resulting differentiable occupancy probabilities are used by the occupancy-level component of the loss function described in Section 3.1.6.

For *hard* rendering, the predicted first-return or no-hit state is obtained using the maximum-score index $\hat{t}_\theta$ defined in Section 3.1.5. If $\hat{t}_\theta < R$, the discrete occupancy prediction is:

$$\hat{y}(r,\theta) = \begin{cases} \text{free}, & r < \hat{t}_\theta, \\ \text{occupied}, & r = \hat{t}_\theta, \\ \text{unknown}, & r > \hat{t}_\theta. \end{cases} \tag{3.35}$$

If $\hat{t}_\theta = R$, the column is classified as no hit and all valid range cells are rendered as free:

$$\hat{y}(r,\theta) = \text{free}, \qquad \forall r \in \{0,\dots,R-1\}. \tag{3.36}$$

The rendering procedure therefore maps each azimuth-wise first-return or no-hit prediction to a visibility-consistent occupancy column. It also ensures that the differentiable representation used during training and the discrete occupancy grid used during evaluation follow the same underlying first-return structure.

3.1.6 Loss Function

The proposed framework is optimized using a task-aligned loss function that combines supervision of the rendered occupancy structure with direct supervision of

the azimuth-wise first-return boundary. The baseline combined objective is first defined below, followed by Radar-Aware Weighting as a controlled modification of the first-return boundary term.

Combined Cost Function

The loss function is designed to supervise two complementary aspects of the structured first-return task. The first is occupancy-level agreement between the rendered prediction and the LiDAR-derived target, while the second is the range-wise accuracy of the predicted first-return boundary. The proposed single-frame baseline therefore combines a Lovasz-based occupancy term with a smoothed CDF-L1 first-return term.

The Lovasz loss is applied to the soft occupancy representation rendered from the predicted first-return distribution, following the rendering formulation described in Section 3.1.5. It provides a differentiable surrogate for intersection over union and therefore directly supports occupancy-level agreement [2]. Because the occupied state corresponds to a single first-return boundary cell in each azimuth column, the Lovasz term is evaluated only over the free and unknown states:

$$\mathcal{C}_{\text{Lovasz}} = \{\text{free}, \text{unknown}\}. \tag{3.37}$$

The final Lovasz term is averaged over the included states that are present within the valid supervision region as:

$$\mathcal{L}_{\text{Lovasz}} = \frac{1}{|\mathcal{C}_{\text{present}}|} \sum_{c \in \mathcal{C}_{\text{present}}} \mathcal{L}^{c}_{\text{Lovasz}}, \qquad \mathcal{C}_{\text{present}} \subseteq \mathcal{C}_{\text{Lovasz}}. \tag{3.38}$$

Valid no-hit columns are rendered as free and therefore remain supervised by the Lovasz term.

Occupancy overlap alone does not directly penalize displacement of the predicted first-return boundary along the range dimension. A smoothed CDF-L1 term is therefore applied directly to the azimuth-wise first-return distribution. Let $P_{\text{pred}}(\theta, k)$ denote the predicted probability that the first return for azimuth bin θ occurs at physical range bin k. The predicted cumulative distribution over the physical range bins is

$$C_{\text{pred}}(r, \theta) = \sum_{k=0}^{r} P_{\text{pred}}(\theta, k). \tag{3.39}$$

The additional no-hit class is excluded from this cumulative sum. For a no-hit target column, the ground-truth cumulative distribution is set to zero throughout the valid range.

For a column with a valid first return at range index t_θ, a one-sided linear smoothing window is applied before the first-return position:

$$C_{\mathrm{gt}}(r,\theta) = \begin{cases} 0, & r \le t_\theta - w, \\ \dfrac{r-(t_\theta - w)}{w}, & t_\theta - w < r < t_\theta, \\ 1, & r \ge t_\theta. \end{cases} \tag{3.40}$$

The smoothing width is set to

$$w = 2 \tag{3.41}$$

range bins. This reduces sensitivity to small discretization differences around the LiDAR-derived first-return position while retaining the ordered range structure of the target. The use of cumulative distributions also provides a range-aware comparison of the predicted and target first-return distributions and is related to the one-dimensional Wasserstein distance [35].

The smoothed CDF-L1 loss is computed over the valid range–azimuth support:

$$\mathcal{L}_{\mathrm{CDF}} = \frac{1}{|\Omega|} \sum_{(r,\theta)\in\Omega} \left| C_{\mathrm{pred}}(r,\theta) - C_{\mathrm{gt}}(r,\theta) \right|, \tag{3.42}$$

where Ω denotes the range–azimuth cells retained by the validity mask defined in Section 3.1.2, which excludes the padded bins and the unreliable azimuth sector from supervision.

Because the Lovasz and CDF terms have different numerical scales, each term is normalized before combination. The general loss is written as:

$$\mathcal{L} = \lambda_{\mathrm{Lovasz}} \frac{\mathcal{L}_{\mathrm{Lovasz}}}{\gamma_{\mathrm{Lovasz}}} + \lambda_{\mathrm{CDF}} \frac{\mathcal{L}_{\mathrm{CDF}}}{\gamma_{\mathrm{CDF}}}, \tag{3.43}$$

where $\lambda_{\mathrm{Lovasz}}$ and λ_{CDF} determine the global balance between the two components, and γ_{Lovasz} and γ_{CDF} denote their normalization values. For the proposed baseline,

$$\lambda_{\mathrm{Lovasz}} = \lambda_{\mathrm{CDF}} = 0.5, \qquad \gamma_{\mathrm{Lovasz}} = 2.0, \qquad \gamma_{\mathrm{CDF}} = 1.0. \tag{3.44}$$

The normalization factors were chosen based on the numerical scales observed when the Lovasz and smoothed CDF terms were evaluated separately. In these experiments, the Lovasz loss was approximately twice the magnitude of the smoothed CDF loss. Accordingly, $\gamma_{\mathrm{Lovasz}} = 2.0$ and $\gamma_{\mathrm{CDF}} = 1.0$ are used to bring the two terms to a comparable scale before combination. After this normalization, equal global weights, $\lambda_{\mathrm{Lovasz}} = \lambda_{\mathrm{CDF}} = 0.5$, are assigned so that occupancy-overlap supervision and first-return localization contribute equally to the combined objective, without imposing an a priori preference for either component. The implemented baseline loss is therefore

$$\mathcal{L} = 0.5 \frac{\mathcal{L}_{\mathrm{Lovasz}}}{2.0} + 0.5 \mathcal{L}_{\mathrm{CDF}}. \tag{3.45}$$

Both loss components are restricted to the valid supervision region defined in Section 3.1.2. Let $M(r,\theta)$ denote the corresponding binary validity mask:

$$M(r,\theta) = \begin{cases} 1, & \text{if the cell belongs to the valid supervision region,} \\ 0, & \text{otherwise.} \end{cases} \tag{3.46}$$

For a cell-wise loss contribution $\ell(r,\theta)$, masking can be expressed as

$$\mathcal{L}_{\text{masked}} = \frac{\sum_{r,\theta} M(r,\theta)\ell(r,\theta)}{\sum_{r,\theta} M(r,\theta)}. \tag{3.47}$$

Essentially, the validity mask prevents padded tensor regions and the unreliable LiDAR-supervision sector from contributing to model optimization.

Radar-Aware Weighting

Radar-Aware Weighting is introduced as a controlled modification of the baseline loss function. The radar input, network architecture, Lovasz term, and global balance between the two loss components remain unchanged:

$$\lambda_{\text{Lovasz}} = \lambda_{\text{CDF}} = 0.5. \tag{3.48}$$

Only the azimuth-wise contribution of the smoothed CDF-L1 term is modified. Let w_θ denote the reliability weight assigned to valid azimuth column θ, and let $\mathcal{L}_{\text{CDF},\theta}$ denote the mean CDF-L1 error over the valid range cells in that column. The weighted boundary term is defined as:

$$\mathcal{L}_{\text{CDF}}^{\text{RA}} = \frac{\sum_{\theta\in\Theta_{\text{valid}}} w_\theta \mathcal{L}_{\text{CDF},\theta}}{\sum_{\theta\in\Theta_{\text{valid}}} w_\theta}. \tag{3.49}$$

The resulting loss function is:

$$\mathcal{L}_{\text{RA}} = 0.5\frac{\mathcal{L}_{\text{Lovasz}}}{2.0} + 0.5\mathcal{L}_{\text{CDF}}^{\text{RA}}. \tag{3.50}$$

The reliability weighting accounts for abrupt ground-truth frontier changes, far-range and no-hit columns, proximity to the unreliable masked sector, and the availability of local radar evidence near the LiDAR-derived first return. The weighting is applied only to the smoothed CDF-L1 term during training and does not modify the inference architecture.

3.1.7 Causal Temporal Feature Fusion

The single-frame framework is extended with a causal temporal feature-fusion module to investigate whether short-term radar history provides complementary evidence for current-frame first-return prediction. For each target frame t, the model receives the current radar cube together with the two preceding radar cubes:

$$\mathcal{X}_t = \{X_{t-2}, X_{t-1}, X_t\}, \tag{3.51}$$

with

$$\mathcal{X}_t \in \mathbb{R}^{B\times T\times R\times D\times A}, \qquad T = 3. \tag{3.52}$$

The supervision target remains the LiDAR-derived first-return occupancy target corresponding to the current frame t.

Figure 3.9 illustrates the temporal feature-fusion architecture.

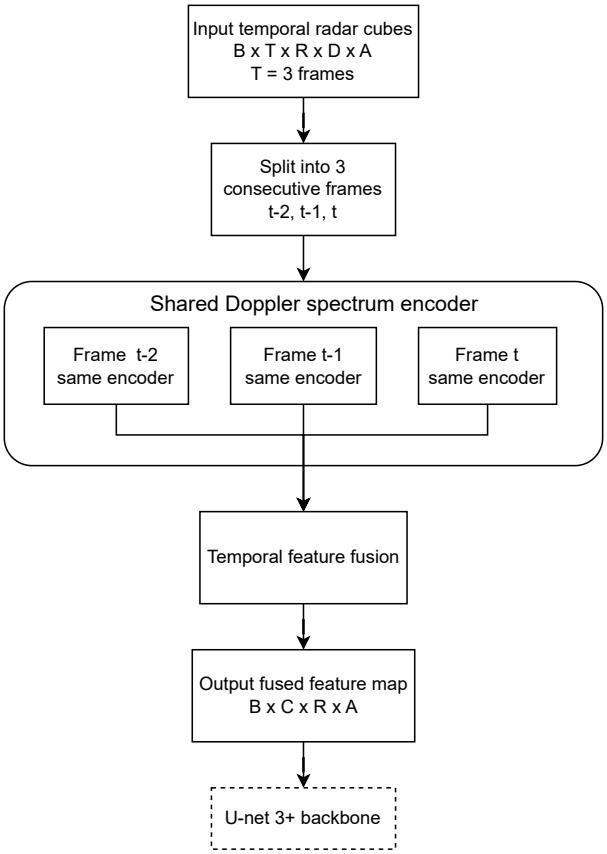

Figure 3.9: Causal temporal feature fusion architecture, which uses the radar cubes at $t-2$, $t-1$, and t. The three frames are independently processed by a shared Range–Azimuth-Preserving Doppler Spectrum Encoder, and the resulting range–azimuth feature maps are fused before spatial first-return prediction using the U-Net 3+ backbone.

Each of the three radar cubes is independently processed by the shared Range–Azimuth-Preserving Doppler Spectrum Encoder:

$$F_i = g_\psi(X_i), \qquad i \in \{t-2, t-1, t\}, \tag{3.53}$$

where the same encoder parameters ψ are used for all three cubes. The resulting feature maps satisfy

$$F_i \in \mathbb{R}^{B\times 16\times R\times A}. \tag{3.54}$$

The temporal module therefore operates on range–azimuth feature representations rather than directly on the raw range–Doppler–azimuth tensors.

Temporal fusion is performed using spatially varying attention weights. A shared scoring network q_η is applied independently to each encoded feature map:

$$s_i = q_\eta(F_i), \qquad s_i \in \mathbb{R}^{B\times 1\times R\times A}. \tag{3.55}$$

The scoring network consists of a 3×3 convolution that maps the 16 input channels to 8 hidden channels, followed by a ReLU activation and a 1×1 convolution that produces one temporal score at each range–azimuth location.

A learnable temporal prior b_i is added to the score associated with each frame. The attention weights are then obtained using a softmax over the three temporal inputs:

$$\alpha_i(r,a) = \frac{\exp(s_i(r,a)+b_i)}{\sum\limits_{j\in\{t-2,t-1,t\}} \exp(s_j(r,a)+b_j)}. \tag{3.56}$$

At every range–azimuth location,

$$\sum_{i\in\{t-2,t-1,t\}} \alpha_i(r,a) = 1. \tag{3.57}$$

The temporal weights are therefore spatially varying, allowing the relative contribution of the current and historical frames to differ across the range–azimuth grid.

The fused feature representation is computed as

$$F_{\text{fused},t} = \sum_{i\in\{t-2,t-1,t\}} \alpha_i \odot F_i, \tag{3.58}$$

where $\odot$ denotes element-wise multiplication with broadcasting over the feature-channel dimension. The fusion preserves the range–azimuth layout:

$$F_{\text{fused},t} \in \mathbb{R}^{B\times 16\times R\times A}. \tag{3.59}$$

The fused representation is subsequently processed by the same U-Net 3+ spatial prediction backbone used by the single-frame model:

$$S_t = h_\omega \left(F_{\text{fused},t} \right), \tag{3.60}$$

where S_t denotes the first-return and no-hit score tensor for the current frame.

For simplicity, no ego-motion compensation or explicit spatial warping is applied before temporal fusion. The historical features are therefore combined in their native range–azimuth tensor coordinates. The extension is consequently interpreted as causal no-warp temporal feature fusion rather than as a motion-compensated multi-frame mapping method. Residual spatial misalignment caused by ego motion or independently moving objects may therefore limit the usefulness of historical features.

The training procedure for this extension, including initialization from the single-frame baseline, parameter freezing, and optimization of the temporal-fusion module, is specified in Section 3.2.3.

3.2 Experimental Setup

This section defines the experimental setup used to evaluate the proposed framework. It first introduces the RaDelft dataset, sensor configuration, and radar preprocessing used in this study. The reference methods are then described, followed by the training configuration for the proposed models and controlled extensions. Finally, the evaluation protocol defines the valid evaluation region, quantitative metrics, and computational-efficiency measurements used for comparison.

