Learning Structurally Consistent Representations for Multi-View Radar Semantic Segmentation

Ali Zia[†*1,2], Muhammad Umer Ramzan[*3], Abdelwahed Khamis[4], Usman Ali[3], and Abdul Rehman[3]

[1] School of Computing, Engineering and Mathematical Sciences, La Trobe University, Melbourne, Australia
[2] La Trobe Institute for Sustainable Agriculture & Food (LISAF), Australia
[3] School of Engineering and Applied Sciences, GIFT University, Pakistan
[4] Commonwealth Scientific and Industrial Research Organisation (CSIRO), Australia
A.Zia@latrobe.edu.au,
{umer.ramzan,usmanali,abdulrehman.naseer}@gift.edu.pk,
abdelwahed.khamis@data61.csiro.au

Abstract. Radar sensors provide reliable perception under adverse weather and lighting conditions, but their sparse, noisy, and weakly semantic measurements make dense semantic segmentation challenging. Most existing radar segmentation methods rely on grid-based encodings and pairwise interactions, which struggle to capture the higher-order relational structure formed by multiple radar returns from the same physical object. We introduce a unified higher-order structural alignment framework for multi-view radar segmentation. The proposed method refines radar feature representations using learnable hypergraphs to capture higher-order dependencies among spatially related responses. To ensure consistency across heterogeneous radar projections, we further align view-specific features using Unbalanced Optimal Transport (UOT), enabling correspondence-free alignment under varying measurement densities and partial observations. An adaptive attention mechanism then fuses complementary radar views while emphasising structurally informative responses under sparsity and noise. The resulting architecture learns structurally consistent representations across Range Angle (RA), Range Doppler (RD), and Angle Doppler (AD) views and is trained using supervised segmentation together with cross-view consistency regularisation. Experiments on the CARRADA and RADIal benchmarks demonstrate consistent improvements over strong radar-specific baselines, achieving **63.8% mIoU on CARRADA and 83.4% mIoU on RADIal**, improving the previous best methods by **+1.7 and +2.3 mIoU**, respectively. These results highlight the importance of higher-order relational modelling for robust radar perception.

Keywords: Representation Learning · Radar Semantic Segmentation · Multi-view Learning · Hypergraph Learning · Optimal Transport · Structured Representation Learning

* Equal Contribution. † Corresponding Author.

1 Introduction

Reliable scene understanding is a fundamental requirement for autonomous driving, particularly under conditions where perception quality directly affects safety-critical decisions. Among commonly used sensing modalities, radar is especially attractive because it remains operational under adverse weather, low illumination, glare, fog, and other degradations that can severely affect cameras and LiDAR [35]. This robustness makes radar an important component of modern perception stacks. However, despite its favourable sensing characteristics, semantic segmentation from radar remains substantially more difficult than in image or point-cloud domains. Radar measurements are inherently sparse, noisy, low-resolution, and weakly semantic, which limits the discriminative content available for dense scene parsing [6, 16]. In practice, radar evidence rarely appears as isolated responses; instead, multiple radar returns from the same object form groups of spatially and spectrally related measurements. Conventional convolutional or pairwise models struggle to capture this structure. For this reason, learning representations that are both structurally reliable and semantically expressive remains a central challenge in radar perception.

A large body of prior work on radar segmentation relies on grid-based encodings or convolutional architectures operating over range-angle, range-Doppler, or related radar tensors [18, 28]. While these formulations enable the use of mature dense prediction pipelines, they often inherit an implicit assumption of locally smooth and densely informative observations. This assumption is poorly matched to radar sensing, where returns are irregular, cluttered, and often fragmented across multiple cells. Moreover, radar reflections originating from a single physical object may be distributed across several neighbouring bins due to object extent, multipath effects, motion, and sensor viewpoint. Under such conditions, local convolutions and pairwise interactions are often insufficient to capture the higher-order dependencies needed for reliable semantic reasoning. Consequently, models may overfit to local artefacts, struggle under severe sparsity, and produce unstable segmentations in challenging scenes.

Recent efforts have sought to improve radar understanding through graph-based learning, self-supervision, and contrastive objectives, aiming to better exploit relational structure and reduce reliance on dense annotations [9, 11, 30]. Despite recent advances in radar perception, two limitations remain. First, most methods rely on convolutional or pairwise interactions that cannot explicitly represent groups of radar responses originating from the same object. Second, multi-view radar observations often exhibit unequal measurement densities and partial overlap, making conventional alignment strategies brittle. Addressing both challenges requires representations that capture higher-order structure while remaining robust to incomplete cross-view correspondence.

To address these limitations, we propose a structurally consistent representation learning framework for radar perception that combines hypergraph reasoning with unbalanced optimal transport alignment. Radar returns from the same physical object often appear as groups of spatially and spectrally related

responses distributed across multiple cells, a structure that conventional convolutional or pairwise models fail to capture.

We therefore represent radar features using *hypergraphs*, where a single hyperedge connects multiple related measurements or embeddings. This formulation enables higher-order relational modelling and promotes region-level coherence, allowing the network to capture object-level structure and remain robust to sparsity, clutter, and ambiguous radar observations.

While hypergraph reasoning captures intra-view structure, robust segmentation also requires consistent representations across radar views. In practice, different projections often exhibit varying measurement densities and partial observations due to occlusions and view-dependent sparsity. To handle these inconsistencies, we employ *Unbalanced Optimal Transport* (UOT) [27] as a correspondence-free alignment mechanism for multi-view radar features. Unlike balanced transport, UOT allows mass variation across distributions, making it well suited for radar data where the number and strength of returns differ between views. Applied to structurally refined features, UOT encourages consistent cross-view organisation while remaining tolerant to missing or spurious measurements. To further improve feature aggregation under sparse sensing conditions, we incorporate an adaptive attention mechanism that dynamically emphasises informative radar responses while suppressing noise. Building on these ideas, we introduce **HyperRadar**, an end-to-end framework for multi-view radar semantic segmentation that integrates higher-order hypergraph reasoning, correspondence-free UOT alignment, and adaptive attention-based refinement. By jointly enforcing intra-view structural coherence and inter-view distributional alignment, HyperRadar produces reliable semantic predictions under sparse and noisy radar observations. This paper makes the following contributions:

- We introduce a **structurally grounded radar representation framework** that models groups of radar returns using **hypergraphs**. This formulation enables explicit higher-order relational reasoning and captures object-level structure that cannot be represented through local convolutions or pairwise interactions.
- We propose a **correspondence-free multi-view alignment mechanism based on Unbalanced Optimal Transport (UOT)**, which aligns radar feature distributions across projections with varying measurement densities and partial observations, improving robustness to realistic sensing inconsistencies.
- We develop a **hypergraph-aware feature refinement strategy with adaptive attention**, which enforces coherent region-level representations and selectively emphasises informative radar responses for improved semantic segmentation under sparse and noisy sensing conditions.

