HyperEcho: Structured Higher Order Representation for Automotive Radar Semantic Segmentation

Abdelwahed Khamis[1], Muhammad Umer Ramzan[2], Usman Ali[2], Ali Zia[3]
[1]CSIRO, Australia
abdelwahed.khamis@data61.csiro.au
[2]GIFT University, Pakistan
[3] La Trobe University, Australia

***Abstract*—Automotive radar is attractive for perception in adverse weather and lighting conditions, but semantic segmentation on radar maps is still difficult. Objects in radar heatmaps are sparse and noisy, appearing as extended blobs or streaks rather than neat contours. Most existing CNN and transformer backbones treat these maps like ordinary images and rely on local filters or pairwise attention, which makes it hard to capture the larger echo patterns that define objects. We propose *HyperEcho*, a backbone that is tailored to radar feature maps. Instead of reasoning only over individual pixels, HyperEcho builds a simple hypergraph on top of the feature map: it learns a few groups of positions along the range and angle/Doppler axes that each capture an echo pattern, runs a lightweight attention mechanism over these groups, and then propagates the result back to the underlying pixels. This axial hypergraph backbone can be used on its own or combined with a lightweight contrastive loss that encourages consistent features across radar views. On the CARRADA benchmark, HyperEcho consistently improves semantic segmentation over strong CNN and transformer baselines. Our best variant sets a new state of the art on both Range–Doppler and Range–Angle segmentation, achieving 62.4% mIoU on RD and 45.1% mIoU on RA, showing that explicitly modelling higher-order echo structure is beneficial for radar scene understanding.**

***Index Terms*—mmWave Radar, Semantic Segmentation, Automotive, Hypergraphs**

I. Introduction

Automotive radar has become an essential sensing modality for perception in autonomous systems due to its robustness to adverse weather, low cost, and ability to measure both range and velocity [13], [8]. Recent radar–vision datasets [12], [14] have enabled learning-based approaches for dense scene understanding, including semantic segmentation [5], [3], [11], [18], [2]. Despite this progress, radar segmentation remains challenging: echoes are extremely sparse and noisy, object signatures appear as fragmented structures spread across range, angle, and Doppler, and the same object may give rise to multiple disconnected reflections due to multipath, partial visibility, or angular resolution limits.

Most existing radar segmentation networks handle these difficulties using either local convolutional operators [5], [3] or pairwise attention mechanisms [2]. While effective to a degree, these architectures fundamentally model interactions *between individual pixels*. This limits their ability to capture the higher–order echo patterns that are typical in radar: a car may produce several spatially separate reflections in a range–angle map, yet these reflections belong to the same underlying object and should be processed jointly. Pairwise operations struggle to associate such groups without many layers, large receptive fields, or heavy computation.

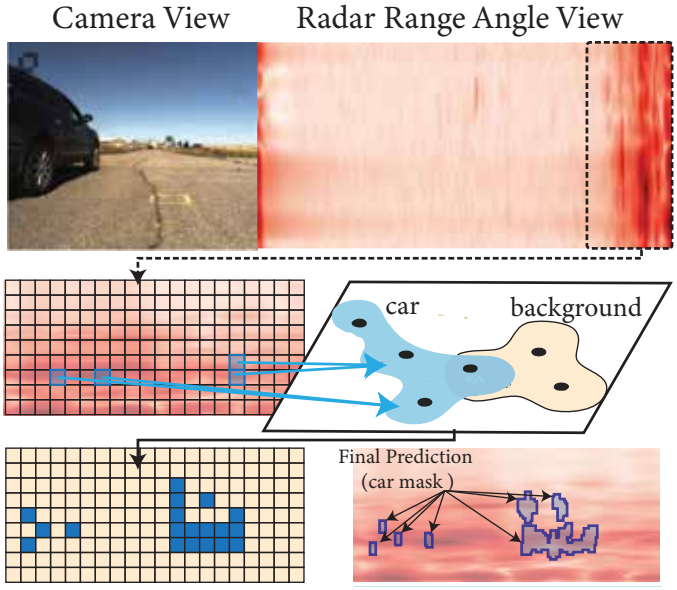

Fig. 1. Radar echoes from a single object often appear as multiple spatially distant peaks in the range–angle map. Hypergraph aggregation groups these scattered reflections into a single latent hyperedge (blue region marked with 'car'), allowing the model to reason about them jointly rather than through local or pairwise interactions only. This higher–order structure helps the network form a coherent car mask despite the fragmented appearance of radar measurements.