3.2.1 Dataset and Sensor Setup

The experiments use the RaDelft dataset [29], which contains synchronized automotive radar, LiDAR, camera, and ego-motion measurements collected in real-world driving scenarios. Radar measurements are used as the model input, while synchronized LiDAR point clouds provide the geometric reference for generating first-return occupancy labels. Camera images are used only for qualitative visualization and interpretation and do not contribute to network input or supervision. Ego-motion measurements are not used by the proposed single-frame or no-warp temporal models, but are used by the odometry-compensated inverse-sensor-model reference described in Section 3.2.2 and used for performance comparison.

Sensor Configuration

The radar measurements are acquired using a Texas Instruments MMWCAS-RF-EVM frequency-modulated continuous-wave (FMCW) imaging radar [14] operating around 76 GHz with 12 transmit and 16 receive antennas. The radar uses a bandwidth of approximately 750 MHz, 256 analog-to-digital converter (ADC) samples per chirp, and 128 chirps per frame, corresponding to a range resolution of approximately 0.2 m and a maximum range of about 51.4 m. The synchronized LiDAR

measurements are provided by a RoboSense Ruby Plus sensor and serve as the geometric reference for supervision.

Figure 3.10 shows representative synchronized observations from the dataset. The camera image provides visual context, the range–azimuth map provides an interpretable two-dimensional view of the radar response, and the LiDAR bird's-eye view illustrates the geometric measurement used for supervision. The displayed range–azimuth map is used only for visualization; the proposed full-RDA model receives the complete range–Doppler–azimuth radar cube.

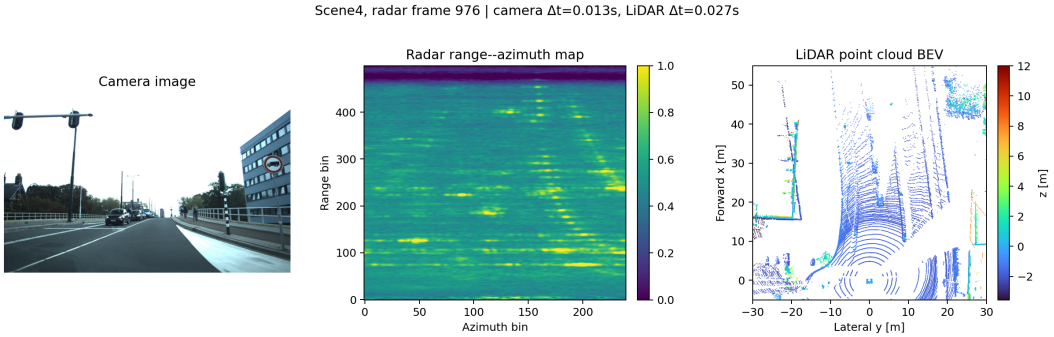

Figure 3.10: Representative synchronized observations from the RaDelft dataset, including the camera view, radar range–azimuth visualization, and LiDAR measurement used as the geometric basis for supervision in the proposed method.

The raw LiDAR point cloud is not directly used as an occupancy label because it contains road-surface returns, height variations, vegetation, elevated structures, and other measurements that do not necessarily correspond to the near-ground first-return obstacle evidence considered in this thesis. It is therefore processed into a radar-aligned first-return target using the pipeline described in Section 3.1.2.

Radar Data Preprocessing

Starting from the radar ADC measurements, the data are first reshaped according to the acquisition structure. A Hann window is applied along the fast-time dimension, followed by a 512-point range FFT to obtain range-resolved responses. A second Hann window is applied along the chirp dimension before a 128-point Doppler FFT. Because the radar employs time-division multiple-access (TDMA) MIMO acquisition, sequential transmitter activation introduces Doppler-dependent phase offsets between transmit subarrays for moving targets. A TDMA phase-correction procedure is therefore applied after Doppler processing and before angular processing to compensate for these motion-induced phase offsets [30].

For construction of the power RDA cube used as model input, the antenna data are first restricted to the virtual-array channels selected for azimuth estimation. A Hann window is applied along this array dimension, followed by a 256-point azimuth FFT. The central 240 azimuth bins are then retained by removing eight edge bins

on each side. The resulting radar representation therefore contains 500 range bins, 128 Doppler bins, and 240 azimuth bins, forming the preprocessed range–Doppler–azimuth radar cube.

The full radar cube forms the primary input to the proposed model and is formalized in Section 3.1.3. The Doppler-compressed range–azimuth variants described in Section 3.1.3 are derived from the same preprocessed cube.

3.2.2 Reference Methods

Two reference methods from the literature are used to contextualize the performance of the proposed framework. A traditional odometry-compensated multi-frame inverse sensor model provides a model-based baseline based on constant false-alarm rate (CFAR) detections and log-odds accumulation, whereas a RaDelft-adapted PolarNet implementation provides a learning-based comparison in polar radar coordinates. Both references are evaluated over the same valid spatial region and with the same quantitative metrics as the proposed framework.

1. Odometry-Compensated Multi-Frame Inverse Sensor Model

A traditional inverse sensor model (ISM) is implemented as a model-based reference for the proposed learning-based framework. For each radar frame, a two-stage ordered-statistic CFAR procedure is applied. The first stage identifies candidate detections on the range–azimuth plane. For each candidate range–azimuth cell, a second one-dimensional OS-CFAR test is then applied along its Doppler spectrum, and the candidate is retained only if a significant Doppler-domain response is also detected. The resulting binary detection mask is denoted by

$$D_t(r,\theta) \in \{0,1\}, \tag{3.61}$$

where t, r, and θ denote the frame index, range bin, and azimuth bin, respectively.

For each azimuth direction containing at least one valid detection, the nearest detected range bin is selected as the first hit:

$$r_{\text{hit}}^{(t)}(\theta) = \min\{r \mid D_t(r,\theta) = 1\}. \tag{3.62}$$

Cells preceding the first hit receive free-space evidence, while a Gaussian-shaped occupied contribution is assigned around the detected boundary. Cells behind the first hit remain unknown. If no detection is present in an azimuth column, the complete column is treated as unknown and contributes neither free nor occupied evidence. This prevents repeated missed detections from being accumulated as free-space observations.

The resulting per-frame ISM is represented in log-odds form. Historical maps are transformed into the coordinate frame of the current radar measurement using interpolated ego-motion information and accumulated over the most recent K frames:

$$L_t^{\mathrm{acc}}(r,\theta) = \sum_{j=0}^{K-1} \widetilde{L}_{t-j\to t}(r,\theta), \tag{3.63}$$

where $\widetilde{L}_{t-j\to t}$ denotes the log-odds evidence from frame $t-j$ after odometry-based alignment to frame t.

The accumulated log-odds map is converted back to occupancy probabilities and thresholded into free, occupied, and unknown states. For first-return evaluation, the nearest occupied cell along each azimuth direction defines the predicted boundary; cells before the boundary are rendered as free and cells behind it as unknown.

Notably, the final reference configuration of this approach uses $K = 25$ frames and treats no-hit azimuth columns as unknown. This configuration is referred to as the *odometry-compensated multi-frame ISM*.

2. RaDelft-Adapted PolarNet Reference

A PolarNet-style model is implemented as a learning-based reference for the proposed framework. PolarNet was originally introduced for binary open-space segmentation in polar radar coordinates [22]. The implementation used in this work retains its main architectural principles, including anisotropic range- and azimuth-wise convolutions, an encoder–decoder structure, and a large range-wise smoothing layer. The input representation, supervision, and training procedure are adapted to the RaDelft first-return task; the model is therefore treated as a RaDelft-adapted reference rather than an exact reproduction of the original benchmark.

The input is a single-channel range–azimuth map obtained by maximum aggregation over the Doppler dimension:

$$X_{\mathrm{RA}}(r,\theta) = \log\left(1 + \max_d |X_{\mathrm{RAD}}(d,r,\theta)|\right). \tag{3.64}$$

Each map is standardized independently and restricted to the valid 500×240 range–azimuth support. The three-state first-return target is converted into binary open-space supervision:

$$Y_{\mathrm{PN}}(r,\theta) = \begin{cases} 1, & Y(r,\theta) = \text{free}, \\ 0, & Y(r,\theta) \in \{\text{occupied}, \text{unknown}\}. \end{cases} \tag{3.65}$$

No-hit columns are therefore labelled entirely as open. The valid spatial region defined in Section 3.1.2 is applied during both training and evaluation.

The encoder uses successive 5×1 range-wise and 1×5 azimuth-wise convolutions, while the decoder reconstructs the range–azimuth output using bilinear interpolation, skip concatenation, and 3×3 convolutions. A 31×1 range-wise smoothing layer followed by a 1×1 output layer produces one open-space logit per cell. A sigmoid threshold of 0.5 is used for binary prediction.

The model is trained using masked binary cross-entropy with a batch-dependent class weight for open-space pixels. AdamW is used with an initial learning rate of

10^{-3}, weight decay 10^{-4}, batch size 8, and a maximum of 80 epochs. The learning rate is reduced when validation mIoU stops improving, and the checkpoint with the highest validation mIoU is retained.

For first-return evaluation, the nearest cell classified as non-open in each valid azimuth column is taken as the predicted boundary. If all valid cells in a column are classified as open, the column is assigned to the no-hit state. The predictions are evaluated using the occupancy and first-return metrics defined in Section 3.2.4.

3.2.3 Training Configuration

This section defines the training protocol for the proposed single-frame framework and its controlled extensions to three frames. Notably, the RaDelft-adapted PolarNet reference follows the separate binary-segmentation training configuration described in Section 3.2.2.

Scene Split and Common Supervision

The dataset is divided at the scene level to prevent frame-level leakage between the training, validation, and test subsets. For the primary experimental split, Scenes 1, 3, 4, 5, and 7 are used for training, Scene 2 is used for validation, and Scene 6 is reserved for final testing. This split is used consistently for the main model development and controlled comparisons. The validation set is used for training monitoring and checkpoint selection, whereas the test set is used exclusively for final evaluation.

To assess whether the reported conclusions depend on the choice of a particular test scene, an additional cross-scene evaluation is conducted in which each of the seven scenes is used once as the held-out test scene. For each split, one of the remaining scenes is used for validation and the other five scenes are used for training. The validation scene is used for checkpoint selection, while the corresponding held-out scene remains unseen during model optimization. This evaluation provides a broader assessment of cross-scene generalization and reduces the dependence of the conclusions on the original Scene 6 test split.

The proposed baseline and its controlled variants use the same LiDAR-derived first-return target construction and the same valid supervision region defined in Section 3.1.2. Consequently, padded tensor regions and the unreliable LiDAR-supervision sector are excluded consistently during optimization. The same supervision formulation is retained across the Doppler-representation comparisons, Radar-Aware Weighting, and Causal Temporal Feature Fusion. Unless explicitly stated otherwise, the primary Scene 6 split is used for the main controlled comparisons, while the additional cross-scene evaluation is used to assess the robustness of the conclusions across different held-out scenes.

Single-Frame Model Training

The proposed single-frame models are optimized using the Adam optimizer [16] with an initial learning rate of 1×10^{-4}.

A ReduceLROnPlateau scheduler monitors the validation loss and reduces the learning rate by a factor of 0.5 when the validation loss stops improving. The minimum learning rate is set to 1×10^{-6}.

The models are optimized using the combined Lovasz and smoothed CDF-L1 objective defined in Section 3.1.6. For the proposed baseline and the controlled first-return variants trained under this objective, the checkpoint with the lowest total validation loss is retained for final evaluation.

The Doppler-compressed input configurations and structured Doppler-encoder comparisons use the same scene split, target construction, valid supervision region, and downstream first-return prediction framework. Radar-Aware Weighting likewise retains the baseline architecture and global loss coefficients, with only the azimuth-wise weighting of the smoothed CDF-L1 term modified as described in Section 3.1.6.

The single-layer Doppler projection configuration was trained using an earlier objective setting and is therefore treated as a reference configuration rather than as a strictly controlled encoder-only ablation, as described in Section 3.1.4.

The RaDelft-adapted PolarNet reference follows its separate optimization and checkpoint-selection protocol, in which the checkpoint with the highest validation mIoU is retained, as specified in Section 3.2.2. In all cases, the test set is excluded from checkpoint selection.

Temporal-Fusion Training

The Causal Temporal Feature Fusion model is initialized from the selected single-frame baseline checkpoint. The Range–Azimuth-Preserving Doppler Spectrum Encoder and the U-Net 3+ prediction backbone are kept frozen, and only the temporal feature-fusion module is optimized at training. This training strategy isolates the effect of temporal feature aggregation from improvements that could otherwise result from re-optimizing the single-frame feature extractor or spatial prediction backbone.

For each training sample, the radar measurements at $t-2$, $t-1$, and t are provided as input, while the supervision target remains the LiDAR-derived first-return target corresponding to the current frame t. The same scene split, valid supervision region, baseline loss function, and validation-loss-based checkpoint-selection criterion are retained. The temporal-fusion architecture itself is defined in Section 3.1.7.

3.2.4 Evaluation Protocol and Metrics

The proposed framework and reference methods are evaluated using a common protocol covering occupancy classification, first-return boundary localization, prediction consistency, and computational efficiency. All quantitative metrics are computed over the same valid spatial support to ensure a fair comparison between methods.

Valid Evaluation Region

Quantitative evaluation is restricted to the valid radar-aligned region defined in Section 3.1.2. The corresponding binary evaluation mask is denoted by:

$$M_{\text{eval}}(r,\theta) = \begin{cases} 1, & \text{if the cell is included in evaluation,} \\ 0, & \text{otherwise.} \end{cases} \tag{3.66}$$

The evaluation region excludes padded tensor areas and is restricted to the physical support containing 500 range bins and 240 azimuth bins. The unreliable LiDAR-supervision sector from approximately $-63°$ to $-37°$, previously defined in Section 3.1.2, is also excluded. The same spatial support is used for all evaluated methods.

Occupancy metrics are computed only over cells satisfying $M_{\text{eval}}(r,\theta) = 1$. First-return distance metrics are computed only for azimuth columns belonging to the corresponding valid angular region.