Extensive experiments on the CARRADA [21] and RADIal [24] benchmarks demonstrate that HyperRadar consistently outperforms strong radar-specific baselines, achieving **63.8% mIoU on CARRADA** and **83.4% mIoU on RADIal**. This corresponds to improvements of $+$**1.7** and $+$**2.3 mIoU** over the

previous best methods, respectively, highlighting the effectiveness of higher-order relational modelling for radar perception.

2 Related Work

Low-cost frequency-modulated continuous-wave (FMCW) radars have been increasingly adopted in machine learning and pattern recognition tasks, including human activity recognition and hand gesture analysis [38, 40, 41]. In autonomous driving, however, LiDAR has historically been preferred due to its high-resolution geometric measurements, and several studies have therefore explored radar-LiDAR fusion within point-cloud-oriented pipelines [1, 7]. Despite these efforts, radar measurements differ fundamentally from LiDAR returns in sensing physics and signal formation. Their low spatial resolution, high sparsity, speckle-like noise, and clutter significantly limit the effectiveness of point-cloud abstractions and architectures designed for denser geometric observations [22, 26].

To address these limitations, recent work has shifted toward image-like radar representations derived from signal processing in the Range, Azimuth, and Doppler domains. While some datasets provide radar data in point-cloud format [2, 26], most contemporary automotive radar benchmarks provide either single-view projections such as Range-Azimuth (RA) or Range-Doppler (RD) maps [24, 29], raw radar signals [24], or full Range-Azimuth-Doppler (RAD) tensors [20, 36]. RAD tensors preserve richer spatial and kinematic structure, but their three-dimensional nature substantially increases memory and computational cost, especially for temporal modelling. Consequently, several methods operate on sliced or projected RAD representations to balance representational richness and efficiency [10, 20].

With the emergence of annotated automotive radar datasets such as CARRADA and RADIal [21,24], a growing number of methods have been proposed for radar semantic segmentation and object detection. Early baselines adapt conventional dense-prediction architectures, including UNet [25] and DeepLabv3+ [3]. Although effective for natural images, these models are not explicitly designed for the sparsity, noise, and clutter characteristic of radar imagery. TMVANet [20] addresses this limitation using a multi-view convolutional architecture with separate encoding, latent interaction, and decoding stages, yielding strong results on both RA and RD segmentation. RAMP-CNN [10], originally introduced for 3D RAD processing, has also influenced radar perception pipelines, though its reliance on volumetric convolutions limits scalability.

More recent methods have explored attention-based and feature-selective mechanisms to better capture long-range dependencies and salient radar responses. T-RODNet [12] employs Swin Transformers [17] for RA-based object detection, demonstrating the value of non-local feature aggregation, but it does not jointly model multi-view segmentation. PeakConv [39] introduces peak-aware convolution to focus on strong radar reflections and improves over TMVANet, albeit with higher parameter count and computational cost. Related sparse at-

tention mechanisms, such as ReLA [37], neighbourhood point attention [32], and SCAN [31], further show that selective feature activation can improve efficiency. However, these methods either target point-cloud data or operate primarily through local/pairwise interactions, and thus do not explicitly model the higher-order relational structure induced by multiple radar returns from the same physical object.

Hypergraphs provide a principled generalisation of graphs in which a single hyperedge can connect more than two vertices, making them well-suited to modelling group-wise dependencies and higher-order structure. Early work on spectral learning with hypergraphs showed their usefulness for clustering, classification, and embedding in settings where pairwise graphs are insufficient [42]. More recently, hypergraph neural models such as HGNN [8] and HyperGCN [33] have demonstrated that hypergraph-based message passing can effectively capture complex multi-way relations while remaining compatible with deep representation learning. These ideas are particularly relevant for radar perception, where meaningful evidence is often distributed across sets of neighbouring bins or returns associated with the same object, extended surface, or motion pattern. Nevertheless, prior radar segmentation methods predominantly rely on convolutional operators, standard attention, or pairwise interactions, and do not explicitly exploit hypergraph structure for semantic segmentation. Our work builds on this line of research by introducing hypergraph-based reasoning directly into multi-view radar representation learning, enabling region-level structural consistency beyond local or pairwise aggregation.

Optimal Transport (OT) has emerged as a powerful tool for comparing and aligning distributions in a geometrically meaningful manner, with broad impact across machine learning and computer vision [23]. Entropically regularised OT and Sinkhorn-based solvers [5] have made OT practical at modern learning scales, while applications such as domain adaptation have shown its effectiveness for correspondence-free alignment between heterogeneous feature distributions [4]. However, classical OT assumes balanced mass preservation, which can be restrictive when observations are incomplete, corrupted, or unevenly distributed. Unbalanced Optimal Transport (UOT) relaxes this assumption by allowing discrepancies in total mass, thereby improving robustness to outliers, missing observations, and partial overlap between distributions [27]. This property is especially desirable for multi-view radar, where different views often contain unequal numbers of valid returns due to occlusion, sparsity, clutter, and viewpoint-dependent visibility. In contrast, our method leverages UOT to align multi-view radar feature distributions without requiring explicit correspondences, a design choice we detail and validate in Sec. 3.3.

In summary, prior radar segmentation methods have improved feature extraction through multi-view convolutions, transformers, and sparse attention, but they remain limited in their ability to model higher-order radar structure and to align heterogeneous multi-view observations under unequal measurement densities. Our approach addresses these two gaps jointly by combining hypergraph-based structural reasoning with unbalanced optimal transport, yielding a repre-

sentation that is better matched to the physical and statistical characteristics of automotive radar data.