In this work, we argue that radar perception benefits from modelling *groups of related positions* directly. Hypergraphs provide exactly this capability: a hyperedge can connect an arbitrary set of spatial positions that together form an echo pattern, enabling a single operation to aggregate and reason over them jointly. This form of higher–order interaction is a natural fit for the anisotropic and structured nature of radar returns.

We therefore introduce **HyperEcho**, a new backbone for radar semantic segmentation that performs learned hypergraph aggregation along the physical axes of radar feature maps (range and angle/Doppler). As shown in Figure 1, instead of operating only on individual pixels, HyperEcho learns latent hyperedges that capture coherent echo patterns and passes messages through these higher–order units. This enables long-range reasoning with far fewer layers than standard 2D convolutions or pairwise attention, while respecting the directional geometry of radar.

HyperEcho can be inserted into existing radar architectures with minimal change and further benefits from a lightweight multi-view contrastive regularizer that encourages consistency across radar modalities. Across both RD and RA views on CARRADA dataset, HyperEcho achieves state-of-the-art or competitive performance, particularly on small and fragmented foreground classes. In this work, we make the following contributions:

- We introduce a radar-specific hypergraph backbone that performs learned higher–order aggregation along range and angle/Doppler axes, enabling compact modelling of extended and fragmented echo patterns.
- We propose a multi-view contrastive regularizer that improves feature consistency across radar modalities without requiring additional sensors.
- We achieve state-of-the-art performance on CARRADA and provide the first demonstration that hypergraph message passing is effective for dense radar perception tasks.

II. Related Works

Recent radar–perception methods have mainly extended image–based architectures to radar maps. TMVA-Net [11] is a multi–view encoder–decoder built entirely from convolutional layers and provides a strong baseline on RD and RA views. RAMP-CNN [3] repurposes a 3D CNN originally designed for RAD tensors, while T-RODNet [4] uses Swin Transformers [6] but operates only on RA inputs. PeakConv [18] focuses the receptive field around local signal peaks to better exploit sparsity, at the cost of increased parameter count. In parallel, sparse attention mechanisms such as ReLA [17] learn to concentrate attention on informative locations or point sets. These approaches improve robustness on noisy, sparse radar data, but they still model interactions largely at the level of individual pixels or pairs of pixels, rather than explicitly capturing higher–order echo patterns, which is the focus of our hypergraph formulation.

TransRadar [2] is the closest to our work. It introduces an adaptive directional transformer for radar semantic segmentation. HyperEcho follows the general multi-view encoder–fusion–decoder design of TransRadar , including axis-aligned directional processing and multi-view supervision, but replaces the Adaptive Directional Attention block with a hypergraph aggregation backbone and augments it with a contrastive projection head for view-consistent representations. Empirically, this higher–order, hypergraph–based design yields stronger performance, especially on small foreground classes and on the more challenging RA view.

III. Radar Semantic Segmentation

Automotive FMCW radar provides, at every time step, several two–dimensional range–parameter maps (such as range–Doppler or range–angle). Semantic segmentation in this setting requires predicting, for each output view, a semantic label for every pixel of the radar map. Given a sequence of recent frames, the task is to infer the spatial layout of objects directly from radar measurements, despite their sparsity, noise, and anisotropy.

Formally, let $X_v^{(t)} \in \mathbb{R}^{H\times W}$ denote the v-th radar view (e.g. RD or RA) at time t, with $v \in \{1,\dots,V\}$. We write

$$X^{(t),V} = \{X_v^{(t)}\}_{v=1}^{V}$$

for the collection of all views at that time step, and use X^V when the explicit time index is not important. Given X^V, the goal of radar semantic segmentation is to assign a semantic label to every pixel of each output view, using only the radar measurements. Next, we introduce the necessary hypergraph background, as it forms the basis of our approach to radar semantic segmentation.