Occupancy Classification Metrics

Occupancy classification is evaluated at the grid-cell level. Although the first-return representation contains free, occupied, and unknown states, the quantitative occupancy evaluation focuses on the distinction between visible free space and non-free space. The occupied and unknown states are therefore grouped into a single *unfree* category:

$$y_{\text{bin}}(r,\theta) = \begin{cases} \text{free}, & y(r,\theta) = \text{free}, \\ \text{unfree}, & y(r,\theta) \in \{\text{occupied}, \text{unknown}\}. \end{cases} \tag{3.67}$$

For each class $c \in \{\text{free}, \text{unfree}\}$, intersection over union is computed as

$$\text{IoU}_c = \frac{\text{TP}_c}{\text{TP}_c + \text{FP}_c + \text{FN}_c}, \tag{3.68}$$

where TP_c, FP_c, and FN_c denote the true-positive, false-positive, and false-negative cell counts for class c, respectively. All counts are restricted to the valid evaluation region.

The reported occupancy metrics are IoU_{free}, $\text{IoU}_{\text{unfree}}$, and their mean:

$$\text{mIoU} = \frac{1}{2}\left(\text{IoU}_{\text{free}} + \text{IoU}_{\text{unfree}}\right). \tag{3.69}$$

These metrics quantify the agreement between the predicted and LiDAR-derived occupancy structure, while the first-return metrics described below directly evaluate the geometric location of the nearest boundary.

First-Return Distance Metrics

First-return distance metrics evaluate the range accuracy of the predicted nearest boundary along each valid azimuth direction. This complements the occupancy metrics, since similar cell-wise overlap can still result from different boundary locations.

Let $t_{n,\theta}$ and $\hat{t}_{n,\theta}$ denote the ground-truth and predicted first-return indices, respectively, for test frame n and valid azimuth bin θ. Physical first-return locations are represented by indices $0, \ldots, R-1$, while the additional index R represents the no-hit state. The predicted index is obtained from the corresponding output representation of each evaluated method.

The range error for frame n and azimuth bin θ is

$$e_{n,\theta} = \left| \hat{t}_{n,\theta} - t_{n,\theta} \right| \Delta r, \tag{3.70}$$

where Δr denotes the radar range-bin resolution.

The frame-level first-return mean absolute error is

$$\mathrm{MAE}_n = \frac{1}{|\Theta_{\mathrm{valid}}|} \sum_{\theta \in \Theta_{\mathrm{valid}}} e_{n,\theta}, \tag{3.71}$$

where Θ_{valid} denotes the azimuth bins retained by the evaluation mask.

For N test frames, the reported mean first-return error is

$$\overline{\mathrm{MAE}} = \frac{1}{N} \sum_{n=1}^{N} \mathrm{MAE}_n. \tag{3.72}$$

To characterize prediction consistency across the test sequence, the population variance of the frame-level MAE values is additionally reported:

$$\mathrm{Var}_{\text{frame-MAE}} = \frac{1}{N} \sum_{n=1}^{N} \left(\mathrm{MAE}_n - \overline{\mathrm{MAE}} \right)^2. \tag{3.73}$$

A lower variance indicates more consistent boundary-localization performance across test frames. This metric describes variation between frame-level errors rather than variation between individual azimuth-column errors within a frame.

Computational Cost Analysis

Computational efficiency is evaluated for the learning-based methods using parameter count, multiply–accumulate operations in GMACs, GPU inference latency, throughput in frames per second, and peak GPU memory usage. Parameter counts include all components required during inference, including modules that were frozen during training.

All models are profiled with batch size one using FP32 inference on an NVIDIA A100 GPU. Each method is evaluated using its corresponding inference input: a single-channel range–azimuth map for the RaDelft-adapted PolarNet reference, one

range–Doppler–azimuth cube for the proposed single-frame baseline, and three consecutive radar cubes for the Causal Temporal Feature Fusion model.

Importantly, latency measurements cover only the neural-network forward pass after GPU warm-up. Data loading, radar preprocessing, CPU-to-GPU transfer, occupancy rendering, metric computation, and visualization are excluded. Throughput is derived from the measured mean forward-pass latency.

For the temporal-fusion model, the reported cost corresponds to the implemented no-cache inference procedure, in which the radar cubes at $t-2$, $t-1$, and t are processed for every prediction. Radar-Aware Weighting is not profiled separately because it modifies only the training objective and therefore has the same inference architecture and computational cost as the proposed single-frame baseline.

3.3 Summary

This chapter presented the methodology for LiDAR-supervised radar-based first-return occupancy grid mapping. The proposed framework converts synchronized LiDAR measurements into radar-aligned first-return supervision and formulates radar occupancy estimation as an azimuth-wise first-return or no-hit prediction problem. The complete range–Doppler–azimuth radar measurement is processed using the proposed Range–Azimuth-Preserving Doppler Spectrum Encoder, followed by a U-Net 3+ spatial backbone and visibility-consistent occupancy rendering. Training combines occupancy-oriented Lovasz supervision with a smoothed CDF-L1 term that directly supervises first-return boundary localization.

Two controlled extensions were additionally defined. Radar-Aware Weighting modifies only the azimuth-wise contribution of the CDF-based boundary term, while Causal Temporal Feature Fusion incorporates short-term radar context through feature-level fusion of three consecutive frames.

The experimental setup was defined using the RaDelft dataset, with an odometry-compensated multi-frame inverse sensor model and a RaDelft-adapted PolarNet serving as model-based and learning-based references, respectively. Common scene splits, valid supervision and evaluation regions, occupancy metrics, first-return distance metrics, and computational-efficiency measurements were specified to support consistent comparison across the evaluated configurations.

The following chapter presents the quantitative and qualitative results of these experiments, including comparisons with the reference methods, Doppler representation and encoding analyses, and evaluation of the two controlled extensions.

Chapter 4

Results

This chapter presents the experimental evaluation of the proposed LiDAR-supervised radar-based first-return occupancy grid mapping framework. The results are organized into three main parts. First, the proposed single-frame method is evaluated on the primary held-out RaDelft test scene against the traditional odometry-compensated multi-frame ISM and the RaDelft-adapted PolarNet reference. A leave-one-scene-out evaluation is then conducted across all seven RaDelft scenes to assess the consistency of the proposed method across different held-out environments. Second, a series of ablation studies and controlled experiments analyse the contributions of the loss formulation, radar input representation, Doppler encoding strategy, and the proposed extensions. Qualitative examples and failure cases are incorporated into the corresponding quantitative analyses where they provide additional insight into model behaviour. Finally, the computational requirements of the learning-based configurations are examined to assess the trade-off between predictive performance and inference efficiency. All reported metrics are computed over the common valid evaluation region defined in Section 3.2.4.

4.1 Evaluation of the Proposed Method

4.1.1 Performance on the Held-Out Test Scene

To provide results at a first glance, over the complete held-out RaDelft test scene (scene #6), the proposed single-frame method achieves a mean (across frames) $\mathrm{IoU}_{\mathrm{free}}$ of 0.79 and a mean (across frames) $\mathrm{IoU}_{\mathrm{unfree}}$ of 0.82, resulting in an mIoU of approximately 0.80. The mean first-return MAE is 5.38 m, while the population variance of the per-frame MAE is 6.52 m^2. The comparable free- and unfree-space IoU values indicate that the overall occupancy performance is not obtained by favouring only one side of the first-return boundary. The boundary-level metric further shows that the transition between visible free space and the unobserved region can be localized more precisely than would be apparent from cell-wise occupancy overlap alone.

To place these numbers in context, Table 4.1 compares the performance of the proposed method with the two reference approaches defined in Section 3.2.2: the odometry-compensated multi-frame ISM and the RaDelft-adapted PolarNet. The ISM represents a conventional model-based radar mapping pipeline based on CFAR detections and temporal log-odds accumulation, whereas PolarNet provides a learning-based reference operating on a Doppler-compressed range–azimuth representation. All three methods are evaluated on the same held-out test scene and within the common valid evaluation region, as discussed in the previous chapter.

Table 4.1: Comparison of the proposed method with the odometry-compensated multi-frame ISM and the RaDelft-adapted PolarNet reference on the RaDelft test scene. Mean quantities are computed across all frames, and the variance is computed across the per-frame first-return MAE values.

Method	$\mathrm{IoU}_{\mathrm{free}}$ ↑	$\mathrm{IoU}_{\mathrm{unfree}}$ ↑	mIoU ↑	MAE [m] ↓	Frame-MAE Var. m^2 ↓
Multi-frame ISM	0.47	0.62	0.54	14.20	–
RaDelft-adapted PolarNet	0.57	0.70	0.64	10.89	10.97
Proposed method	**0.79**	**0.82**	**0.80**	**5.38**	**6.52**

The proposed method substantially outperforms both reference approaches. Relative to the multi-frame ISM, the mIoU increases from 0.54 to 0.80, while the mean first-return MAE decreases from 14.20 m to 5.38 m. Compared with PolarNet, the proposed method improves the mIoU from 0.64 to 0.80 and reduces the mean first-return MAE from 10.89 m to 5.38 m. The lower frame-MAE variance compared with PolarNet additionally indicates more consistent boundary localization across the test sequence.

The performance gap is consistent with the different ways in which the methods use the radar measurements. The ISM first reduces the radar observation to sparse CFAR detections and then determines the nearest detected cell along each azimuth direction. Missed detections may therefore shift the estimated frontier to a larger range, while false detections may introduce an erroneously close boundary. Multi-frame log-odds accumulation increases the amount of evidence available over time, but it cannot recover information that has already been removed by the detection thresholding stage.

PolarNet provides a stronger reference than the ISM but also differs substantially from the proposed framework. It compresses the Doppler dimension through maximum aggregation before inference and predicts binary open space at the range–azimuth cell level. In contrast, the proposed method retains the complete Doppler-resolved measurement and formulates the output directly as one first-return or no-hit state per azimuth column. The predicted occupancy grid is then rendered according to the corresponding visibility structure. The comparison therefore reflects the complete proposed framework rather than the contribution of any single component; these components are examined separately in Section 4.2.

Aggregate metrics averaged over all frames do not show how prediction quality varies throughout the test sequence. Figures 4.1 and 4.2 therefore show the frame-wise mIoU and first-return MAE for all three methods. Faint curves represent the original per-frame measurements, whereas the solid curves show a centred 30-frame moving average, corresponding to approximately 3 s at the 10 Hz dataset frame rate, to make the local performance trends easier to interpret. The mean values reported in the legends are calculated from the original per-frame measurements, as reported in Table 4.1.

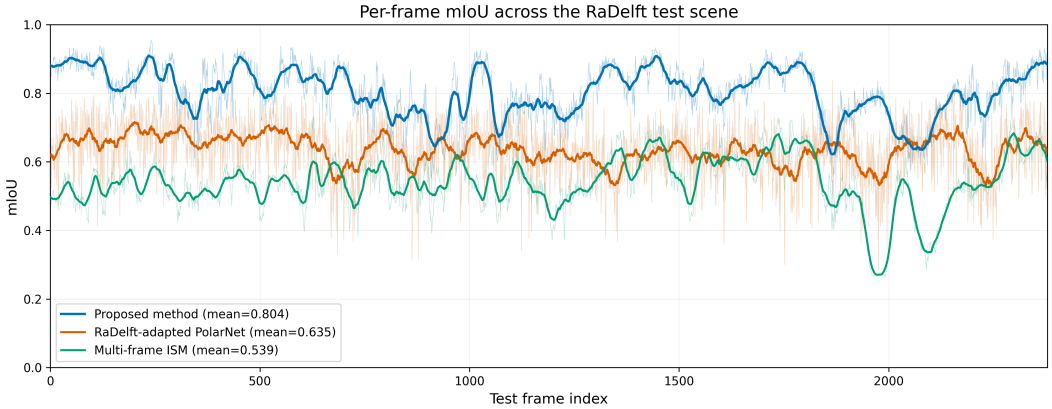

Figure 4.1: Per-frame mIoU of the proposed method, the RaDelft-adapted PolarNet, and the multi-frame ISM across the RaDelft test scene. Faint curves show the original frame-wise measurements, while solid curves show a centred 30-frame moving average. The mean values in the legend are computed from the original per-frame measurements.

Figure 4.1 shows that the proposed method maintains consistently higher occupancy agreement over most of the test sequence. Its local trend generally remains around or above an mIoU of 0.8, with several intervals approaching 0.9. Performance nevertheless varies with the observed scene. Several extended intervals exhibit lower occupancy agreement, most notably around approximately frames 800–1100 and 1800–2100. Inspection of sampled camera frames from these intervals indicates that they frequently correspond to road segments with less clearly defined lateral boundaries, including tree-lined roads, vegetation, grass verges, and locally irregular or discontinuous roadside structures. Such environments can produce more fragmented first-return geometry and less spatially distinct radar responses than scenes with continuous and well-defined road boundaries. The reference methods exhibit related variations in several of these intervals, further suggesting that part of the performance degradation is associated with scene-dependent sensing difficulty rather than isolated model errors.

The sequence-level comparison further shows that the advantage of the proposed method is not restricted to a small number of favourable frames. Although the difference between the methods becomes smaller in some challenging intervals, the

proposed method remains above the two references over most of the sequence. PolarNet generally forms the intermediate performance level, whereas the multi-frame ISM exhibits lower overall occupancy agreement and stronger degradation in several scene segments.

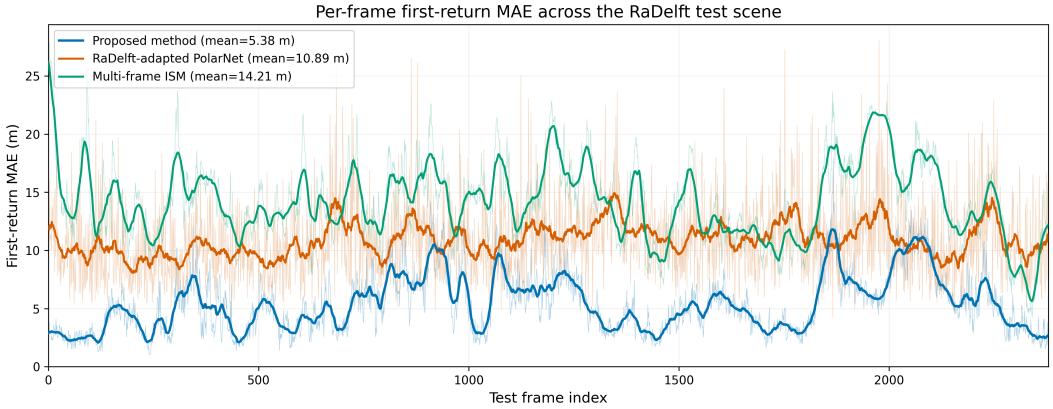

Figure 4.2: Per-frame first-return MAE of the proposed method, the RaDelft-adapted PolarNet, and the multi-frame ISM across the RaDelft test scene. Faint curves show the original frame-wise measurements, while solid curves show a centred 30-frame moving average. The mean values in the legend are computed from the original per-frame measurements.