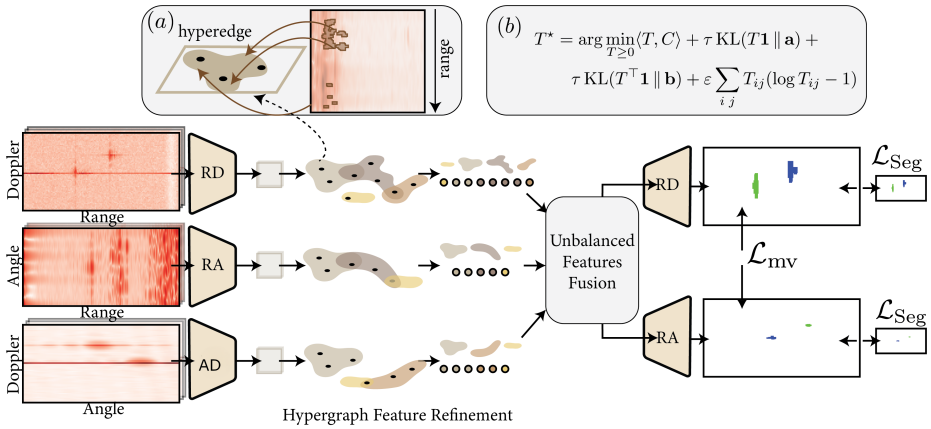

Fig. 1: Architecture of the proposed HyperRadar framework. Three radar views (RD, RA, AD) are processed by independent encoders and refined using view-specific learnable hypergraphs. (a) Hypergraph refinement: for an extended object like a car with distant reflections spanning multiple range cells, the hypergraph captures these non-local dependencies and groups them into a single hyperedge. (b) Unbalanced Optimal Transport (UOT) aligns the refined features across views. The fused tokens are refined via adaptive attention and subsequently decoded into dense segmentation masks. The network is trained end-to-end using joint segmentation and cross-view consistency objectives.

3 Methodology

3.1 Overview

Radar semantic segmentation is challenging due to sparse, noisy, and weakly semantic measurements, as well as view-dependent missing returns. While local convolutions and pairwise fusion capture short-range patterns, they are limited in modelling the higher-order structure induced by multiple returns from the same physical object. We propose **HyperRadar**, a compact multi-view framework that (i) extracts view-specific features from three RAD projections, (ii) performs *per-view* learnable hypergraph refinement to capture higher-order dependencies, (iii) aligns and fuses refined features across views via *unbalanced optimal transport* (UOT), and (iv) decodes RD and RA semantic masks with cross-view consistency regularisation. The overall pipeline is illustrated in Figure 1.

Given a radar cube at time t_0, we form three 2D projections from the 3D Range–Angle–Doppler (RAD) representation: Range–Doppler (RD), Range–Angle (RA), and Angle–Doppler (AD). For each view $v \in \{\mathrm{RD}, \mathrm{RA}, \mathrm{AD}\}$, the input is $x^{(v)} \in \mathbb{R}^{T \times H_v \times W_v}$, where T is the temporal window and H_v, W_v are the spatial dimensions. A view-specific encoder maps each input to a latent representation:

$$F^{(v)} = E_v(x^{(v)}) \in \mathbb{R}^{C \times H_d^{(v)} \times W_d^{(v)}}, \tag{1}$$

where C is the latent channel dimension. Flattening $F^{(v)}$ yields tokens $X^{(v)} \in \mathbb{R}^{N_v \times C}$ with $N_v = H_d^{(v)} W_d^{(v)}$, where each row corresponds to one latent location.

3.2 Learnable Hypergraph-Based Feature Refinement

For each view v, we refine latent tokens using a learnable hypergraph to capture higher-order interactions among spatially and spectrally related radar responses. We define a hypergraph $\mathcal{G}^{(v)} = (\mathcal{V}, \mathcal{E}^{(v)})$ over the N_v tokens, where $\mathcal{V} = \{1, \ldots, N_v\}$ and $\mathcal{E}^{(v)} = \{e_1, \ldots, e_M\}$ contains M learnable hyperedges. Instead of fixed neighbourhoods, node-to-hyperedge associations are learned via a soft incidence matrix

$$H^{(v)} = \mathrm{softmax}(X^{(v)} W_h) \in \mathbb{R}^{N_v \times M},$$

where $W_h \in \mathbb{R}^{C \times M}$ is trainable and the softmax is applied along the hyperedge dimension so each token distributes its membership across hyperedges. Hyperedge embeddings are computed by weighted aggregation $Z^{(v)} = (H^{(v)})^\top X^{(v)}$, transformed by a lightweight two-layer mapping $\phi(\cdot)$, and projected back to tokens through the same incidence structure. The refinement is expressed as

$$X_{\mathrm{hg}}^{(v)} = X^{(v)} + \gamma\, H^{(v)} \phi\Big((H^{(v)})^\top X^{(v)}\Big), \tag{2}$$

where γ is a learnable scalar. This residual hypergraph update allows tokens participating in the same higher-order structures to exchange contextual information, improving robustness to clutter, sparsity, and fragmented returns.

3.3 Unbalanced Optimal Transport for Multi-View Feature Fusion

Although hypergraph refinement improves within-view structural coherence, features remain misaligned across views due to unequal measurement densities and partial observations. We therefore align and fuse hypergraph-refined features across views using unbalanced optimal transport (UOT), which relaxes strict mass preservation and is more suitable for radar sparsity.

For a pair of views (p, q), let $X_{\mathrm{hg}}^{(p)} \in \mathbb{R}^{N_p \times C}$ and $X_{\mathrm{hg}}^{(q)} \in \mathbb{R}^{N_q \times C}$ denote refined tokens. We define the transport cost matrix $C^{(p,q)} \in \mathbb{R}^{N_p \times N_q}$ by

$$C_{ij}^{(p,q)} = \|x_i^{(p)} - x_j^{(q)}\|_2^2,$$

where $x_i^{(p)}$ and $x_j^{(q)}$ are token embeddings from views p and q. With reference masses $\mathbf{a} = \frac{1}{N_p}\mathbf{1} \in \mathbb{R}^{N_p}$ and $\mathbf{b} = \frac{1}{N_q}\mathbf{1} \in \mathbb{R}^{N_q}$, we solve the entropically regularised UOT problem

$$T^{(p,q)} = \arg\min_{T \geq 0} \langle T, C^{(p,q)} \rangle + \tau\, \mathrm{KL}(T\mathbf{1} \,\|\, \mathbf{a}) + \tau\, \mathrm{KL}(T^\top \mathbf{1} \,\|\, \mathbf{b}) + \varepsilon \sum_{i,j} T_{ij}(\log T_{ij} - 1), \tag{3}$$

where τ controls marginal relaxation, ε is the entropic regularisation weight, and $\mathrm{KL}(\cdot\|\cdot)$ denotes Kullback–Leibler divergence. We compute $T^{(p,q)}$ using a differentiable generalised Sinkhorn solver. The mass imbalance is inherent to radar physics: For a radar cube $\mathcal{S}[r, d, l]$ with range r, Doppler d, and antenna element l, the RD view integrates $\sum_l |\mathcal{S}[r, d, l]|^2$ over the full antenna array while RA integrates over a narrower aperture, causing a car's $\sim$3-bin RD peak to smear into 15+ RA bins and violating balanced-OT mass conservation. UOT's marginal penalty τ absorbs this mismatch, we use 30 Sinkhorn iterations on normalised downsampled latent tokens, with hyperparameter sensitivity reported in Sec. 4.4.