A. Hypergraph Background

Automotive radar measurements are sparse and highly structured: a single physical object often produces several reflections that are spread over non-adjacent pixels in range–angle or range–Doppler space. For example, a car may generate multiple bright returns at different angles and ranges due to its metallic corners and curved surfaces. These reflections are spatially distant on the radar grid, yet they arise from the same object and should ideally be processed jointly. Figure 1 illustrates this effect: several separated peaks in the radar map belong to the same car and should be grouped together before downstream classification or segmentation.

a) Limitations of common alternatives.: Standard 2D convolutions struggle with this behaviour because they aggregate information only through small local windows; relating distant but correlated peaks requires many stacked layers and large receptive fields, which is computationally inefficient for high-resolution radar maps. Transformers [4] alleviate this by allowing pairwise interactions across the entire image, but they do so through attention between *individual* pixels. Pairwise attention is powerful but not object-centric: it cannot jointly model a *set* of related reflections as a single unit, and its quadratic cost in the number of pixels is prohibitive on dense feature maps.

Graph-based models [16] provide another alternative, but graphs still connect pairs of nodes and therefore encode only binary relations. Radar echoes, however, often form *multi-point patterns*—elongated streaks, curved contours, or compact clusters—whose joint structure cannot be captured by edge-based interactions alone.