The first-return errors in Figure 4.2 show a corresponding pattern. The proposed method maintains the lowest local MAE trend over most of the test sequence. Higher-error intervals are still present, including several regions where the smoothed MAE approaches or exceeds 10 m, but these increases are generally accompanied by substantially larger errors in one or both reference methods.

The temporal variation in the two metrics is broadly consistent. Intervals with reduced mIoU often coincide with increased first-return MAE, indicating that deterioration in occupancy agreement is associated with greater displacement of the predicted first-return boundary. Nevertheless, the metrics capture different properties of the prediction. mIoU measures disagreement over the free–unfree regions, whereas first-return MAE directly measures the range displacement of the nearest predicted boundary. Frames with similar occupancy overlap can therefore still differ substantially in boundary-localization error.

The quantitative trends are further illustrated by representative qualitative examples. In the following three figures, the yellow region denotes the LiDAR-derived free-space region, the white curve denotes the corresponding ground-truth first-return frontier, and the cyan curve denotes the predicted frontier. The shaded azimuth sector corresponds to the region excluded from training and evaluation because of unreliable LiDAR supervision.

First, figure 4.3 shows a representative high-performing frame.

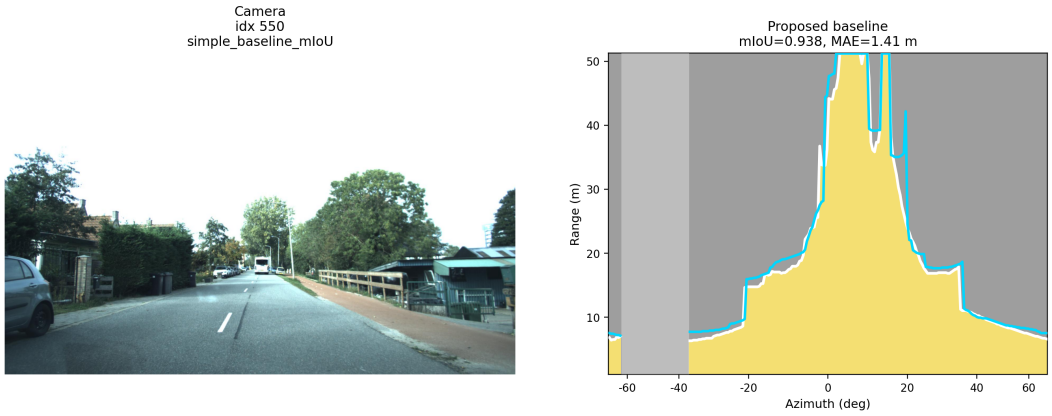

Figure 4.3: Representative high-performing prediction of the proposed method. The yellow region denotes the LiDAR-derived free-space region, the white curve denotes the ground-truth first-return frontier, and the cyan curve denotes the predicted frontier. The proposed method achieves an mIoU of 0.938 and a first-return MAE of 1.41 m in this frame.

In this relatively regular road scene, the predicted frontier follows the LiDAR-derived boundary closely over most azimuth directions. Both the gradual variation of the road-side boundary and several larger range transitions are reproduced with limited displacement. The close alignment between the white and cyan curves is consistent with the high mIoU of 0.938 and the low first-return MAE of 1.41 m. This example illustrates that, when the radar observations provide sufficiently clear spatial evidence, the proposed method can recover both the global free-space geometry and local variations in the first-return boundary.

Then, a more challenging example is shown in Figure 4.4.

In the more challenging scene, the ground-truth frontier contains stronger local variation and several abrupt changes in first-return range across neighbouring azimuth directions. The prediction still captures the dominant large-scale free-space structure, particularly where the boundary varies gradually, but larger discrepancies occur around sharp range transitions and narrow local structures. In several azimuth regions, the predicted frontier remains spatially smooth while the LiDAR-derived frontier changes rapidly, resulting in larger local range errors. The corresponding mIoU of 0.758 and first-return MAE of 6.69 m are therefore lower and higher, respectively, than in the high-performing example, while still showing that the principal scene geometry is retained.

Figure 4.5 further compares the three methods on a common representative scene. The camera image is included only to provide visual context.

The qualitative comparison is consistent with the quantitative results. The ISM captures the broad free-space structure but produces a more fragmented and displaced frontier in several azimuth regions. PolarNet yields a smoother prediction, although noticeable deviations from the LiDAR-derived boundary remain. The pro-

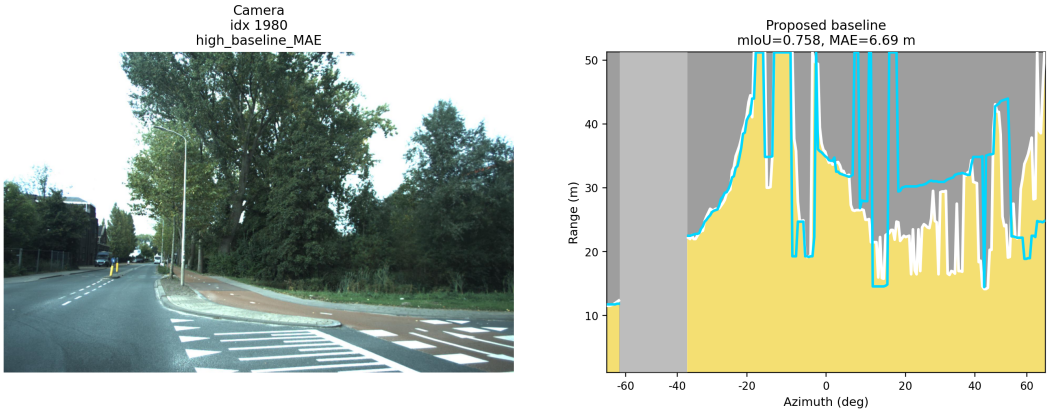

Figure 4.4: Representative challenging prediction of the proposed method. The yellow region denotes the LiDAR-derived free-space region, the white curve denotes the ground-truth first-return frontier, and the cyan curve denotes the predicted frontier. The proposed method achieves an mIoU of 0.758 and a first-return MAE of 6.69 m in this frame.

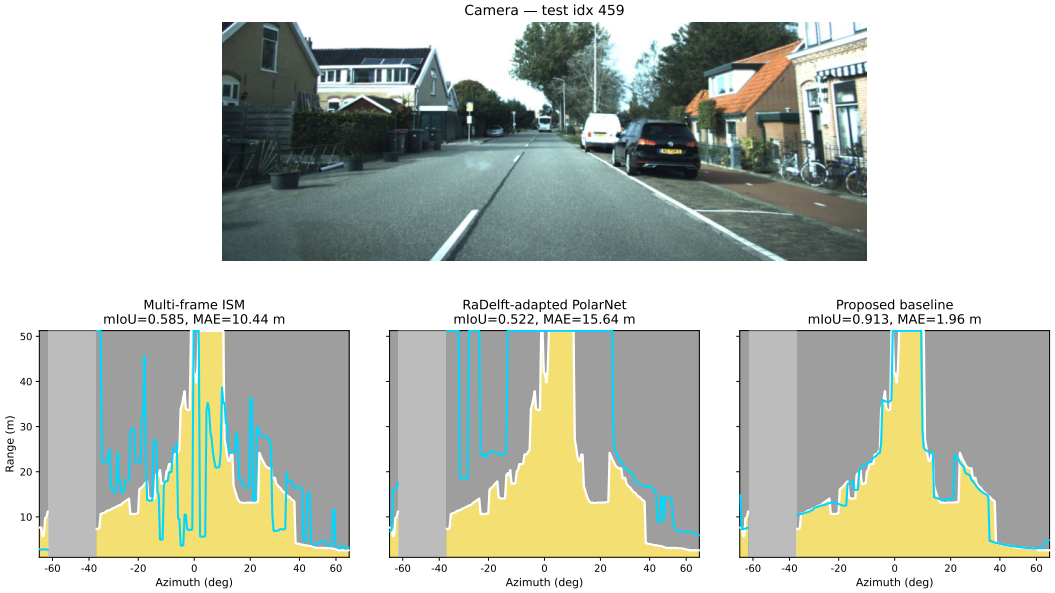

Figure 4.5: Qualitative comparison of the odometry-compensated multi-frame ISM, the RaDelft-adapted PolarNet, and the proposed method. In each polar representation, the white curve denotes the LiDAR-derived ground-truth first-return frontier and the cyan curve denotes the predicted frontier.

posed method follows the dominant first-return structure more closely over a larger portion of the field of view and better preserves local changes in boundary range.

Overall, the proposed method provides more accurate and consistent occupancy estimation and first-return localization than both the traditional model-based reference and the adapted learning-based reference. Its strongest predictions occur when the first-return structure is supported by clear radar evidence, while abrupt boundary changes, narrow structures, and locally ambiguous radar responses remain more challenging. The following section examines the contributions of the individual design choices and controlled extensions through a series of ablation studies.

4.1.2 Cross-Scene Generalization

The preceding evaluation uses scene #6 as the held-out test scene. To examine whether the performance of the proposed method depends strongly on this particular test scene, an additional leave-one-scene-out evaluation is conducted across all seven RaDelft scenes. In each fold, one complete scene is held out for testing, while the remaining scenes are used for model development according to the corresponding training and validation split. The same proposed single-frame architecture and evaluation protocol are retained across all folds. This experiment therefore provides a scene-level assessment of generalization to previously unseen driving environments.

Table 4.2 summarizes the results. The final row reports the macro mean and standard deviation across the seven held-out scenes, such that each scene contributes equally irrespective of its number of frames.

Table 4.2: Leave-one-scene-out evaluation of the proposed single-frame method across the seven RaDelft scenes. Each row reports the performance obtained when the corresponding scene is held out for testing. The final row gives the macro mean and standard deviation across scenes.

Held-out scene	$\mathrm{IoU}_{\mathrm{free}}$ ↑	$\mathrm{IoU}_{\mathrm{unfree}}$ ↑	mIoU ↑	MAE [m] ↓	Frame-MAE Var. m^2 ↓
Scene 1	0.76	0.69	0.72	7.58	6.86
Scene 2	0.78	0.81	0.80	5.43	5.74
Scene 3	0.73	0.90	0.81	3.97	3.17
Scene 4	0.76	0.77	0.77	6.50	3.84
Scene 5	0.79	0.83	0.81	5.14	4.56
Scene 6	0.78	0.82	0.80	5.38	6.54
Scene 7	0.70	0.81	0.76	6.28	5.73
Macro mean ± std	**0.76 ± 0.03**	**0.81 ± 0.06**	**0.78 ± 0.03**	**5.75 ± 1.07**	**5.21 ± 1.28**

Across the seven held-out scenes, the proposed method achieves a macro-average mIoU of 0.78 ± 0.03 and a first-return MAE of 5.75 ± 1.07 m. Although some scene-dependent variation is present, the occupancy agreement remains within a relatively compact range across most scenes. The mean values are also close to those obtained on the primary held-out test scene, where the proposed method achieves an mIoU

of approximately 0.80 and a first-return MAE of 5.38 m. This indicates that the performance observed in the main evaluation is not restricted to a single favourable test sequence.

The largest difference is observed between Scenes 1 and 3. Scene 3 provides the strongest boundary-localization result, with an mIoU of 0.81 and a first-return MAE of 3.97 m, whereas Scene 1 is the most challenging, with an mIoU of 0.72 and an MAE of 7.58 m. Inspection of the LiDAR-derived first-return distributions shows that Scene 1 contains substantially more long-range and no-hit azimuth columns, corresponding predominantly to more open road environments in which the visible free-space region extends farther from the radar. In contrast, Scene 3 contains more nearby roadside and urban structures, resulting in generally shorter first-return distances and a much lower proportion of no-hit rays. These differences make precise frontier localization more demanding in Scene 1, since an erroneous early return along a long or no-hit ray can produce a large range error. Local abrupt changes and narrow structures remain challenging in both scenes and contribute to additional frame-level variation.

Overall, the leave-one-scene-out results provide additional evidence that the proposed single-frame framework generalizes across the different RaDelft scenes rather than relying on the characteristics of the primary test scene. At the same time, the remaining variation indicates that first-return localization is sensitive to the underlying scene geometry and sensing conditions, particularly in environments with long visible ranges and a high proportion of no-hit directions. Together with the fixed-test-scene results above, this supports the use of the proposed architecture as the principal configuration for the subsequent ablation and controlled-extension analyses.

4.2 Ablation Studies and Discussion

Having established the performance of the complete proposed method, this section examines the design choices that contribute to its behaviour. The analysis first investigates the loss function, including the contribution of its individual terms and the effect of transferring the proposed formulation to the RaDelft-adapted PolarNet architecture. The radar representation and Doppler-processing strategy are then analysed through comparisons between fixed Doppler-compressed inputs, alternative full-RDA encoders, and a corresponding transfer of the proposed Doppler Spectrum Encoder to PolarNet.

These experiments are complemented by a component-level study on PolarNet, in which the proposed Doppler encoder and adapted learning objective are introduced individually and jointly. This provides an additional cross-architecture test of whether the observed improvements arise from universally transferable components or from their integration within the complete structured first-return framework. The section subsequently evaluates Radar-Aware Weighting and Causal Temporal Feature Fusion as controlled extensions of the selected single-frame model. Qualitative

examples and failure cases are incorporated into the corresponding analyses where they help explain the quantitative behaviour. Finally, the computational cost of the learning-based configurations is examined to assess the trade-off between predictive performance and inference efficiency.

4.2.1 Analysis of the Loss Function

The loss function is analysed by comparing the proposed combined formulation with several alternative supervision strategies. The evaluated configurations include the Huber loss [13], the reverse Huber (BerHu) loss [18], Lovasz-only supervision [2], CDF-only supervision, smoothed-CDF-only supervision, and the proposed combination of normalized Lovasz and smoothed CDF-L1 losses. All configurations use the same radar input, Doppler encoder, prediction backbone, scene split, supervision target, and evaluation protocol, so that the comparison focuses on the effect of the overall proposed loss.

Table 4.3 summarizes the results.

Table 4.3: Comparison of different loss function formulations on the RaDelft test scene. The variance is computed across the per-frame first-return MAE values.