We obtain an aligned contribution from view q to view p by transporting features:

$$\widehat{X}^{(p \leftarrow q)} = T^{(p,q)} X_{\mathrm{hg}}^{(q)} \in \mathbb{R}^{N_p \times C}. \tag{4}$$

The *unbalanced features fusion* module then aggregates aligned evidence with the native representation:

$$X_{\mathrm{fused}}^{(p)} = X_{\mathrm{hg}}^{(p)} + \sum_{q \in \mathcal{N}(p)} \beta_{p,q}\, \widehat{X}^{(p \leftarrow q)}, \tag{5}$$

where $\mathcal{N}(p)$ denotes the other views and $\beta_{p,q}$ are learnable scalar fusion weights. In our setting, we form fused representations for the supervised output views $p \in \{\mathrm{RD}, \mathrm{RA}\}$ using contributions from the remaining views.

3.4 Adaptive Attention and Decoding

The fused representations remain subject to varying reliability across scenes and views. We therefore apply an adaptive attention mechanism before decoding to emphasise informative responses. In practice, this module can be applied multiple times in sequence to progressively refine the fused features. For each supervised view $p \in \{\mathrm{RD}, \mathrm{RA}\}$, we compute a global descriptor by average pooling $g^{(p)} = \frac{1}{N_p} \sum_{i=1}^{N_p} X_{\mathrm{fused},i}^{(p)} \in \mathbb{R}^C$ and obtain channel-wise gates via a small function $\psi(\cdot)$. The attended features are

$$\widetilde{X}^{(p)} = \psi(g^{(p)}) \odot X_{\mathrm{fused}}^{(p)}, \tag{6}$$

where $\odot$ denotes channel-wise multiplication with broadcasting over tokens. The attended tokens are reshaped back to a latent map and decoded by shallow view-specific decoders:

$$S^{(p)} = D_p\Big(\mathrm{reshape}(\widetilde{X}^{(p)})\Big), \tag{7}$$

where $S^{(p)}$ are dense semantic logits for view p.

3.5 Training Objectives

We train the network end-to-end using supervised segmentation together with cross-view consistency between RD and RA. For each supervised view $p \in \mathcal{V}_{\text{out}} = \{\text{RD}, \text{RA}\}$, the segmentation loss combines cross-entropy and soft Dice:

$$\mathcal{L}_{\text{seg}}^{(p)} = \mathcal{L}_{\text{ce}}^{(p)} + \lambda_{\text{dice}} \mathcal{L}_{\text{dice}}^{(p)}, \qquad \mathcal{L}_{\text{seg}} = \sum_{p \in \mathcal{V}_{\text{out}}} \mathcal{L}_{\text{seg}}^{(p)}.$$

Here, $\mathcal{L}_{\text{dice}}$ is computed on softmax-normalised class probabilities to mitigate foreground–background imbalance.

To enforce agreement between RD and RA predictions along their shared range dimension, let $P^{(p)} = \text{softmax}(S^{(p)})$ denote class probabilities. We extract range-wise confidence profiles using a max operator over the non-shared dimension:

$$Q^{(\text{RA})} = \max_{\text{azimuth}} P^{(\text{RA})}, \qquad Q^{(\text{RD})} = \text{Rot}_{180}\left(\max_{\text{doppler}} P^{(\text{RD})} \right),$$

where $\text{Rot}_{180}(\cdot)$ aligns the RD range axis with the RA convention. We define two complementary regularisers:

$$\mathcal{L}_{\text{coh}} = \| Q^{(\text{RD})} - Q^{(\text{RA})} \|_2^2, \qquad \mathcal{L}_{\text{mv}} = \text{Huber}\left(Q^{(\text{RD})}, Q^{(\text{RA})} \right). \tag{8}$$

The final objective is

$$\mathcal{L} = \mathcal{L}_{\text{seg}} + \lambda_{\text{coh}} \mathcal{L}_{\text{coh}} + \lambda_{\text{mv}} \mathcal{L}_{\text{mv}}, \tag{9}$$

where λ_{dice}, λ_{coh}, and λ_{mv} control the contribution of each term. All modules are optimised jointly in a single end-to-end training procedure.

4 Experiments

4.1 Datasets

We evaluate the proposed method on two public automotive radar benchmarks: CARRADA [21] and RADIal [24]. CARRADA is used as the primary benchmark for multi-class radar semantic segmentation, while RADIal is used to assess generalisation under a different sensing setup and to compare with prior work on both segmentation and detection.

CARRADA: CARRADA [21] contains synchronised radar–camera recordings collected in urban driving scenes, with 12,666 annotated frames. Labels are generated semi-automatically and provided directly in the radar domain for the Range–Angle (RA) and Range–Doppler (RD) views. The dataset includes four classes: *pedestrian*, *cyclist*, *car*, and *background*. Radar observations are represented using 2D projections (RA, RD, and Angle–Doppler (AD)) derived from the original 3D RAD tensor. The RA maps are of size $1 \times 256 \times 256$, while the

RD and AD maps are $1 \times 256 \times 64$. Following prior work, we use these 2D projections to reduce computational cost while preserving complementary spatial and kinematic information.

RADIal: RADIal [24] is a higher-resolution automotive radar benchmark with 8,252 annotated frames. In contrast to CARRADA, it provides a single radar representation and does not support multi-view radar inputs. Its annotations are obtained by projecting labels from aligned RGB images into the radar domain rather than being directly defined in radar coordinates. The radar input has resolution $32 \times 512 \times 256$, reflecting the dataset's finer-grained spatial representation. RADIal considers two semantic categories, *free space* and *occupied space* (vehicles), and is commonly used for both segmentation and detection evaluation.