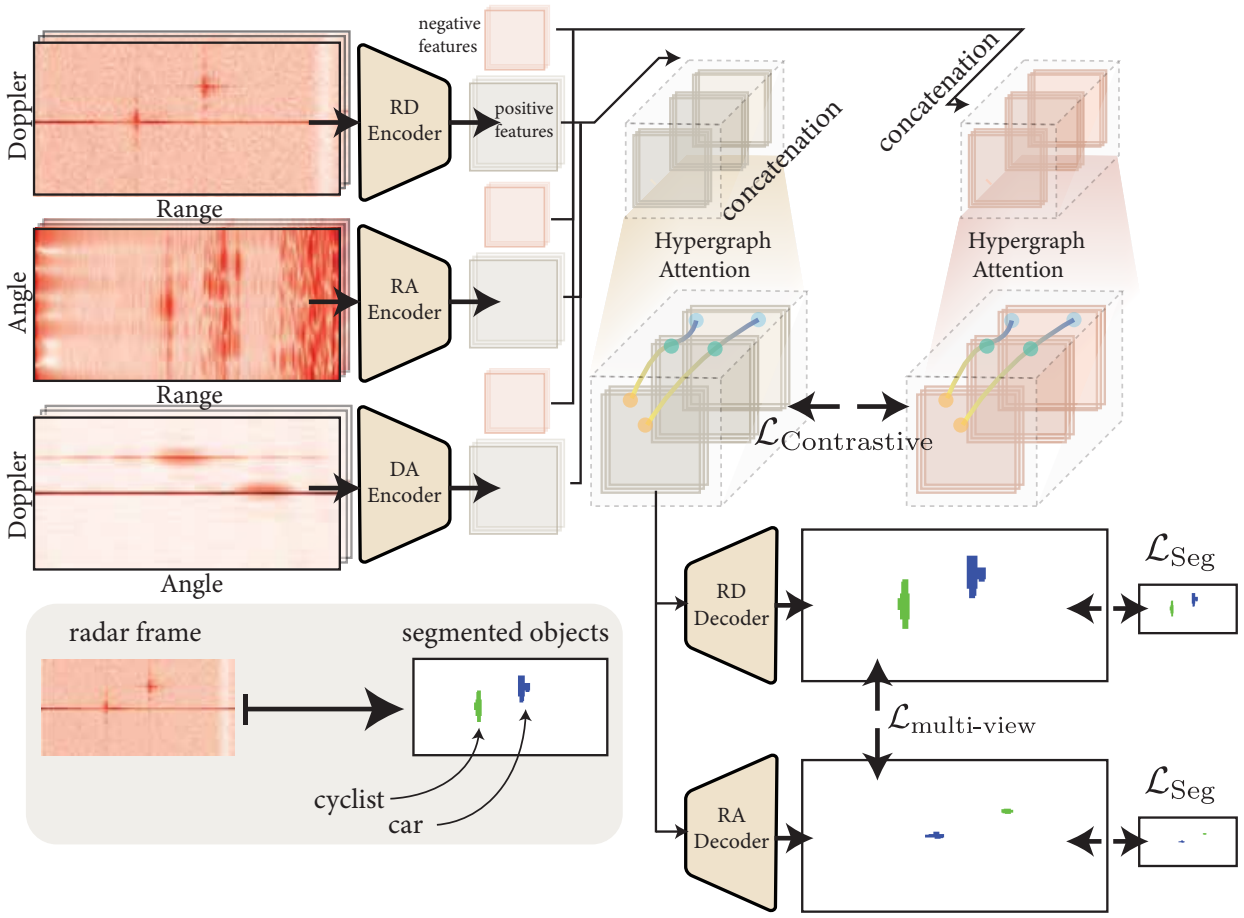

Fig. 2. **HyperEcho architecture.** Multi-view radar inputs (e.g., range–Doppler, range–angle, Doppler–angle) are encoded separately and their feature maps are concatenated into a fused latent representation. This fused tensor is processed by a shared Hypergraph Attention backbone that captures higher-order echo patterns; two augmented versions of the fused features are used to compute a contrastive loss $\mathcal{L}_{\text{contrastive}}$. The backbone output is decoded into RD and RA segmentation maps with lightweight decoders, trained using per-view segmentation losses $\mathcal{L}_{\text{Seg}}$ and a multi-view consistency loss $\mathcal{L}_{\text{multi-view}}$, as illustrated for an example frame containing a car and a cyclist.

b) Motivation for hypergraphs.: Hypergraphs [19] generalise graphs by allowing a single hyperedge to connect an arbitrary subset of positions. This enables the model to treat an entire echo pattern—even when discontinuous or spatially scattered—as a coherent higher-order entity. In the radar setting, this is precisely what is needed: the reflections marked in Figure 1 can be grouped under one hyperedge, allowing information to flow between them as a single unit before being returned to their original spatial locations.

c) Minimal formalism used in later sections.: A hypergraph is defined as $\mathcal{H} = (V, \mathcal{E})$, where $V = \{1, \ldots, T\}$ is a set of nodes (here: radar positions along an axis) and $\mathcal{E} = \{e_1, \ldots, e_M\}$ is a set of hyperedges, each connecting an arbitrary subset of nodes. We represent the connectivity using an incidence matrix $H \in \{0,1\}^{T \times M}$, where $H_{im} = 1$ if node i belongs to hyperedge e_m.

Given node features $X \in \mathbb{R}^{T \times C}$, a hypergraph message-passing step aggregates information from nodes to hyperedges and broadcasts it back:

$$E = H^\top X, \qquad X' = X + H\,\phi(E), \tag{1}$$

where $E \in \mathbb{R}^{M \times C}$ are hyperedge features and ϕ is a learnable transformation. This pattern captures higher-order interactions: a hyperedge collects information from *all* its nodes, processes them jointly, and returns a coordinated update to those same positions.

In Section III-B we instantiate this mechanism directly on radar feature maps. Instead of hand-constructed hyperedges, HyperEcho learns a soft incidence matrix that groups positions along range and angle/Doppler axes into latent hyperedges corresponding to coherent echo patterns. The forward–backward structure in (1) is reused verbatim in the backbone, enabling long-range aggregation through a small number of data-driven hyperedges while preserving the original spatial layout.

B. HyperEcho Architecture

The design of HyperEcho is guided by the structure of automotive radar data and the requirements of semantic segmentation. Radar maps are extremely sparse and noisy, yet objects appear as *coherent echo patterns* that extend along physically meaningful axes (range, angle, Doppler). A good backbone must therefore (i) extract robust low–level fea-

tures from each radar view, (ii) connect distant but related echoes along these axes, and (iii) remain efficient enough for real–time use. HyperEcho addresses these points with three components: a multi–view encoder and fusion stage, an axial hypergraph backbone that implements the higher–order updates from Section III-A, and lightweight decoders with radar–aware training losses.

a) Multi–View Encoder and Fusion.: We start from the per–acquisition radar views $\{X_v\}_{v=1}^{V}$ defined in the problem formulation (e.g. RD and RA). The goal of the encoder stage is not to perform long–range reasoning, but to denoise and compress each view into a common latent grid where subsequent hypergraph operations can act.