Learning objective	IoU_{free} ↑	IoU_{unfree} ↑	mIoU ↑	MAE [m] ↓	Frame-MAE Var. m^2 ↓
BerHu	0.73	0.76	0.75	7.14	10.51
Huber	0.74	0.77	0.76	6.77	10.63
Lovasz	0.78	0.81	0.79	5.60	7.28
CDF	0.78	0.80	0.79	5.74	6.58
Smoothed CDF	0.78	0.81	0.80	5.57	6.90
Proposed combined loss	**0.79**	**0.82**	**0.80**	**5.38**	**6.52**

A clear difference is observed between the generic regression losses and those that explicitly reflect the occupancy or first-return structure of the task. BerHu and Huber produce the weakest results among the evaluated configurations. BerHu achieves an mIoU of 0.75 and a first-return MAE of 7.14 m, while Huber improves these values to 0.76 and 6.77 m, respectively. Both configurations also exhibit substantially larger frame-MAE variance than the task-aligned alternatives. These results indicate that directly penalizing boundary-distance errors with a generic regression loss does not sufficiently capture the structured free–occupied–unknown relationship induced by the first-return formulation.

In contrast, both occupancy-oriented and cumulative first-return supervision provide substantially stronger performance. Lovasz reaches an mIoU of 0.79 and a first-return MAE of 5.60 m. The relatively high IoU_{unfree} of 0.81 is consistent with the role of the Lovasz term in directly optimizing occupancy-region overlap. CDF also performs strongly, with an mIoU of 0.79, an MAE of 5.74 m, and a frame-MAE variance of 6.58 m^2. This shows that direct supervision of the cumulative first-

return structure is sufficient to recover much of the required occupancy geometry even without an explicit occupancy-overlap term.

Applying linear smoothing to the CDF target improves the mean predictive accuracy relative to the unsmoothed formulation. Smoothed CDF increases the mIoU from 0.79 to 0.80 and reduces the mean first-return MAE from 5.74 m to 5.57 m. This behaviour is consistent with the motivation for smoothing the LiDAR-derived boundary target: small range-bin discrepancies around the first-return position are penalized less abruptly, reducing sensitivity to discretization and local supervision uncertainty. The frame-MAE variance increases slightly from 6.58 m^2 to 6.90 m^2, however, indicating that the improvement is primarily reflected in average localization accuracy rather than across-frame consistency.

The strongest overall result is obtained by combining the normalized Lovasz term with the smoothed CDF-L1 term. The combined loss reaches an mIoU of 0.80, a mean first-return MAE of 5.38 m, and a frame-MAE variance of 6.52 m^2. Relative to Lovasz, the combined loss reduces the mean boundary error by approximately 0.22 m and lowers the frame-MAE variance. Relative to smoothed CDF, it further improves both occupancy overlap and mean boundary localization.

These results support the complementary roles of the two loss components. The Lovasz term provides occupancy-oriented supervision of the rendered free and unknown regions, whereas the smoothed CDF-L1 term directly constrains the range-wise location of the first-return boundary. Neither component alone clearly dominates across all metrics. Their combination instead provides the most favourable balance between cell-wise occupancy agreement, boundary-localization accuracy, and prediction consistency, motivating its use as the default loss function of the proposed method.

4.2.2 Analysis of Radar Representation and Doppler Encoding

The initial radar processing steps, i.e., before the backbone network, are analysed from two complementary perspectives. First, the effect of retaining the Doppler-resolved radar measurement is examined by comparing fixed Doppler-compressed range–azimuth representations with the full range–Doppler–azimuth input used by the proposed method. Second, given the full RDA representation, alternative strategies for encoding the Doppler dimension are compared before range–azimuth spatial prediction.

Effect of Radar Input Representation

The first experiment examines the effect of compressing the Doppler dimension before spatial prediction. Two fixed range–azimuth representations are considered. Doppler-mean aggregation averages the response over all Doppler bins (mean pool), whereas Doppler-maximum aggregation retains the strongest response at each range–azimuth location (max pool). Both variants produce a single-channel range–azimuth input that is passed directly to the U-Net 3+ first-return prediction backbone.

These configurations are compared with the proposed full-RDA setting, in which the complete Doppler spectrum is retained and transformed into range–azimuth features by the proposed Doppler Spectrum Encoder. All configurations use the same scene split, LiDAR-derived supervision, valid evaluation region, and first-return evaluation protocol. Since the compressed-input variants remove both the Doppler dimension and the corresponding learnable Doppler encoder, this experiment compares complete radar-input processing configurations rather than constituting a strictly representation-only ablation.

Table 4.4: Comparison of fixed Doppler-compressed range–azimuth representations and the proposed full-RDA configuration on the RaDelft test scene. The variance is computed across the per-frame first-return MAE values.

Input configuration	$\mathrm{IoU}_{\mathrm{free}}$ ↑	$\mathrm{IoU}_{\mathrm{unfree}}$ ↑	mIoU ↑	MAE [m] ↓	Frame-MAE Var. m^2 ↓
Doppler mean R–A	0.72	0.78	0.75	7.12	10.00
Doppler maximum R–A	0.75	0.80	0.77	6.35	8.87
Full R–D–A with proposed encoder	**0.79**	**0.82**	**0.80**	**5.38**	**6.52**

Among the fixed Doppler-compression strategies, maximum aggregation consistently outperforms mean aggregation. Doppler maximum reaches an mIoU of 0.774 and a mean first-return MAE of 6.35 m, compared with 0.750 and 7.12 m, respectively, for Doppler mean. The frame-MAE variance is also reduced from 10.00 m^2 to 8.87 m^2. These results indicate that retaining the strongest Doppler response at each spatial location is more effective than averaging the complete Doppler spectrum for the present task. Mean aggregation can suppress localized responses by averaging them with weaker Doppler bins, whereas maximum aggregation preserves at least the dominant response at each range–azimuth location.

The full-RDA configuration provides a further improvement across all reported metrics. Relative to Doppler maximum, the proposed configuration increases the mIoU from 0.774 to approximately 0.80, reduces the mean first-return MAE by 0.97 m, and decreases the frame-MAE variance from 8.87 m^2 to 6.52 m^2. The reduction in both mean error and across-frame variance indicates that retaining the Doppler-resolved measurement benefits not only average boundary localization but also the consistency of the predictions across the test sequence.

Figure 4.6 provides a representative qualitative example in which the difference between the three input configurations is particularly clear. The yellow region denotes the LiDAR-derived free-space region, the white curve denotes the ground-truth first-return frontier, and the cyan curve denotes the corresponding model prediction.

In this frame, all three configurations recover the dominant large-scale free-space structure, but clear differences appear in the localization of individual first-return transitions. The proposed full-RDA configuration follows the LiDAR-derived frontier closely over most azimuth directions, including the gradual boundary variation around the central field of view and several sharper range transitions. This results

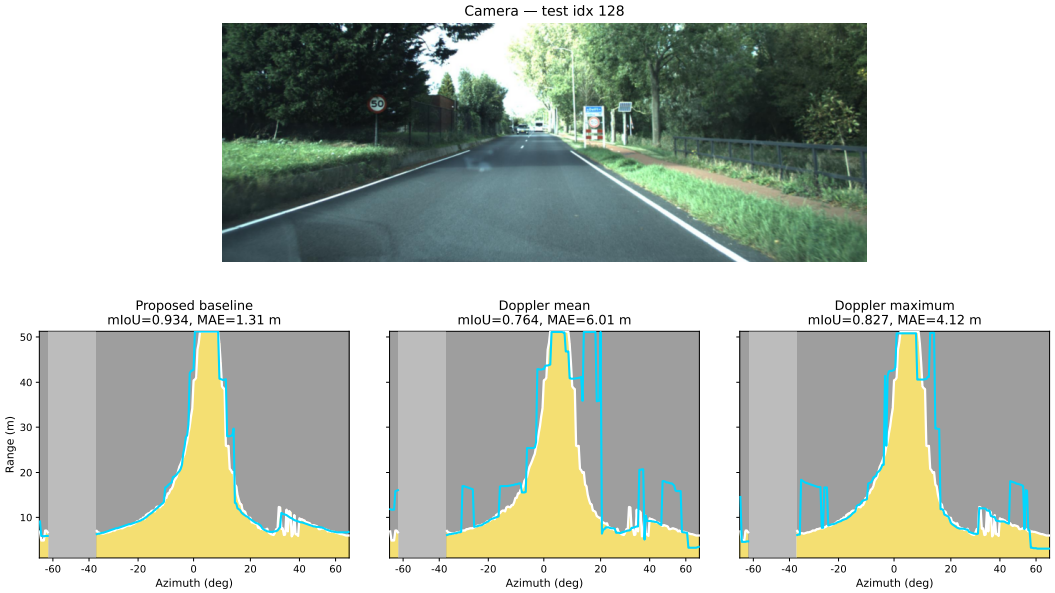

Figure 4.6: Qualitative comparison of the proposed full-RDA configuration with Doppler-mean and Doppler-maximum range–azimuth inputs. The yellow region denotes the LiDAR-derived free-space region, the white curve denotes the ground-truth first-return frontier, and the cyan curve denotes the predicted frontier. The proposed configuration (left) achieves an mIoU of 0.934 and a first-return MAE of 1.31 m, compared with 0.764 and 6.01 m for Doppler mean (middle), and 0.827 and 4.12 m for Doppler maximum (right).

in an mIoU of 0.934 and a first-return MAE of only 1.31 m.

Doppler-mean aggregation exhibits substantially larger deviations from the ground-truth frontier. In particular, several abrupt predicted range changes appear in regions where the ground-truth boundary varies more smoothly, resulting in an mIoU of 0.764 and an MAE of 6.01 m. Doppler maximum preserves the boundary more effectively than Doppler mean, consistent with the aggregate results, but still introduces several local first-return displacements and obtains an mIoU of 0.827 and an MAE of 4.12 m.

The qualitative comparison therefore reflects the trend observed over the complete test scene. Fixed Doppler compression can preserve sufficient spatial information to recover the overall free-space geometry, particularly when maximum aggregation is used, but it discards the internal structure of the Doppler spectrum before learning. By retaining the complete Doppler-resolved measurement, the proposed configuration allows the network to learn which spectral responses are informative for first-return estimation instead of committing to a fixed mean or maximum projection in advance.

Overall, the results support the use of the full RDA representation in the pro-

posed framework. However, because retaining the Doppler dimension also requires a learnable Doppler encoder, the improvement cannot be attributed solely to the input representation, i.e. to the fact that there are more Doppler values than just a mean or maximum value. The effect of the Doppler encoding architecture processing such data is therefore examined separately in the following experiment.

Effect of Doppler Encoder Design

The second experiment compares alternative strategies for transforming the retained Doppler spectrum into range–azimuth features. The evaluated configurations are the single-layer Doppler projection, the Roldán-style three-dimensional Doppler encoder adapted from Roldán et al. [29], and the proposed Range–Azimuth-Preserving Doppler Spectrum Encoder. All three configurations process the full RDA cube and provide 16 output feature channels to the subsequent U-Net 3+ prediction backbone. Their architectural definitions are given in Section 3.1.4. Table 4.5 summarizes their quantitative performance on the held-out test scene.

Table 4.5: Comparison of full-RDA Doppler encoder configurations on the RaDelft test scene. The variance is computed across the per-frame first-return MAE values.

Encoder configuration	$\mathrm{IoU}_{\mathrm{free}}$ ↑	$\mathrm{IoU}_{\mathrm{unfree}}$ ↑	mIoU ↑	MAE [m] ↓	Frame-MAE Var. m^2 ↓
Single-layer Doppler projection	0.75	0.80	0.77	6.35	9.44
Roldán-style Doppler encoder [29]	**0.79**	**0.82**	**0.80**	**5.37**	6.70
Proposed Doppler Spectrum Encoder	**0.79**	**0.82**	**0.80**	5.38	**6.52**

The single-layer Doppler projection obtains an mIoU of 0.77, a mean first-return MAE of 6.35 m, and a frame-MAE variance of 9.44 m^2. Both structured Doppler encoders provide substantially stronger results, increasing the mIoU to 0.80 and reducing the mean first-return MAE to approximately 5.4 m. This comparison indicates that a structured transformation of the Doppler spectrum is more effective than directly projecting all Doppler bins through a single 1×1 operation.

The single-layer configuration should, however, be interpreted as a reference configuration rather than as a strictly controlled encoder-only ablation, because it was trained using an earlier objective setting. The comparison therefore provides supporting evidence for structured Doppler processing, while the relative comparison between the Roldán-style and proposed encoders is more directly informative.

The Roldán-style encoder and the proposed Doppler Spectrum Encoder achieve almost identical aggregate predictive performance. The Roldán-style configuration obtains a marginally lower mean first-return MAE of 5.37 m, compared with 5.38 m for the proposed encoder, while both achieve an mIoU of 0.80. The proposed encoder provides a slightly lower frame-MAE variance of 6.52 m^2, compared with 6.70 m^2 for the Roldán-style alternative. The differences in predictive accuracy are therefore negligible at the reported precision.

The main distinction between the two structured encoders is however computational. The Roldán-style configuration requires 139.6 GMACs, whereas the proposed configuration requires 125.3 GMACs. This corresponds to a reduction of approximately 10.24% in computational workload while maintaining essentially the same predictive accuracy. The proposed encoder additionally restricts its three-dimensional convolutions to the Doppler dimension, preserving the range–azimuth geometry until the subsequent two-dimensional spatial backbone.

Overall, the radar-front-end experiments support the two design choices used in the final proposed method. First, retaining the complete Doppler-resolved measurement provides stronger performance than applying fixed mean or maximum Doppler compression before learning. Second, structured Doppler encoding substantially outperforms the simple projection reference, while the proposed Range–Azimuth-Preserving Doppler Spectrum Encoder achieves accuracy comparable to the Roldán-style alternative at lower computational cost. These results motivate the use of the full-RDA input and proposed Doppler encoder in the selected single-frame configuration.

4.2.3 Component Analysis on PolarNet

To further examine whether the gains of the proposed framework can be attributed to individual components in isolation, selected components are transferred to the RaDelft-adapted PolarNet reference. This experiment is intended as a cross-architecture component analysis rather than a strict ablation, since PolarNet differs from the proposed framework in both its spatial backbone and its binary cell-wise prediction formulation. The proposed Doppler encoder and the proposed learning objective are therefore evaluated individually and jointly within the PolarNet framework. In addition, the Roldán-style Doppler encoder is transferred to PolarNet to assess whether the effect of learned Doppler processing depends on the downstream architecture.

For the loss-transfer experiment, the proposed loss is adapted to PolarNet's binary free-space output. Since PolarNet does not explicitly predict a normalized first-return distribution over range and no-hit states, the CDF-related supervision cannot be transferred in exactly the same form as in the proposed structured head. The corresponding result should therefore be interpreted as an adapted transfer of the loss function rather than an identical loss-only replacement.