4.2 Evaluation Metrics

We follow standard evaluation protocols used in radar perception. For semantic segmentation, we report Intersection over Union (IoU) and Dice (F1) scores, together with their class-wise means. Mean IoU (mIoU) is the primary metric on RADIal. Reporting both mIoU and Dice provides a balanced assessment of region overlap and class-level prediction quality. For object detection on RADIal, we follow prior work [24] and report Average Precision (AP), Average Recall (AR), and box regression errors.

4.3 Implementation Details

All models are implemented in `PyTorch` and trained on a single NVIDIA RTX 5090 GPU. On CARRADA, we use a batch size of 6 and input sequences of 5 consecutive radar frames. Optimisation is performed with Adam [15] using an initial learning rate of 1×10^{-4}. We employ an exponential learning rate schedule with a decay step of 10 epochs. The final configuration uses $8\times$ cascaded blocks of the proposed adaptive attention module. During inference, the batch size is set to 1 while keeping the same temporal context.

For RADIal, we replace the original FFTRadNet [24] backbone with the proposed architecture while retaining its segmentation and detection heads. Since RADIal is single-view, the cross-view UOT module, AD encoder, and consistency losses ($\mathcal{L}_{\text{coh}}$, $\mathcal{L}_{\text{mv}}$) are disabled, only the single-view encoder, hypergraph refinement, adaptive attention, and FFTRadNet heads are retained, isolating backbone generalisability from the multi-view design. Unless otherwise stated, we use the same optimisation strategy as on CARRADA to ensure a fair comparison across datasets.

4.4 Quantitative Results

Semantic Segmentation Results on CARRADA: Table 1 reports semantic segmentation results on the CARRADA test split for both RD and RA views.

Table 1: Semantic segmentation performance on the test split of the CARRADA dataset, shown for the RD (Range-Doppler) and RA (Range-Angle) views. Columns from left to right are the view (RD/RA), the name of the method, the intersection-over-union (IoU) score of the four different classes with their mean, and the Dice score for the same classes. Best results are shown in **black bold**, and second-best results are highlighted in **bold blue**. The Background class is near-saturated across most methods, all values tied at the column maximum are jointly highlighted.

View	Method	IoU (%)					Dice (%)				
		Bkg.	Ped.	Cycl.	Car	mIoU	Bkg.	Ped.	Cycl.	Car	mDice
RD	FCN-8s [19]	**99.7**	47.7	18.7	52.9	54.7	99.8	24.8	16.5	26.9	66.3
	U-Net [25]	**99.7**	51.1	**33.4**	37.7	55.4	99.8	67.5	**50.0**	54.7	68.0
	DeepLabv3+ [3]	**99.7**	43.2	11.2	49.2	50.8	**99.9**	60.3	20.2	66.0	61.6
	RSS-Net [14]	99.3	0.1	4.1	25.0	32.1	99.7	0.2	7.9	40.0	36.9
	RAMP-CNN [10]	**99.7**	48.8	23.2	54.7	56.6	**99.9**	65.6	37.7	70.8	68.5
	MVNet [20]	98.0	0.0	3.8	14.1	29.0	99.0	0.0	7.3	24.8	32.8
	TMVA-Net [20]	**99.7**	52.6	29.0	53.4	58.7	99.8	68.9	45.0	69.6	70.9
	PeakConv [39]	-	-	-	-	60.7	-	-	-	-	72.5
	TransRadar [6]	**99.7**	56.7	30.2	61.7	62.1	99.8	72.4	46.4	**76.3**	73.4
	HyperRadar (OT)	**99.7**	57.0	30.6	62.1	62.4	99.8	72.6	46.8	75.8	73.8
	HyperRadar (Hypergraph)	**99.7**	**57.2**	30.9	**62.3**	**62.5**	99.8	**72.9**	47.1	76.1	**74.0**
	HyperRadar (HyperGraph + OT)	**99.7**	**59.2**	**32.4**	**63.8**	**63.8**	99.8	**74.5**	**48.9**	**77.6**	**75.2**
RA	FCN-8s [19]	99.8	14.8	0.0	23.3	34.5	99.9	25.8	0.0	37.8	40.9
	U-Net [25]	99.8	22.4	**8.8**	0.0	32.8	99.9	25.8	0.0	37.8	40.9
	DeepLabv3+ [3]	**99.9**	3.4	5.9	21.8	32.7	99.9	6.5	11.1	35.7	38.3
	RSS-Net [14]	99.5	7.3	5.6	15.8	32.1	99.8	13.7	10.5	27.4	37.8
	RAMP-CNN [10]	99.8	1.7	2.6	7.2	27.9	99.9	3.4	5.1	13.5	30.5
	MVNet [20]	98.8	0.1	1.1	6.2	26.8	99.0	0.0	7.3	24.8	28.5
	TMVA-Net [20]	99.8	26.0	8.6	30.7	41.3	99.9	41.3	**15.9**	47.0	51.0
	PeakConv [39]	-	-	-	-	42.9	-	-	-	-	**53.3**
	TransRadar [6]	**99.9**	29.9	6.5	35.3	42.9	99.9	46.0	12.2	52.2	52.6
	HyperRadar (OT)	**99.9**	30.1	6.8	35.6	43.1	99.9	46.3	12.6	52.5	52.8
	HyperRadar (Hypergraph)	**99.9**	**30.4**	7.1	**35.9**	**43.3**	99.9	**46.7**	12.9	**52.9**	53.1
	HyperRadar (HyperGraph + OT)	**99.9**	**31.8**	**8.4**	**37.5**	**44.4**	**99.9**	**48.2**	**14.6**	**54.4**	**54.3**

Across both views, the full **HyperRadar** model achieves the best overall performance, with the clearest gains on sparse foreground classes such as *pedestrian* and *cyclist*, indicating the benefit of combining higher-order structural modelling with cross-view alignment.

In the RD view, HyperRadar achieves the best mIoU/mDice of **63.8**/**75.2**%, improving over TransRadar (62.1/73.4%) by +**1.7**/+**1.8** points. Gains are most evident for foreground classes, with pedestrian, cyclist, and car IoU improving to **59.2**%, **32.4**%, and **63.8**%, respectively. In the more challenging RA view, HyperRadar again yields the best mIoU/mDice of **44.4**/**54.3**%, compared with 42.9/52.6% for TransRadar. The largest relative gain is for cyclists, where IoU increases from 6.5% to **8.4%**, while pedestrian and car IoU also improve to **31.8**% and **37.5**%.