For each view v we use a compact convolutional encoder E_v producing a feature map $F_v = E_v(X_v)$ at a reduced spatial resolution. We deliberately avoid heavy backbones (e.g. large ResNets or full vision transformers), which would dominate the computation and make it difficult to attribute improvements to the hypergraph reasoning itself. A separate encoder per view allows the network to adapt to view–specific artifacts (e.g. sidelobes in RD vs. RA) without entangling them prematurely.

The encoded views are then fused by channel–wise concatenation, yielding a single latent tensor F aligned in range and angle (or range and Doppler). We chose concatenation over additive fusion or early cross–attention so that each view retains its own feature subspace while still sharing a common spatial grid.

b) Axial Hypergraph Backbone.: Once the multi–view features are fused, the main challenge is to link scattered echoes into coherent object hypotheses. Standard 2D convolutions can only do this gradually, via many layers and large receptive fields, while full image–level transformers are too expensive for high–resolution radar maps. Moreover, neither is tailored to the fact that echoes tend to organise *along* range and angle/Doppler axes rather than isotropically.

To exploit this structure, HyperEcho employs a stack of L *axial hypergraph* blocks. Each block applies the message–passing pattern of Equation (1) along one spatial axis at a time. Intuitively, we treat each row or column of the fused feature map as a one–dimensional sequence of nodes, let the model discover a small number of latent hyperedges that capture echo patterns along that axis, and then send messages back from those hyperedges to the original positions.

Axial view. Rather than building a hypergraph over the full $H_d \times W_d$ grid, we construct simpler one–dimensional hypergraphs. Inspired by TransRadar [2], we operate along the physical axes. Let $F \in \mathbb{R}^{B\times C\times H_d\times W_d}$ denote the fused feature map, where B is the batch size, C the number of channels, and H_d, W_d the downsampled spatial dimensions. For the range axis, we reinterpret each column of F as a sequence of $T = H_d$ nodes with C–dimensional features; for the angle/Doppler axis, we do the same with each row ($T = W_d$). This is implemented by permuting and reshaping F so that the last two dimensions are "position along the chosen axis" and "feature channel", matching the node representation in Section III-A. This axial construction is a compromise between full 2D hypergraphs (too expensive) and standard axial self–attention (which still only models pairwise interactions).

Hyperedge interactions and scattering. For each 1D node sequence obtained from the fused feature map F along a range or angle/Doppler axis we write $F_{\text{axis}} \in \mathbb{R}^{T\times C}$ for the corresponding T positions with C-dimensional features. Using the learned incidence matrix $H \in \mathbb{R}^{T\times M}$, we first aggregate node features into M latent hyperedges

$$E = H^{\top} F_{\text{axis}}, \tag{2}$$

and process these hyperedge features with a lightweight transformer block: multi–head self–attention across hyperedges, followed by a feed–forward network, yields updated features $\tilde{E}$. We then scatter the result back to the node sequence via the same incidence matrix,

$$\Delta F_{\text{axis}} = H\tilde{E}, \tag{3}$$

and add it with a gated residual connection $F^{+}_{\text{axis}} = F_{\text{axis}} + \gamma\,\psi(\Delta F_{\text{axis}})$, where ψ is a linear projection and γ is a scalar gate initialised to zero. Thus each block behaves as an identity at the start of training and only gradually introduces higher–order interactions, which stabilises optimisation and prevents the hypergraph machinery from overriding reliable low–level cues too early.

The same procedure is applied to 1D sequences along both spatial axes inside each block

c) Decoders and Training Objective.: Starting from the shared latent map $F^{\star}$ produced by the axial hypergraph backbone, we attach a small decoder D_v for each output view $v \in \mathcal{V}_{\text{out}}$. Each decoder simply upsamples $F^{\star}$ back to the radar resolution obtain semantic logits $S_v = D_v(F^{\star})$ as required in the problem formulation. We deliberately keep the decoders lightweight so that most of the capacity is concentrated in the shared backbone.

Training uses a composite loss tailored to radar segmentation. A standard per-pixel segmentation loss (cross-entropy or Dice [9]) is complemented by (i) a foreground/background term to counteract the extreme sparsity and class imbalance of radar returns, and (ii) a multi-view consistency term that encourages RD and RA predictions to agree along their shared range axis. These components encourage features that are both pixel-wise discriminative and geometrically consistent across views.