Table 4.6 summarizes the results.

Replacing PolarNet's fixed Doppler-compressed input with the proposed Doppler encoder produces only a small change in aggregate accuracy. The mIoU increases from 0.64 to 0.64, while the mean first-return MAE decreases slightly from 10.89 m to 10.83 m. The effect is more apparent in the frame-wise error variance, which decreases from 10.97 m^2 to 8.44 m^2. The free-space IoU also increases from 0.57 to 0.60, although this is partly offset by a reduction in unfree-space IoU. These results indicate that the richer Doppler representation changes the behaviour of PolarNet and improves prediction consistency, but does not by itself produce a substantial improvement in mean occupancy accuracy.

Table 4.6: Cross-architecture component analysis on the RaDelft-adapted PolarNet reference. The proposed Doppler encoder and the adapted proposed loss function are transferred individually and jointly to the PolarNet framework. The variance is computed across the per-frame first-return MAE values.

PolarNet configuration	IoU_{free} ↑	IoU_{unfree} ↑	mIoU ↑	MAE [m] ↓	Frame-MAE Var. m^2 ↓
Original PolarNet	0.57	**0.70**	0.64	10.89	10.97
+ Proposed Doppler Encoder	0.60	0.69	0.64	10.83	8.44
+ Adapted Proposed Loss	0.52	0.70	0.61	11.59	14.51
+ Proposed Encoder + Adapted Loss	0.58	**0.70**	0.64	10.78	9.07
+ Roldán-style Doppler Encoder	**0.61**	0.69	**0.65**	**10.55**	**7.94**

The adapted proposed loss transfers less favourably. When applied to the original PolarNet architecture, the mIoU decreases to 0.61, the mean first-return MAE increases to 11.59 m, and the frame-MAE variance rises to 14.51 m^2. The degradation is concentrated primarily in free-space prediction, while the unfree-space IoU remains close to that of the original PolarNet. This behaviour is consistent with the mismatch between the structured first-return formulation for which the proposed objective was designed and PolarNet's independent binary cell-wise output. In the proposed framework, the CDF term acts on a normalized distribution over mutually exclusive first-return and no-hit states, which guarantees a valid monotonic cumulative structure. PolarNet does not impose this representation, so the transferred objective acts only as an approximation to the original supervision mechanism.

Using both the proposed Doppler encoder and the adapted loss recovers much of the original PolarNet performance. The resulting mIoU returns to 0.64, while the mean first-return MAE improves slightly to 10.78 m and the frame-MAE variance decreases to 9.07 m^2. The partial recovery suggests that the effect of the transferred loss depends on the representation provided to the downstream network, but the combined modification still does not reproduce the gains observed in the complete proposed framework.

Among the transferred Doppler encoders, the Roldán-style encoder provides the strongest result within PolarNet. It increases the mIoU to 0.65, reduces the mean first-return MAE to 10.55 m, and yields the lowest frame-MAE variance of 7.94 m^2. This does not imply that the Roldán-style encoder is generally superior to the proposed encoder, since the two achieve nearly identical predictive accuracy within the proposed U-Net 3+ framework. Instead, the difference suggests that the effectiveness of a Doppler encoder depends on how its output features are exploited by the downstream spatial architecture.

Overall, the cross-architecture results show that the proposed components should not be interpreted as universally transferable plug-in improvements. Learned Doppler encoding can provide modest gains and improved consistency within PolarNet, but the full performance improvement of the proposed method cannot be reproduced

by transferring the encoder or learning objective alone. The results therefore support the view that the main performance gains arise from the joint integration of Doppler-resolved feature extraction, multi-scale spatial processing, structured first-return prediction, and task-aligned supervision in the proposed approach.

4.2.4 Controlled Extensions

Two controlled extensions are evaluated to examine whether the selected single-frame framework can be further improved through supervision reweighting or short-term temporal context. Radar-Aware Weighting modifies only the azimuth-wise contribution of the smoothed CDF-L1 term, while Causal Temporal Feature Fusion incorporates in a new architecture radar measurements from $t-2$, $t-1$, and t. Both extensions are evaluated relative to the same proposed single-frame baseline.

Table 4.7 summarizes the quantitative results.

Table 4.7: Comparison of the proposed single-frame baseline and the two controlled extensions on the RaDelft test scene. The variance is computed across the per-frame first-return MAE values.

Configuration	IoU_{free} ↑	$\text{IoU}_{\text{unfree}}$ ↑	mIoU ↑	MAE [m] ↓	Frame-MAE Var. m^2 ↓
Proposed baseline	0.79	0.82	0.80	5.38	6.52
Radar-Aware Weighting	0.78	0.82	0.80	5.37	7.30
Causal Temporal Feature Fusion	0.79	0.82	0.80	**5.35**	**5.98**

Radar-Aware Weighting

Radar-Aware Weighting introduces azimuth-wise reliability weights into the smoothed CDF-L1 term while leaving the network architecture, Lovasz term, and global loss coefficients unchanged. The weighting accounts for factors associated with supervision reliability, including abrupt first-return changes, far-range and no-hit columns, proximity to the unreliable LiDAR sector, and local radar evidence near the LiDAR-derived first return.

As shown in Table 4.7, the extension produces almost unchanged aggregate performance. The mIoU remains 0.80 at the reported precision, while the mean first-return MAE decreases only marginally from 5.38 m to 5.37 m. At the same time, the frame-MAE variance increases from 6.52 m^2 to 7.30 m^2.

The weighting strategy therefore does not provide a consistent improvement over the baseline. Although down-weighting potentially unreliable azimuth columns can slightly alter the learned boundary behaviour, the increased across-frame variance indicates that these changes do not translate into more stable performance over the complete test scene. This suggests that supervision reliability alone is not the dominant limitation of the selected single-frame framework.

Causal Temporal Feature Fusion

The temporal extension incorporates radar measurements from $t-2$, $t-1$, and t through feature-level fusion. The Doppler Spectrum Encoder and U-Net 3+ backbone are initialized from the selected single-frame checkpoint and remain frozen, while only the temporal-fusion module is optimized.

The occupancy-level results remain effectively unchanged relative to the single-frame baseline. The reported mIoU remains 0.80, with $\mathrm{IoU}_{\mathrm{free}} = 0.79$ and $\mathrm{IoU}_{\mathrm{unfree}} = 0.82$. However, the mean first-return MAE decreases from 5.38 m to 5.35 m, while the frame-MAE variance decreases from 6.52 m^2 to 5.98 m^2.

The unchanged occupancy metrics indicate that temporal fusion does not substantially modify the overall free–unfree segmentation structure. Its main effect is instead observed at the boundary level, where the slightly lower mean error and variance indicate a modest improvement in first-return localization and consistency. Among the two controlled extensions, temporal fusion therefore provides the more favourable aggregate result, although the improvement over the single-frame baseline remains rather limited.

Figure 4.7 compares the per-frame first-return MAE distributions of the three configurations to observe more details than just a mean value across frames as in the table.

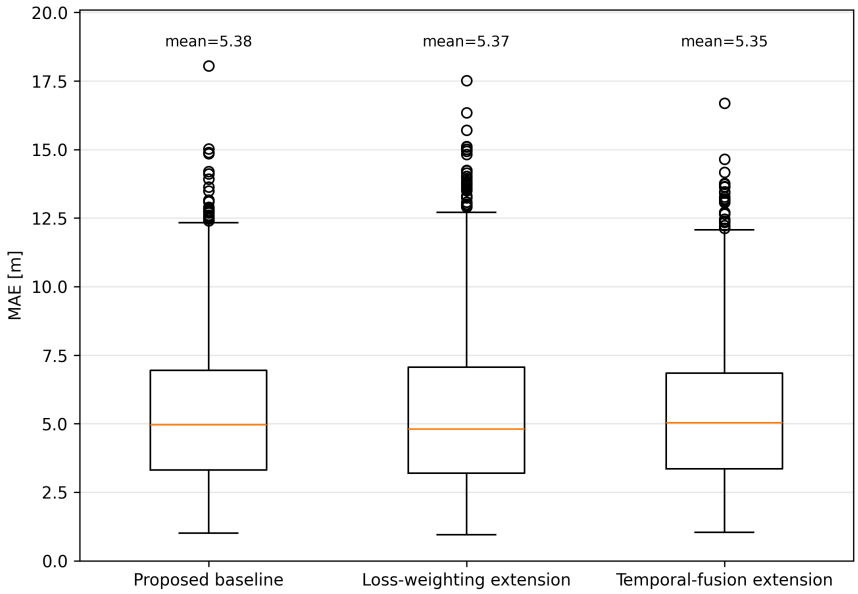

Figure 4.7: Per-frame first-return MAE distributions of the proposed baseline, Radar-Aware Weighting, and Causal Temporal Feature Fusion on the RaDelft test scene.

The distributions overlap substantially, which is consistent with the small differences in their mean errors. Causal Temporal Feature Fusion exhibits the lowest

mean MAE and the lowest across-frame variance, whereas Radar-Aware Weighting produces the largest variance. The substantial overlap nevertheless shows that neither extension changes the overall error distribution fundamentally; most of the frame-level behaviour remains shared with the single-frame baseline.

The limited aggregate differences are also apparent qualitatively. Figure 4.8 shows a representative frame with a relatively clear road structure and well-defined free-space boundary.

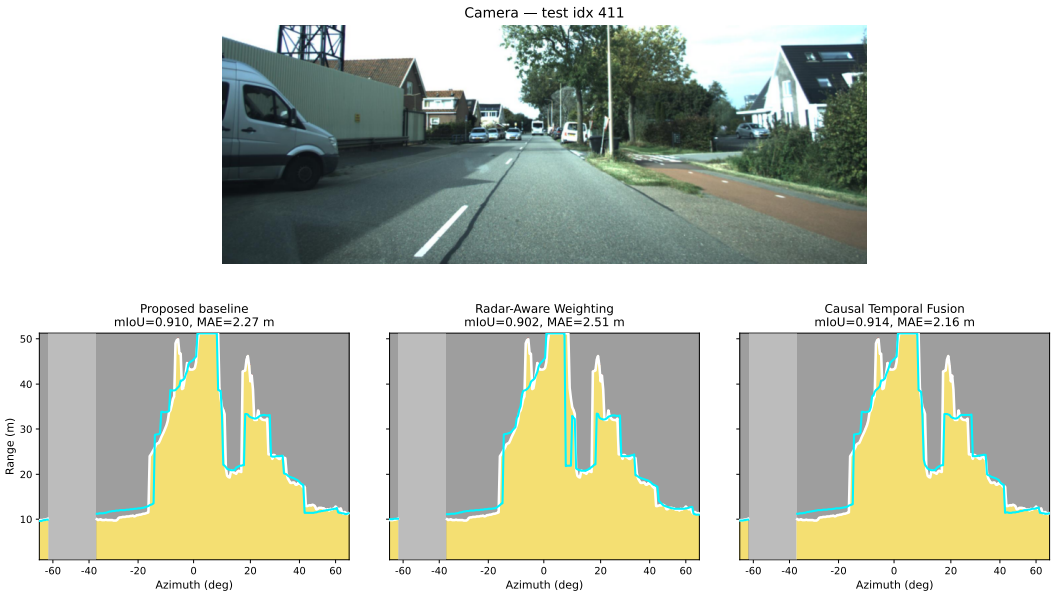

Figure 4.8: Qualitative comparison of the proposed baseline (left), Radar-Aware Weighting (middle), and Causal Temporal Feature Fusion (right) in a representative test frame. The camera image is included for visual context. In each polar panel, the white curve denotes the LiDAR-derived ground-truth first-return frontier, while the cyan curve denotes the predicted frontier.

The three predictions are visually similar and recover the dominant free-space structure over most of the evaluated field of view. Radar-Aware Weighting produces only small local changes relative to the baseline, while temporal fusion slightly modifies the predicted frontier in several regions. These differences are consistent with the quantitative results, which show only minor changes in aggregate occupancy and boundary performance.

A more challenging example is shown in Figure 4.9. In this scene, the road boundary is less regular and the first-return frontier contains stronger local variation.

Despite the increased scene complexity, all three configurations recover the dominant free-space contour and retain several localized first-return changes. For this individual frame, Radar-Aware Weighting achieves the closest agreement with the LiDAR-derived frontier, with an mIoU of 0.878 and a first-return MAE of 2.82 m.

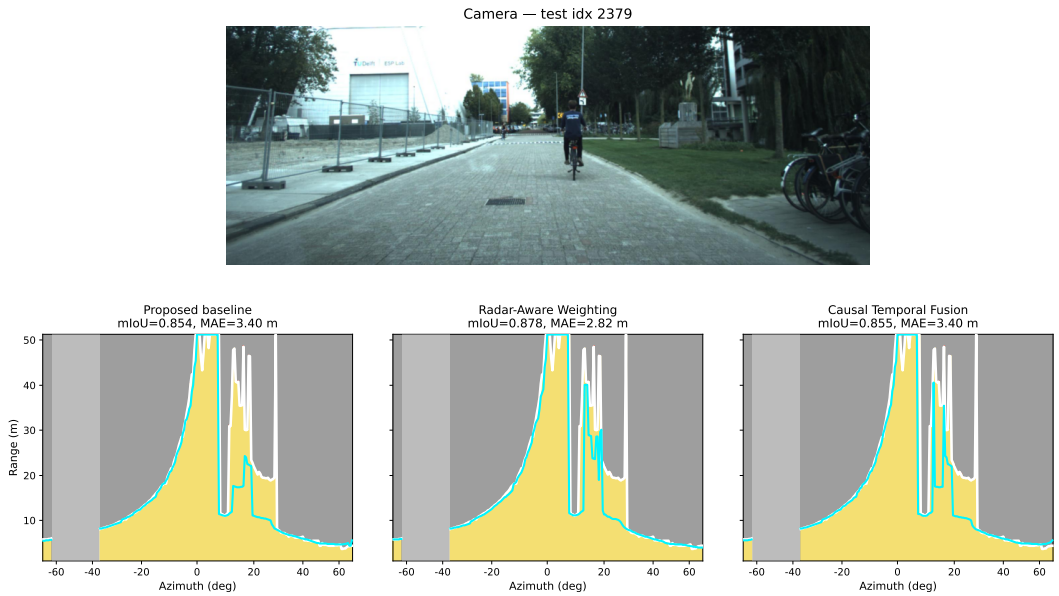

Figure 4.9: Qualitative comparison of the proposed baseline (left), Radar-Aware Weighting (middle), and Causal Temporal Feature Fusion (right) in a challenging urban scene. The camera image is included for visual context. In each polar range–azimuth panel, the white curve denotes the LiDAR-derived ground-truth first-return frontier, while the cyan curve denotes the predicted frontier.