The ablation results further show that the two proposed components are complementary. In RD, OT alone and hypergraph refinement alone achieve 62.4% and 62.5% mIoU, respectively, while their combination reaches **63.8%**. The same trend holds in RA (43.1%, 43.3%, and **44.4%**), confirming that hypergraph

Table 2: Results on the RADIal benchmark. We compare different backbones for joint detection and segmentation. AP and AR denote average precision and average recall, respectively; R and A denote range and angle errors. Best results are shown in **black bold**, and second-best results are highlighted in **bold blue**.

Backbone	AP (%) ↑	AR (%) ↑	R (m) ↓	A (°) ↓	mIoU (%) ↑
Pixor [34]	96.6	81.7	**0.10**	0.20	–
FFTRadNet [24]	96.8	82.2	**0.11**	0.17	74.0
C-M DNN [13]	96.9	83.5	–	–	80.4
TransRadar [6]	**97.3**	**98.4**	**0.11**	**0.10**	**81.1**
HyperRadar	**98.5**	**98.9**	**0.10**	**0.08**	**83.4**

refinement improves intra-view structural coherence, whereas OT strengthens cross-view consistency; together they deliver the most reliable segmentation across both radar views.

Semantic Segmentation Results on RADIal: Table 2 reports results on the RADIal benchmark for both detection and segmentation. The proposed **HyperRadar** achieves the best overall performance across all reported metrics, demonstrating that the learned representation generalises effectively beyond CARRADA and remains robust under a different sensing configuration.

Compared with the strongest prior method, TransRadar, HyperRadar improves AP from 97.3% to **98.5%** and AR from 98.4% to **98.9%**. At the same time, it reduces the angle error from 0.10° to **0.08°**, while matching the best reported range error of **0.10 m**. These gains indicate that the proposed backbone not only improves detection confidence and recall, but also yields more precise localisation.

For segmentation, HyperRadar achieves the highest mIoU of **83.4%**, outperforming TransRadar (81.1%) by +**2.3** points and C-M DNN (80.4%) by +**3.0** points. The margin over FFTRadNet is larger still (+**9.4** points), highlighting the benefit of the proposed representation over earlier radar-specific backbones. Since RADIal provides a single-view radar setting, these improvements show that the proposed architecture remains effective even when the full multi-view configuration is not available.

Overall, the consistent gains across AP, AR, localisation error, and mIoU show that HyperRadar provides a stronger and more transferable backbone for radar perception. The results suggest that the proposed structural refinement and feature alignment strategy improve both dense semantic understanding and downstream detection quality.

Computational Cost and Runtime: We also compare the computational footprint of **HyperRadar** against TransRadar in terms of GFLOPs, inference throughput (FPS), and peak GPU memory. HyperRadar requires slightly fewer computations (**1132.13** vs. 1138.57 GFLOPs) and lower GPU memory (**3.18**

Table 3: Component ablation and UOT sensitivity on CARRADA. Default: ε=0.05, τ=0.8; no loss divergence across all settings.

(a) Component ablation

HG	UOT	Attn.	RD	RA	Avg.
–	–	✓	62.1	42.9	52.5
–	✓	–	61.7	42.2	52.0
–	✓	✓	62.4	43.1	52.8
✓	–	–	61.9	42.4	52.2
✓	–	✓	62.5	43.3	52.9
✓	✓	–	63.0	43.7	53.4
✓	✓	✓	**63.8**	**44.4**	**54.1**

(b) UOT sensitivity

ε	τ	RD	RA	Avg.
0.01	0.8	63.4	44.0	53.7
0.05	**0.8**	**63.8**	**44.4**	**54.1**
0.10	0.8	63.5	44.1	53.8
0.05	0.5	63.3	43.8	53.6
0.05	1.0	63.6	44.2	53.9

vs. 3.25 GB), while also achieving higher inference speed (**10.55** vs. 9.6 FPS). These results show that the proposed structural modelling improves segmentation performance without increasing computational cost and, in practice, offers a modest efficiency gain over the strongest baseline.

Table 3(a) shows each component of HyperRadar is complementary: the attention-only baseline obtains **52.5** Avg.; adding UOT or HG individually reaches **52.8** and **52.9**; the full HG + UOT + Attention model achieves **54.1** Avg. (**63.8**/**44.4** RD/RA), a +**1.6** Avg. gain. Table 3(b) confirms robustness: $\varepsilon \in \{0.01, 0.05, 0.10\}$ and $\tau \in \{0.5, 0.8, 1.0\}$ keep Avg. mIoU within **53.6**–**54.1** with no loss divergence, and performance plateaus at M=8 hyperedges, ruling out hyperedge-count sensitivity.

4.5 Qualitative Results on CARRADA

Figure 2 presents qualitative comparisons on two representative scenes from the CARRADA test split, showing the RGB image together with RD and RA semantic predictions from different methods. As seen in both examples, the proposed **HyperRadar** produces predictions that are more compact, better localised, and more semantically consistent than the competing methods, particularly for sparse foreground objects and weak radar returns.

In the first scene, which contains vehicles with relatively weak and spatially concentrated responses, HyperRadar recovers cleaner foreground structure in both RD and RA views while preserving the extent of the *car* regions more faithfully. Compared with TransRadar, TMVA-Net, MVNet, and U-Net, our predictions exhibit fewer fragmented activations and less background leakage. Competing methods either miss parts of the target, produce overly sparse responses, or introduce noisy foreground regions, whereas HyperRadar remains closer to the ground-truth shape and location across both radar projections.