For the variants that use contrastive regularisation, we add a lightweight projection head on top of the backbone features that compresses each radar view into a single embedding per sample. During training, we apply an InfoNCE contrastive loss [10]: embeddings from different radar views of the same scene are treated as positives, while embeddings from other scenes in the mini–batch act as negatives, encouraging features that are consistent across radar modalities yet well separated between scenes.

TABLE I

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 NUMBER OF PARAMETERS IN MILLIONS, THE INTERSECTION-OVER-UNION (IOU) SCORE OF THE FOUR DIFFERENT CLASSES WITH THEIR MEAN, AND THE DICE SCORE FOR THE SAME CLASSES.

View	Method	IoU (%)					Dice (%)				
		Bkg.	Ped.	Cycl.	Car	mIoU	Bkg.	Ped.	Cycl.	Car	mDice
RD	FCN-8s [7]	99.7	47.7	18.7	52.9	54.7	99.8	64.8	16.5	26.9	66.3
	U-Net [15]	99.7	51.1	33.4	37.7	55.4	99.8	67.5	50.0	54.7	68.0
	DeepLabv3+ [1]	99.7	43.2	11.2	49.2	50.8	99.9	60.3	20.2	66.0	61.6
	RSS-Net [5]	99.3	0.1	4.1	25.0	32.1	99.7	0.2	7.9	40.0	36.9
	RAMP-CNN [3]	99.7	48.8	23.2	54.7	56.6	99.9	65.6	37.7	70.8	68.5
	MVNet [11]	98.0	0.0	3.8	14.1	29.0	99.0	0.0	7.3	24.8	32.8
	TMVA-Net [11]	99.7	52.6	29.0	53.4	58.7	99.8	68.9	45.0	69.6	70.9
	PeakConv [18]	-	-	-	-	60.7	-	-	-	-	72.5
	TransRadar [2]	99.7	56.68	30.22	**61.71**	62.09	99.8	72.35	46.41	**76.32**	73.43
	HyperEcho$_{HG}$	99.7	55.26	20.76	57.87	58.4	99.8	71.18	34.38	73.31	69.68
	HyperEcho$_{CL}$	99.7	57.7	31.36	57.66	61.6	99.8	73.16	47.75	73.14	73.48
	HyperEcho	99.7	**60.33**	**31.86**	57.7	**62.42**	99.8	**75.26**	**48.32**	73.21	**74.16**
RA	FCN-8s [7]	99.8	14.8	0.0	23.3	34.5	99.9	25.8	0.0	37.8	40.9
	U-Net [15]	99.8	22.4	8.8	0.0	32.8	99.9	25.8	0.0	37.8	40.9
	DeepLabv3+ [1]	99.9	3.4	5.9	21.8	32.7	99.9	6.5	11.1	35.7	38.3
	RSS-Net [5]	99.5	7.3	5.6	15.8	32.1	99.8	13.7	10.5	27.4	37.8
	RAMP-CNN [3]	99.8	1.7	2.6	7.2	27.9	99.9	3.4	5.1	13.5	30.5
	MVNet [11]	98.8	0.1	1.1	6.2	26.8	99.0	0.0	7.3	24.8	28.5
	TMVA-Net [11]	99.8	26.0	8.6	30.7	41.3	99.9	41.3	15.9	47.0	51.0
	PeakConv [18]	-	-	-	-	42.9	-	-	-	-	53.3
	TransRadar [2]	99.86	29.88	6.50	35.28	42.88	99.90	46.01	12.20	52.16	52.58
	HyperEcho$_{HG}$	99.86	**31.26**	10.55	**38.74**	**45.10**	99.34	**47.63**	19.09	**55.85**	**55.62**
	HyperEcho$_{CL}$	99.86	22.92	8.11	30.37	40.32	99.93	37.29	15.01	46.59	49.70
	HyperEcho	99.86	26.12	**11.96**	31.25	42.30	99.93	41.43	**23.23**	47.62	52.50