The proposed baseline and temporal-fusion extension obtain first-return MAEs of 3.40 m, with mIoU values of 0.854 and 0.855, respectively.

This frame illustrates that the weighting strategy can improve particular local configurations even though it does not provide a consistent aggregate advantage. The effect should therefore be interpreted as frame-specific rather than as evidence that Radar-Aware Weighting systematically improves the baseline.

The similarities between the three configurations also extend to difficult failure cases. Figure 4.10 shows a representative frame in which all three methods produce inaccurate first-return estimates.

The LiDAR-derived frontier in this frame contains multiple narrow structures and abrupt range changes across neighbouring azimuth directions. In contrast, all three predicted frontiers are substantially smoother, and several near-range first returns are missed or merged with more distant structures. The proposed baseline obtains an mIoU of 0.607 and a first-return MAE of 12.29 m, while Radar-Aware Weighting and Causal Temporal Feature Fusion obtain MAEs of 12.51 m and 12.96 m, respectively.

The shared failure indicates that neither extension addresses the principal difficulty in this example. Reweighting the supervision cannot recover a boundary when the underlying radar evidence for a narrow structure is weak or spatially ambiguous.

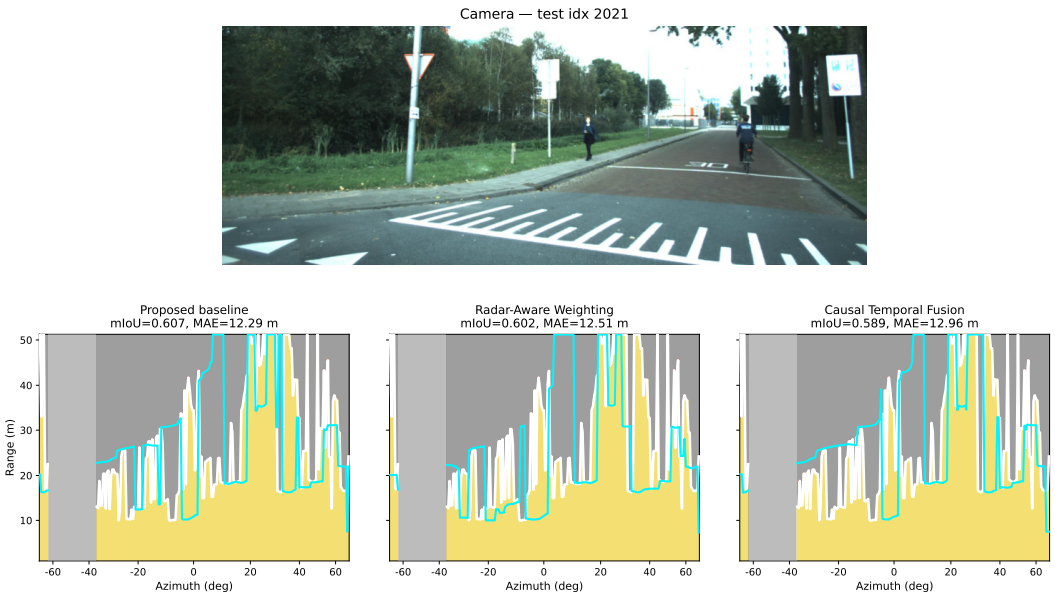

Figure 4.10: Representative failure case for the proposed baseline (left), Radar-Aware Weighting (middle), and Causal Temporal Feature Fusion (right). The camera image is included for visual context. In each polar range–azimuth panel, the white curve denotes the LiDAR-derived ground-truth first-return frontier, while the cyan curve denotes the predicted frontier.

Similarly, the no-warp temporal fusion module does not explicitly align historical radar features to the current coordinate frame, so small objects and rapidly changing local structures may not benefit from the additional temporal context.

More generally, the remaining errors are concentrated around narrow objects, rapid first-return transitions, and spatially ambiguous radar responses. Angular resolution, local smoothing in the convolutional prediction framework, and motion-related effects in the TDMA MIMO radar measurements may all contribute to these cases, although the qualitative results alone do not allow their individual effects to be isolated.

Overall, the controlled-extension experiments indicate that the selected single-frame framework already captures most of the performance achievable with the tested modifications. Radar-Aware Weighting changes local predictions but does not improve aggregate consistency, while Causal Temporal Feature Fusion provides only a modest reduction in boundary error and variance. These results motivate retaining the single-frame configuration as the principal proposed method and treating both extensions as exploratory refinements rather than replacements for the baseline.

4.2.5 Computational Efficiency

In addition to predictive accuracy, the computational requirements of the learning-based methods are evaluated in terms of parameter count, GMACs, GPU inference latency, throughput, and peak GPU memory usage. The measurements follow the protocol defined in Section 3.2.4: all models are evaluated with batch size one using FP32 inference on an NVIDIA A100 GPU, and only the neural-network forward pass is included. Radar preprocessing, data transfer, occupancy rendering, metric computation, and visualization are excluded.

Table 4.8 summarizes the results for the RaDelft-adapted PolarNet, the proposed single-frame method, and the Causal Temporal Feature Fusion extension. Radar-Aware Weighting is not reported separately because it modifies only the training objective and therefore has the same inference architecture and computational cost as the proposed single-frame model.

Table 4.8: Model-only computational efficiency of the learning-based methods under batch-size-one FP32 inference on an NVIDIA A100 GPU.

Method	Params [M]	GMACs	Latency [ms]	FPS	Peak memory [MB]
RaDelft-adapted PolarNet	1.2	16.4	10.13 ± 0.05	98.70	146.6
Proposed method	6.5	125.3	148.07 ± 0.23	6.75	796.9
Causal Temporal Feature Fusion	6.5	156.8	278.13 ± 0.49	3.60	2333.8

PolarNet is substantially lighter and faster than the proposed method. Its single-channel range–azimuth input and compact encoder–decoder architecture require only 1.2 million parameters and 16.4 GMACs, resulting in a mean forward-pass latency of 10.13 ms and a throughput of 98.70 FPS. By comparison, the proposed single-frame method requires 6.5 million parameters and 125.3 GMACs, with a latency of 148.07 ms and a throughput of 6.75 FPS.

This additional computational cost accompanies the substantial improvement in predictive performance reported in Section 4.1. PolarNet achieves an mIoU of 0.64 and a first-return MAE of 10.89 m, whereas the proposed method reaches an mIoU of 0.80 and reduces the MAE to 5.38 m. The comparison therefore reveals a clear accuracy–efficiency trade-off: PolarNet is considerably more computationally efficient, while the proposed framework provides substantially more accurate occupancy and first-return estimation.

The temporal-fusion extension increases the computational burden further. Although the trainable fusion module itself introduces only a small additional parameter overhead at the reported precision, inference requires processing three consecutive radar cubes rather than a single frame. Under the implemented no-cache inference procedure, the computational cost increases to 156.8 GMACs, the latency reaches 278.13 ms, and the peak GPU memory usage increases to 2333.8 MB. The resulting throughput is reduced to 3.60 FPS.

The additional temporal cost is not accompanied by a comparable increase in predictive accuracy. As shown in Section 4.2.4, temporal fusion reduces the mean first-return MAE only from 5.38 m to 5.35 m and the frame-MAE variance from 6.52 m^2 to 5.98 m^2, while the reported mIoU remains 0.80. The improvement in boundary localization is therefore rather modest relative to the substantial increase in latency and memory usage.

Overall, the proposed single-frame method provides the most favourable balance between predictive performance and computational cost among the evaluated full-RDA configurations. PolarNet remains preferable when computational efficiency is the primary requirement, but at the price of a considerable loss in occupancy and boundary accuracy. Causal Temporal Feature Fusion provides a small improvement in first-return consistency at substantially higher inference cost, supporting the use of the single-frame configuration as the principal proposed method.

4.3 Summary

This chapter evaluated the proposed LiDAR-supervised radar-based first-return occupancy mapping framework in terms of predictive accuracy, cross-scene performance, model design, controlled extensions, and computational cost. The proposed single-frame method outperformed both the traditional multi-frame ISM and the RaDelft-adapted PolarNet reference in occupancy agreement and first-return localization. Leave-one-scene-out evaluation showed reasonably consistent occupancy performance across the seven RaDelft scenes, with a macro-average mIoU of 0.78 ± 0.03 and a first-return MAE of 5.75 ± 1.07 m. Larger errors were mainly associated with long visible ranges, no-hit directions, abrupt boundary changes, and locally ambiguous radar responses.

The ablation studies supported the main design choices. Lovasz and CDF-based supervision outperformed generic regression losses, with the combined Lovasz and smoothed CDF-L1 loss providing the strongest balance between occupancy overlap and boundary accuracy. Full Doppler-resolved input outperformed Doppler mean and maximum aggregation, while structured Doppler encoding improved over a single-layer projection. The proposed Doppler Spectrum Encoder achieved accuracy comparable to the Roldán-style alternative at lower computational cost.

The controlled extensions provided only limited gains. Radar-Aware Weighting did not yield a consistent aggregate benefit, whereas Causal Temporal Feature Fusion slightly improved first-return accuracy and across-frame consistency at substantially higher inference cost. Narrow structures and rapidly varying boundaries remained challenging.

Overall, the proposed single-frame full-RDA framework provides a favourable balance between prediction accuracy, cross-scene robustness, and computational cost among the evaluated full-RDA methods.

Chapter 5

Conclusion and Future Work

5.1 Conclusion

This thesis investigated LiDAR-supervised radar-only first-return occupancy estimation from automotive radar measurements. The work was structured around three research questions concerning visibility-consistent prediction, Doppler-resolved radar encoding, and task-aligned supervision.

With respect to **RQ1**, the results show that radar occupancy estimation can be effectively formulated as an azimuth-wise first-return or no-hit prediction problem. The corresponding LiDAR-derived supervision and visibility-consistent rendering enforce the ordering of free space, the nearest occupied boundary, and the unknown region behind it directly through the output representation. Using this formulation, the selected single-frame model achieved an mIoU of 0.80 and a mean first-return MAE of 5.38 m, outperforming both the odometry-compensated inverse sensor model and the RaDelft-adapted PolarNet reference. This supports the contribution of a structured first-return formulation rather than independent cell-wise occupancy prediction.

Regarding **RQ2**, retaining the complete range–Doppler–azimuth measurement was more effective than fixed Doppler-mean or Doppler-maximum compression. The proposed Range–Azimuth-Preserving Doppler Spectrum Encoder preserved the spatial alignment required by the occupancy target while learning from the Doppler spectrum at multiple scales. Its predictive performance was comparable to the structured encoder adapted from the literature while requiring lower computational cost. These findings support the use of Doppler-resolved radar information before spatial prediction, rather than removing the Doppler dimension through fixed aggregation.

For **RQ3**, the combined Lovasz and smoothed CDF-L1 objective provided complementary supervision of occupancy overlap and first-return boundary localization. The controlled extensions, however, offered limited additional benefit. Radar-Aware Weighting did not produce a consistent improvement, while Causal Temporal Feature Fusion slightly reduced first-return error and its frame-to-frame variation but introduced substantially greater computational cost without improving occupancy

overlap. The selected single-frame formulation therefore provided the strongest overall trade-off among the evaluated configurations.

The principal contribution of this thesis is consequently not any individual use of LiDAR supervision, Doppler information, multi-scale convolution, or temporal fusion in isolation, but their integration into a radar-to-occupancy framework centred on structured first-return prediction. The results demonstrate that radar-aligned first-return supervision, Doppler-resolved encoding, and task-aligned learning together provide a viable basis for radar-only polar occupancy mapping, while also showing that remaining errors are dominated by difficult spatial structures and ambiguous radar responses rather than by a single isolated component of the framework.

5.2 Limitations

Several limitations should be considered when interpreting the results.

First, the LiDAR-derived occupancy maps are sensor-based reference labels rather than perfect representations of the environment. LiDAR and radar differ in visibility, angular resolution, material response, penetration, and occlusion behaviour. Consequently, an obstacle that is clearly visible to LiDAR may produce weak or ambiguous radar evidence, while radar may observe responses that are not represented by the LiDAR first-return target. The validity mask reduces the influence of the most unreliable supervision region, but it does not remove all cross-sensor discrepancies.

Second, a more demanding validation would evaluate the framework across independent automotive radar datasets and sensor configurations. Such an evaluation would require datasets providing Doppler-resolved radar measurements, preferably the full range–Doppler–azimuth representation or sufficiently low-level radar data to reproduce it, together with synchronized LiDAR measurements, reliable sensor calibration, and multiple independent driving scenes. These requirements are necessary to reproduce the proposed LiDAR-supervised first-return targets and to evaluate the same radar representation consistently. Cross-dataset and cross-sensor generalization therefore remains an important direction for future work.

Third, the first-return representation deliberately reduces each azimuth column to one nearest boundary or a no-hit outcome. This produces a physically consistent free–occupied–unknown structure, but it does not describe multiple surfaces along the same viewing direction, object semantics, object motion, or the full three-dimensional scene. Furthermore, the occupancy evaluation groups occupied and unknown cells into a common unfree class. The reported IoU values therefore evaluate the separation between free and unfree regions rather than independent recognition of all three occupancy states.

Fourth, the quality of the radar input remains dependent on the signal processing applied before network inference. The RaDelft radar uses a TDMA MIMO acquisition scheme, and residual motion-induced phase errors or angle–Doppler coupling may affect the available Doppler structure [41]. These effects were not isolated ex-

perimentally, and the current model operates on the resulting preprocessed radar cube without explicitly estimating its calibration uncertainty.

Fifth, the temporal-fusion extension combines historical features without explicit ego-motion compensation or learned spatial alignment. Range–azimuth features from $t-2$, $t-1$, and t may therefore correspond to different physical locations when the ego vehicle or surrounding objects move. In addition, the implemented inference procedure repeatedly processes all three frames without caching historical features. The reported temporal result consequently reflects a no-warp, no-cache implementation rather than an optimized streaming temporal model.

Finally, the proposed baseline remains computationally more expensive than the compact PolarNet reference. Although the additional cost is accompanied by substantially higher accuracy, the measured throughput indicates that further optimization would be required for higher-rate online deployment. The present work therefore establishes the prediction framework and its accuracy characteristics, but does not provide a fully optimized embedded implementation.

5.3 Future Work

Future work can extend the proposed framework in several directions.