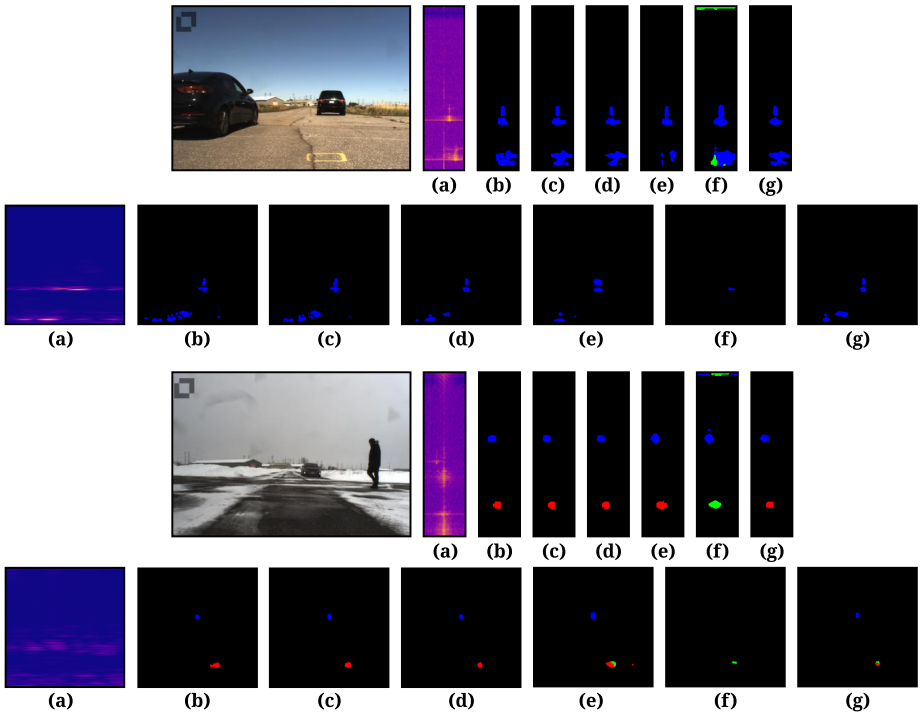

Fig. 2: Qualitative comparisons on two test scenes from the CARRADA test split, showing the RGB camera view alongside semantic segmentation outputs from different methods. For each scene, the top row corresponds to the RD view and the bottom row to the RA view. (a) RD/RA inputs, (b) ground truth, (c) Our method, (d) TransRadar [6], (e) TMVA-Net [20], (f) MVNet [20], and (g) U-Net [25]. All RD outputs are rotated for visual consistency. Colours denote semantic classes: blue for Car, green for Cyclist, red for Pedestrian, and black for background.

This is especially visible in the RD view, where our method better preserves the main object response while avoiding scattered false positives.

In the second scene, which is more challenging due to the presence of multiple object categories and weaker radar evidence, HyperRadar again yields the most coherent predictions. In particular, it better separates the *pedestrian* and *car* regions and preserves their relative spatial layout in both RD and RA. Baseline methods show stronger confusion between sparse foreground classes, with some methods suppressing small targets and others producing incomplete or noisy masks. By contrast, HyperRadar produces sharper and more stable activations, indicating improved robustness to sparsity and class ambiguity.

Overall, the qualitative results support the quantitative findings: HyperRadar more reliably captures sparse foreground objects, reduces fragmented predictions, and yields masks that are visually closer to the ground truth across both RD and RA views. These visual improvements are consistent with the proposed

higher-order structural modelling and cross-view alignment, which help the network maintain coherent object-level predictions under noisy radar sensing.

5 Limitations and Future Work

Although **HyperRadar** achieves strong results, it has some limitations. First, the current formulation operates on 2D RA/RD/AD projections rather than the full RAD tensor, which is efficient but may discard some fine-grained 3D structure. Second, the cross-view regularisation mainly exploits the shared range dimension between RA and RD, providing only a partial geometric constraint and not explicitly modelling longer temporal consistency. Finally, our evaluation is limited to CARRADA and RADIal, which, although widely used, do not cover the full diversity of radar sensors and deployment conditions. Future work will therefore focus on improving efficiency through sparse or hierarchical formulations, extending the method to richer 3D and temporal modelling, and exploring multi-modal and multi-task settings such as radar-camera fusion and joint segmentation, detection, and tracking.

6 Conclusion

We introduced **HyperRadar**, a compact and structurally grounded framework for multi-view radar semantic segmentation. By combining learnable hypergraph-based feature refinement, unbalanced optimal transport for correspondence-free cross-view alignment, and adaptive attention-based fusion, the proposed method addresses key limitations of existing radar segmentation approaches that rely primarily on local or pairwise interactions.

HyperRadar is designed to better match the sparse, noisy, and partially observed nature of automotive radar data, enabling more coherent feature learning within each view and more robust alignment across views. Extensive experiments on CARRADA and RADIal show that this design consistently improves over strong radar-specific baselines, yielding better segmentation accuracy and stronger overall perception performance with an efficient architecture.

These results highlight the value of higher-order relational modelling for radar perception and suggest that structurally aware representation learning is a promising direction for robust scene understanding in adverse sensing conditions. Future work will extend this framework to longer temporal reasoning, multi-modal fusion, and unified perception pipelines for segmentation, detection, and tracking.

References

1. Bansal, K., Rungta, K., Bharadia, D.: Radsegnet: A reliable approach to radar camera fusion. arXiv preprint arXiv:2208.03849 (2022)
2. Barnes, D., Gadd, M., Murcutt, P., Newman, P., Posner, I.: The oxford radar robotcar dataset: A radar extension to the oxford robotcar dataset. In: 2020 IEEE international conference on robotics and automation (ICRA). pp. 6433–6438. IEEE (2020)
3. Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H.: Encoder-decoder with atrous separable convolution for semantic image segmentation. In: Proceedings of the European conference on computer vision (ECCV). pp. 801–818 (2018)
4. Courty, N., Flamary, R., Tuia, D., Rakotomamonjy, A.: Optimal transport for domain adaptation. IEEE transactions on pattern analysis and machine intelligence **39**(9), 1853–1865 (2016)
5. Cuturi, M.: Sinkhorn distances: Lightspeed computation of optimal transport. Advances in neural information processing systems **26** (2013)
6. Dalbah, Y., Lahoud, J., Cholakkal, H.: Transradar: Adaptive-directional transformer for real-time multi-view radar semantic segmentation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 353–362 (2024)
7. Feng, D., Haase-Schütz, C., Rosenbaum, L., Hertlein, H., Glaeser, C., Timm, F., Wiesbeck, W., Dietmayer, K.: Deep multi-modal object detection and semantic segmentation for autonomous driving: Datasets, methods, and challenges. IEEE Transactions on Intelligent Transportation Systems **22**(3), 1341–1360 (2020)
8. Feng, Y., You, H., Zhang, Z., Ji, R., Gao, Y.: Hypergraph neural networks. In: Proceedings of the AAAI conference on artificial intelligence. vol. 33, pp. 3558–3565 (2019)
9. Fent, F., Bauerschmidt, P., Lienkamp, M.: Radargnn: Transformation invariant graph neural network for radar-based perception. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 182–191 (2023)
10. Gao, X., Xing, G., Roy, S., Liu, H.: Ramp-cnn: A novel neural network for enhanced automotive radar object recognition. IEEE Sensors Journal **21**(4), 5119–5132 (2020)
11. Hao, Y., Madani, S., Guan, J., Alloulah, M., Gupta, S., Hassanieh, H.: Bootstrapping autonomous driving radars with self-supervised learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15012–15023 (2024)
12. Jiang, T., Zhuang, L., An, Q., Wang, J., Xiao, K., Wang, A.: T-rodnet: Transformer for vehicular millimeter-wave radar object detection. IEEE Transactions on Instrumentation and Measurement **72**, 1–12 (2022)
13. Jin, Y., Deligiannis, A., Fuentes-Michel, J.C., Vossiek, M.: Cross-modal supervision-based multitask learning with automotive radar raw data. IEEE Transactions on Intelligent Vehicles **8**(4), 3012–3025 (2023)
14. Kaul, P., De Martini, D., Gadd, M., Newman, P.: Rss-net: Weakly-supervised multi-class semantic segmentation with fmcw radar. In: 2020 IEEE Intelligent Vehicles Symposium (IV). pp. 431–436. IEEE (2020)
15. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)