IV. EXPERIMENTAL EVALUATION

A. Dataset, Metrics, and Model Variants

a) Dataset.: We evaluate HyperEcho on the CARRADA dataset [11], which contains synchronized FMCW radar and camera recordings from a roadside automotive scenario. Following the standard protocol, we consider the RD (range–Doppler) and RA (range–angle) radar views and predict dense semantic labels for four classes: background, pedestrian, cyclist, and car. All methods are trained on the official training split and evaluated on the test split, using only radar inputs at test time.

b) Metrics.: Performance is reported using class-wise intersection-over-union (IoU) and Dice scores for each of the four classes, together with their mean values (mIoU and mDice). IoU measures the overlap between predicted and ground-truth regions and is particularly sensitive to false positives, while Dice emphasises overlap for small objects and is more forgiving to boundary noise. Both metrics are computed per class on RD and RA views separately, and then averaged over the foreground classes to obtain mIoU and mDice shown in Table I.

c) HyperEcho variants.: To disentangle the contributions of the hypergraph backbone and the contrastive objective, we evaluate three flavours of our model. HyperEcho$_{HG}$ replaces the TransRadar attention blocks with the axial hypergraph backbone described in Section III-B, but is trained only with the segmentation and multi-view consistency losses. HyperEcho$_{CL}$ keeps a lightweight directional backbone (Section III-B) and augments it with the projection head g_ϕ and an InfoNCE contrastive loss applied to embeddings from two stochastic views of the same radar sample; no hypergraph reasoning is used in this variant. Finally, HyperEcho combines both components: the axial hypergraph backbone is regularised by the same contrastive objective, yielding the full model reported as **HyperEcho** in Table I.

B. Results on CARRADA

Table I summarises semantic segmentation performance on the CARRADA test split. On the RD view, generic CNN baselines (FCN-8s, U-Net, DeepLabv3+) achieve high background IoU but struggle on pedestrians and cyclists, while radar-specific architectures (RSS-Net, RAMP-CNN, MVNet, TMVA-Net, PeakConv) close part of this gap. TransRadar

further improves foreground performance and sets a strong reference, particularly for cars.

Our HyperEcho variants match or surpass this reference while keeping a similar parameter budget. On RD, HyperEcho$_{HG}$ already reaches 58.4 mIoU and 69.68 mDice, demonstrating that replacing pairwise attention by axial hypergraph aggregation is competitive on its own. HyperEcho$_{CL}$ improves rare classes, especially cyclist, and achieves 61.6 mIoU and 73.48 mDice, indicating that aligning embeddings from different views of the same radar sample helps stabilise the learned features. The full model HyperEcho attains the best overall RD scores with 62.42 mIoU and 74.16 mDice, slightly outperforming TransRadar (62.09 mIoU, 73.43 mDice). The largest gains are on pedestrians and cyclists, indicating that the combination of hypergraph reasoning and contrastive regularisation helps the network better separate small, structured foreground echoes from background clutter.

On the RA view, which is substantially more challenging, the benefits of the hypergraph backbone are even clearer. HyperEcho$_{HG}$ achieves 45.10 mIoU and 55.62 mDice, improving over TransRadar by +2.22 mIoU and +3.04 mDice and yielding the best RA performance among all methods. It consistently boosts all foreground classes, with noticeable gains for pedestrians and cars. In contrast, HyperEcho$_{CL}$ alone underperforms on RA, suggesting that contrastive supervision without higher-order aggregation is insufficient in this very sparse, elongated representation. The full HyperEcho model recovers part of this gap and achieves the best cyclist Dice, but the strongest RA performance is obtained when the hypergraph backbone is present.

Overall, these results show that axial hypergraph aggregation is the primary driver of accuracy gains, particularly on the harder RA view and on small foreground classes, while the contrastive objective provides complementary regularisation that is most beneficial on the RD view.

C. Computational Cost and Runtime

We compare the computational footprint of our method against TransRadar in terms of FLOPs, inference throughput (FPS), and peak GPU memory usage. The results in Table II indicate that our approach achieves slightly lower GFLOPs and GPU memory while improving FPS.

TABLE II
COMPUTATIONAL FOOTPRINT. LOWER GFLOPS AND GPU MEMORY ARE BETTER; HIGHER FPS IS BETTER.