A first direction is to improve the reliability and expressiveness of the supervision target. Instead of assigning one deterministic first-return label to every valid azimuth column, future models could explicitly represent uncertainty in the LiDAR-to-radar correspondence. Confidence values could account for LiDAR sampling density, local frontier discontinuities, radar evidence near the target boundary, and expected cross-sensor visibility differences. Synchronized camera information could additionally provide complementary semantic supervision, following prior work on camera-supervised radar perception [9]. Multi-frame LiDAR accumulation or probabilistic first-return targets could also provide more stable supervision for narrow or intermittently observed structures, provided that dynamic objects and ego motion are handled appropriately.

A second direction is motion-compensated temporal modelling. Historical radar features could be transformed into the current coordinate frame using ego-motion estimates before temporal fusion. Learned feature alignment, deformable sampling, optical-flow-like displacement estimation, or attention over spatial neighbourhoods could further address residual motion and moving objects. A streaming implementation could cache the encoded features of previous frames and reuse them for subsequent predictions, reducing the latency and memory overhead observed in the current no-cache implementation.

The temporal training strategy could also be expanded beyond optimizing only the fusion module. After an initial frozen-backbone stage, selected layers of the Doppler encoder or spatial backbone could be fine-tuned using a lower learning rate. This would allow the representation to adapt to temporal input while limiting the risk of degrading the single-frame representation learned by the baseline. Com-

parisons between frozen, partially fine-tuned, and fully trainable temporal models would help clarify whether the limited improvement observed in this thesis is primarily caused by the fusion mechanism, the absence of spatial alignment, or the restricted optimization strategy.

A third direction is the development of more efficient Doppler-resolved encoders. Factorized convolutions, grouped or depthwise operations, learned spectral pooling, low-rank decomposition, and sparse processing could reduce the computational cost of full-RDA inference. Knowledge distillation could transfer information from the full-RDA model to a smaller range–azimuth model, while quantization and structured pruning could support deployment on automotive computing platforms. Future efficiency studies should evaluate complete end-to-end latency, including radar preprocessing, memory transfer, occupancy rendering, and temporal caching, rather than only the neural-network forward pass.

Further work could also investigate richer radar representations. Complex-valued or phase-aware processing may preserve information lost when only magnitude-based inputs are used. The robustness of the current TDMA processing could also be improved through more accurate Doppler-ambiguity resolution and transmitter-dependent motion-phase compensation, as investigated in prior TDM-MIMO radar studies [1]. Improved array calibration could further reduce residual phase inconsistencies before Doppler encoding. The present first-return formulation could additionally be extended towards elevation-aware or three-dimensional occupancy prediction, although this would require denser and more reliable supervision as well as substantially greater computational resources.

The output formulation could also be extended to better represent ambiguous radar observations. Improving the calibration of the predicted first-return distribution, including the no-hit probability, could provide a more informative measure of prediction confidence. Multiple candidate boundaries could additionally be considered in cases where the radar evidence supports several plausible first-return locations. Dynamic-state estimation and semantic information could further enrich the representation while retaining the visibility-consistent relationship between free space, the nearest observable surface, and the region behind it.

Finally, broader experimental validation is required beyond the RaDelft dataset. The proposed framework should be evaluated on independent datasets collected with different radar configurations and sensor setups. Testing under rain, fog, night-time conditions, dense traffic, and different road geometries would provide a more complete assessment of robustness. Cross-dataset transfer and fine-tuning experiments would also help clarify which components of the learned representation are specific to the RaDelft sensor configuration and which generalize to other automotive radar systems.

Together, these directions would extend the present work from a single-frame LiDAR-supervised first-return mapping framework towards a more uncertainty-aware, temporally aligned, computationally efficient, and broadly validated radar occupancy perception system.

Bibliography

[1] Jonathan Bechter, Fabian Roos, and Christian Waldschmidt. Compensation of motion-induced phase errors in tdm mimo radars. *IEEE Microwave and Wireless Components Letters*, 27(12):1164–1166, 2017.

[2] Maxim Berman, Amal Rannen Triki, and Matthew B Blaschko. The lovász-softmax loss: A tractable surrogate for the optimization of the intersection-over-union measure in neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4413–4421, 2018.

[3] Mario Bijelic, Tobias Gruber, and Werner Ritter. Seeing through fog without seeing fog: Deep multimodal sensor fusion in unseen adverse weather. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11682–11692, 2020.

[4] Christophe Coué, Cédric Pradalier, Christian Laugier, Thierry Fraichard, and Pierre Bessière. Bayesian occupancy filtering for multitarget tracking: an automotive application. *The International Journal of Robotics Research*, 25(1): 19–30, 2006.

[5] Johan Degerman, Thomas Pernstål, and Klas Alenljung. 3d occupancy grid mapping using statistical radar models. In *2016 IEEE Intelligent Vehicles Symposium (IV)*, pages 902–908. IEEE, 2016.

[6] Fangqiang Ding, Xiangyu Wen, Yunzhou Zhu, Yiming Li, and Chris Xiaoxuan Lu. Radarocc: Robust 3d occupancy prediction with 4d imaging radar. *Advances in Neural Information Processing Systems*, 37:101589–101617, 2024.

[7] Mariella Dreissig, Dominik Scheuble, Florian Piewak, and Joschka Boedecker. Survey on lidar perception in adverse weather conditions. *arXiv preprint arXiv:2304.06312*, 2023.

[8] Alberto Elfes. Using occupancy grids for mobile robot perception and navigation. *Computer*, 22(6):46–57, 1989. doi: 10.1109/2.30720.

[9] Christopher Grimm, Tai Fei, Ernst Warsitz, Ridha Farhoud, Tobias Breddermann, and Reinhold Haeb-Umbach. Warping of radar data into camera image for cross-modal supervision in automotive applications. *IEEE Transactions on Vehicular Technology*, 71(9):9435–9449, 2022.

[10] Stefan Haag, Bharanidhar Duraisamy, Daniel Pfrommer, Wolfgang Koch, Martin Fritzsche, and Jurgen Dickmann. Unify: Multi-belief bayesian grid framework based on automotive radar. *arXiv preprint arXiv:2104.11979*, 2021.

[11] Kejia Hu, Mohammed Alsakabi, John M. Dolan, and Ozan K. Tonguz. Reproducing and extending radelft 4d radar with camera-assisted labels, 2025. URL `https://arxiv.org/abs/2512.02394`.

[12] Huimin Huang, Lanfen Lin, Ruofeng Tong, Hongjie Hu, Qiaowei Zhang, Yutaro Iwamoto, Xianhua Han, Yen-Wei Chen, and Jian Wu. Unet 3+: A full-scale connected unet for medical image segmentation. In *ICASSP 2020-2020 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 1055–1059. Ieee, 2020.

[13] Peter J Huber. Robust estimation of a location parameter. In *Breakthroughs in statistics: Methodology and distribution*, pages 492–518. Springer, 1992.

[14] Texas Instruments Inc. Design guide: Tidep-01012—imaging radar using cascaded mmwave sensor reference design (rev. a), 2019. URL `https://www.ti.com/lit/ug/tiduen5a/tiduen5a.pdf`.

[15] Yi Jin, Marcel Hoffmann, Anastasios Deligiannis, Juan-Carlos Fuentes-Michel, and Martin Vossiek. Semantic segmentation-based occupancy grid map learning with automotive radar raw data. *IEEE Transactions on Intelligent Vehicles*, 9 (1):216–230, 2023.

[16] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[17] Pou-Chun Kung, Chieh-Chih Wang, and Wen-Chieh Lin. Radar occupancy prediction with lidar supervision while preserving long-range sensing and penetrating capabilities. *IEEE Robotics and Automation Letters*, 7(2):2637–2643, 2022.

[18] Iro Laina, Christian Rupprecht, Vasileios Belagiannis, Federico Tombari, and Nassir Navab. Deeper depth prediction with fully convolutional residual networks. In *2016 Fourth international conference on 3D vision (3DV)*, pages 239–248. IEEE, 2016.

[19] Seungjae Lee, Hyungtae Lim, and Hyun Myung. Patchwork++: Fast and robust ground segmentation solving partial under-segmentation using 3d point cloud. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 13276–13283. IEEE, 2022.

[20] Bence Major, Daniel Fontijne, Amin Ansari, Ravi Teja Sukhavasi, Radhika Gowaikar, Michael Hamilton, Sean Lee, Slawomir Grzechnik, and Sundar Subramanian. Vehicle detection with automotive radar using deep learning on range-azimuth-doppler tensors. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pages 924–932. IEEE, 2019.

[21] Hans Moravec and Alberto Elfes. High resolution maps from wide angle sonar. In *Proceedings. 1985 IEEE international conference on robotics and automation*, volume 2, pages 116–121. IEEE, 1985.

[22] Farzan Erlik Nowruzi, Dhanvin Kolhatkar, Prince Kapoor, Elnaz Jahani Heravi, Fahed Al Hassanat, Robert Laganiere, Julien Rebut, and Waqas Malik. Polarnet: Accelerated deep open space segmentation using automotive radar in polar domain. *arXiv preprint arXiv:2103.03387*, 2021.

[23] Çağan Önen, Ashish Pandharipande, Geethu Joseph, and Nitin Jonathan Myers. Occupancy grid mapping for automotive driving exploiting clustered sparsity. *IEEE Sensors Journal*, 24(7):9240–9250, 2024. doi: 10.1109/JSEN.2023.3342463.

[24] Arthur Ouaknine, Alasdair Newson, Patrick Pérez, Florence Tupin, and Julien Rebut. Multi-view radar semantic segmentation. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 15651–15660. IEEE, 2021.

[25] Andras Palffy, Jiaao Dong, Julian FP Kooij, and Dariu M Gavrila. Cnn based road user detection using the 3d radar cube. *IEEE Robotics and Automation Letters*, 5(2):1263–1270, 2020.

[26] Hwanhee Park, Yonghee Lee, Seonmin Cho, Seongwook Lee, Sangho Nam, Seungpyo Ha, and Byoungkwon Park. Deep learning-based occupancy grid mapping for free-space detection in automotive radar systems. In *2025 IEEE Radar Conference (RadarConf25)*, pages 1236–1241. IEEE, 2025.

[27] Jakub Porębski. Customizable inverse sensor model for bayesian and dempster-shafer occupancy grid frameworks. In *Advanced, Contemporary Control: Proceedings of KKA 2020—The 20th Polish Control Conference, Łódź, Poland, 2020*, pages 1225–1236. Springer, 2020.

[28] Julien Rebut, Arthur Ouaknine, Waqas Malik, and Patrick Pérez. Raw high-definition radar for multi-task learning. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17000–17009. IEEE, 2022.

[29] Ignacio Roldan, Andras Palffy, Julian FP Kooij, Dariu M Gavrila, Francesco Fioranelli, and Alexander Yarovoy. A deep automotive radar detector using the radelft dataset. *IEEE Transactions on Radar Systems*, 2:1062–1075, 2024.

[30] Christian M Schmid, Reinhard Feger, Clemens Pfeffer, and Andreas Stelzer. Motion compensation and efficient array design for tdma fmcw mimo radar systems. In *2012 6th European Conference on Antennas and Propagation (EUCAP)*, pages 1746–1750. IEEE, 2012.

[31] Liat Sless, Bat El Shlomo, Gilad Cohen, and Shaul Oron. Road scene understanding by occupancy grid learning from sparse radar clusters using semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, 2019.

[32] Michael Slutsky and Daniel Dobkin. Dual inverse sensor model for radar occupancy grids. In *2019 IEEE Intelligent Vehicles Symposium (IV)*, pages 1760–1767. IEEE, 2019.

[33] Arvind Srivastav and Soumyajit Mandal. Radars for autonomous driving: A review of deep learning methods and challenges. *arXiv preprint arXiv:2306.09304*, 2023.

[34] Botao Sun, Ignacio Roldan, and Francesco Fioranelli. Automatic labelling & semantic segmentation with 4d radar tensors. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2025.

[35] SS Vallender. Calculation of the wasserstein distance between probability distributions on the line. *Theory of Probability & Its Applications*, 18(4):784–786, 1974.

[36] Patrik Viktor and Gabor Kiss. Sensors in self-driving vehicles: A detailed literature review and new trends. *Sensors (Basel, Switzerland)*, 26(7):2153, 2026.

[37] Zihang Wei, Rujiao Yan, and Matthias Schreier. Deep radar inverse sensor models for dynamic occupancy grid maps. *arXiv preprint arXiv:2305.12409*, 2023.

[38] Klaudius Werber, Matthias Rapp, Jens Klappstein, Markus Hahn, Jürgen Dickmann, Klaus Dietmayer, and Christian Waldschmidt. Automotive radar gridmap representations. In *2015 IEEE MTT-S International Conference on Microwaves for Intelligent Mobility (ICMIM)*, pages 1–4. IEEE, 2015.

[39] Rob Weston, Sarah Cen, Paul Newman, and Ingmar Posner. Probably unknown: Deep inverse sensor modelling radar. In *2019 international conference on robotics and automation (ICRA)*, pages 5446–5452. IEEE, 2019.

[40] Tianwei Yin, Xingyi Zhou, and Philipp Krahenbuhl. Center-based 3d object detection and tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11784–11793, 2021.

[41] Sen Yuan, Tworit Dash, Ignacio Roldan, Francesco Fioranelli, and Alexander Yarovoy. A doppler de-aliasing technique for fmcw-tdm mimo automotive radar. *IEEE Transactions on Instrumentation and Measurement*, 2025.

[42] Sen Yuan, Francesco Fioranelli, and Alexander Yarovoy. High-resolution imaging algorithms for automotive radar: Challenges in real driving scenarios. *IEEE Aerospace and Electronic Systems Magazine*, 40(7):30–43, 2025. doi: 10.1109/MAES.2025.3550301.

[43] Sen Yuan, Changheng Li, and Francesco Fioranelli. Robust joint doppler disambiguation and phase-compensated doa estimation for fmcw-tdm mimo radar. *IEEE Transactions on Radar Systems*, 2026.

[44] Ekim Yurtsever, Jacob Lambert, Alexander Carballo, and Kazuya Takeda. A survey of autonomous driving: Common practices and emerging technologies. *IEEE Access*, 8:58443–58469, 2020. doi: 10.1109/ACCESS.2020.2983149.

[45] Ao Zhang, Farzan Erlik Nowruzi, and Robert Laganiere. Raddet: Range-azimuth-doppler based radar object detection for dynamic road users. In *2021 18th Conference on Robots and Vision (CRV)*, pages 95–102. IEEE, 2021.