16. Li, S., Hong, Z., Chen, Y., Hu, L., Qin, J.: Get it for free: Radar segmentation without expert labels and its application in odometry and localization. IEEE Robotics and Automation Letters (2025)
17. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 10012–10022 (2021)
18. Lombacher, J., Laudt, K., Hahn, M., Dickmann, J., Wöhler, C.: Semantic radar grids. In: 2017 IEEE intelligent vehicles symposium (IV). pp. 1170–1175. IEEE (2017)
19. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3431–3440 (2015)
20. Ouaknine, A., Newson, A., Pérez, P., Tupin, F., Rebut, J.: Multi-view radar semantic segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 15671–15680 (2021)
21. Ouaknine, A., Newson, A., Rebut, J., Tupin, F., Pérez, P.: Carrada dataset: Camera and automotive radar with range-angle-doppler annotations. In: 2020 25th International Conference on Pattern Recognition (ICPR). pp. 5068–5075. IEEE (2021)
22. Palffy, A., Dong, J., Kooij, J.F., Gavrila, D.M.: Cnn based road user detection using the 3d radar cube. IEEE Robotics and Automation Letters **5**(2), 1263–1270 (2020)
23. Peyré, G., Cuturi, M.: Computational optimal transport: With applications to data science. Now Foundations and Trends (2019)
24. Rebut, J., Ouaknine, A., Malik, W., Pérez, P.: Raw high-definition radar for multi-task learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 17021–17030 (2022)
25. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. pp. 234–241. Springer (2015)
26. Schumann, O., Hahn, M., Scheiner, N., Weishaupt, F., Tilly, J.F., Dickmann, J., Wöhler, C.: Radarscenes: A real-world radar point cloud data set for automotive applications. In: 2021 IEEE 24th International Conference on Information Fusion (FUSION). pp. 1–8. IEEE (2021)
27. Séjourné, T., Peyré, G., Vialard, F.X.: Unbalanced optimal transport, from theory to numerics. Handbook of Numerical Analysis **24**, 407–471 (2023)
28. Sless, L., El Shlomo, B., Cohen, G., Oron, S.: 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. pp. 0–0 (2019)
29. Wang, Y., Jiang, Z., Gao, X., Hwang, J.N., Xing, G., Liu, H.: Rodnet: Radar object detection using cross-modal supervision. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 504–513 (2021)
30. Xiong, W., Liu, J., Xia, Y., Huang, T., Zhu, B., Xiang, W.: Contrastive learning for automotive mmwave radar detection points based instance segmentation. In: 2022 IEEE 25th International Conference on Intelligent Transportation Systems (ITSC). pp. 1255–1261. IEEE (2022)
31. Xu, S., Wan, R., Ye, M., Zou, X., Cao, T.: Sparse cross-scale attention network for efficient lidar panoptic segmentation. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 36, pp. 2920–2928 (2022)

32. Xue, R., Wang, J., Ma, Z.: Efficient lidar point cloud geometry compression through neighborhood point attention. arXiv preprint arXiv:2208.12573 (2022)
33. Yadati, N., Nimishakavi, M., Yadav, P., Nitin, V., Louis, A., Talukdar, P.: Hypergcn: A new method for training graph convolutional networks on hypergraphs. Advances in neural information processing systems **32** (2019)
34. Yang, B., Luo, W., Urtasun, R.: Pixor: Real-time 3d object detection from point clouds. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. pp. 7652–7660 (2018)
35. Yao, S., Guan, R., Huang, X., Li, Z., Sha, X., Yue, Y., Lim, E.G., Seo, H., Man, K.L., Zhu, X., et al.: Radar-camera fusion for object detection and semantic segmentation in autonomous driving: A comprehensive review. IEEE Transactions on Intelligent Vehicles **9**(1), 2094–2128 (2023)
36. Zhang, A., Nowruzi, F.E., Laganiere, R.: Raddet: Range-azimuth-doppler based radar object detection for dynamic road users. In: 2021 18th Conference on Robots and Vision (CRV). pp. 95–102. IEEE (2021)
37. Zhang, B., Titov, I., Sennrich, R.: Sparse attention with linear units. arXiv preprint arXiv:2104.07012 (2021)
38. Zhang, G., Li, H., Wenger, F.: Object detection and 3d estimation via an fmcw radar using a fully convolutional network. In: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 4487–4491. IEEE (2020)
39. Zhang, L., Zhang, X., Zhang, Y., Guo, Y., Chen, Y., Huang, X., Ma, Z.: Peakconv: Learning peak receptive field for radar semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 17577–17586 (2023)
40. Zhang, Z., Tian, Z., Zhang, Y., Zhou, M., Wang, B.: u-deephand: Fmcw radar-based unsupervised hand gesture feature learning using deep convolutional auto-encoder network. IEEE Sensors Journal **19**(16), 6811–6821 (2019)
41. Zhang, Z., Tian, Z., Zhou, M.: Latern: Dynamic continuous hand gesture recognition using fmcw radar sensor. IEEE Sensors Journal **18**(8), 3278–3289 (2018)
42. Zhou, D., Huang, J., Schölkopf, B.: Learning with hypergraphs: Clustering, classification, and embedding. Advances in neural information processing systems **19** (2006)