Model	GFLOPs	FPS	GPU Memory (GBs)
TransRadar [2]	1138.57	9.6	3.25
HypeEcho	1132.13	10.55	3.18

V. CONCLUSION

HyperEcho introduces axis-aligned hypergraph aggregation as an efficient way to capture the structured echo patterns of automotive radar. By grouping distant reflections, HyperEcho achieves state-of-the-art results on CARRADA. This suggests that explicitly modelling higher–order structure is a valuable direction for future radar perception systems.

REFERENCES

[1] CHEN, L.-C., ZHU, Y., PAPANDREOU, G., SCHROFF, F., AND ADAM, H. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)* (2018), pp. 801–818.

[2] DALBAH, Y., LAHOUD, J., AND 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 (WACV)* (January 2024), pp. 353–362.

[3] GAO, X., XING, G., ROY, S., AND LIU, H. Ramp-cnn: A novel neural network for enhanced automotive radar object recognition. *IEEE Sensors Journal 21*, 4 (2020), 5119–5132.

[4] JIANG, T., ZHUANG, L., AN, Q., WANG, J., XIAO, K., AND WANG, A. T-rodnet: Transformer for vehicular millimeter-wave radar object detection. *IEEE Transactions on Instrumentation and Measurement 72* (2022), 1–12.

[5] KAUL, P., DE MARTINI, D., GADD, M., AND NEWMAN, P. Rss-net: Weakly-supervised multi-class semantic segmentation with fmcw radar. In *2020 IEEE Intelligent Vehicles Symposium (IV)* (2020), IEEE, pp. 431–436.

[6] LIU, Z., LIN, Y., CAO, Y., HU, H., WEI, Y., ZHANG, Z., LIN, S., AND GUO, B. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision* (2021), pp. 10012–10022.

[7] LONG, J., SHELHAMER, E., AND DARRELL, T. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (2015), pp. 3431–3440.

[8] MAJOR, B., FONTIJNE, D., ANSARI, A., TEJA SUKHAVASI, R., GOWAIKAR, R., HAMILTON, M., LEE, S., GRZECHNIK, S., AND SUBRAMANIAN, S. Vehicle detection with automotive radar using deep learning on range-azimuth-doppler tensors. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops* (2019), pp. 0–0.

[9] MILLETARI, F., NAVAB, N., AND AHMADI, S.-A. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 fourth international conference on 3D vision (3DV)* (2016), Ieee, pp. 565–571.

[10] OORD, A. V. D., LI, Y., AND VINYALS, O. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748* (2018).

[11] OUAKNINE, A., NEWSON, A., PÉREZ, P., TUPIN, F., AND REBUT, J. Multi-view radar semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2021), pp. 15671–15680.

[12] OUAKNINE, A., NEWSON, A., REBUT, J., TUPIN, F., AND PÉREZ, P. Carrada dataset: Camera and automotive radar with range- angle-doppler annotations. In *2020 25th International Conference on Pattern Recognition (ICPR)* (2021), pp. 5068–5075.

[13] PATOLE, S. M., TORLAK, M., WANG, D., AND ALI, M. Automotive radars: A review of signal processing techniques. *IEEE Signal Processing Magazine 34*, 2 (2017), 22–35.

[14] REBUT, J., OUAKNINE, A., MALIK, W., AND PÉREZ, P. Raw high-definition radar for multi-task learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2022), pp. 17021–17030.

[15] RONNEBERGER, O., FISCHER, P., AND BROX, T. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention* (2015), Springer, pp. 234–241.

[16] WU, Z., PAN, S., CHEN, F., LONG, G., ZHANG, C., AND YU, P. S. A comprehensive survey on graph neural networks. *IEEE transactions on neural networks and learning systems 32*, 1 (2020), 4–24.

[17] XU, S., WAN, R., YE, M., ZOU, X., AND CAO, T. Sparse cross-scale attention network for efficient lidar panoptic segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence* (2022), vol. 36, pp. 2920–2928.

[18] ZHANG, L., ZHANG, X., ZHANG, Y., GUO, Y., CHEN, Y., HUANG, X., AND MA, Z. Peakconv: Learning peak receptive field for radar semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2023), pp. 17577–17586.

[19] ZHOU, D., HUANG, J., AND SCHÖLKOPF, B. Learning with hypergraphs: Clustering, classification, and embedding. In *Advances in Neural Information Processing Systems* (2006).