SAGE-RSS: STAGE-ALIGNED GLOBAL–LOCAL EVIDENCE REFINEMENT FOR MULTI-VIEW RADAR SEMANTIC SEGMENTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Radar semantic segmentation (RSS) supports robust all-weather perception, yet radar frequency maps exhibit anisotropic structures, weak spectral responses, and temporal dependencies that generic image operators do not address. High-precision RSS must recover sparse target evidence while integrating context across heterogeneous radar projections. We present SAGE-RSS, a stage-aligned global–local evidence refinement framework. A Progressive Support Encoder (PSE) gathers multi-scale support through parallel channel groups and an identity path; a Global–Local Fusion Refiner (GLFR) combines an efficient transform-space branch with residual spatial enhancement; and a Context-Conditioned Axis-aware Decoder (CCAD) performs compact global readout and parallel multi-kernel reconstruction. On CARRADA, SAGE-RSS obtains 63.91% RD mIoU, 46.74% RA mIoU, and 55.33% mean-view mIoU. Under the same profiling protocol, the model runs at 33.30 FPS. The parallel branches and compact context aggregation provide real-time inference and make SAGE-RSS suitable for latency-sensitive and edge-side radar perception applications.

1 INTRODUCTION

Automotive FMCW radar supports perception under low visibility and adverse illumination (Wang et al., 2021). Its echoes are transformed into Range–Doppler (RD), Range–Angle (RA), and Angle–Doppler (AD) projections. RD represents range and radial velocity, RA represents spatial position, and AD links angular direction with Doppler velocity. Radar semantic segmentation (RSS) assigns each frequency-map pixel to background, pedestrian, cyclist, or car (Kaul et al., 2020; Ouaknine et al., 2021a), providing dense scene interpretation from radar alone.

These projections are informative but difficult to segment. Target responses are sparse and may occupy very different spatial extents, while clutter, sidelobes, and multipath surround them. Moreover, range, angle, and Doppler axes have different resolutions and semantics (Richards et al., 2005). The challenge is not simply to enlarge the receptive field. Repeated aggregation can improve semantic context while making an already weak response harder to recover: downsampling removes compact support, cross-view mixing combines evidence with different coordinate semantics, and decoding must reconstruct boundaries from a compressed representation. A multi-view network must therefore accumulate sufficient context without erasing the local evidence that identifies a target.

We formulate RSS as *stage-wise sparse-evidence preservation*. The central insight is that evidence loss occurs in different forms at different stages, and each stage should introduce the necessary context while retaining a low-cost route for local information. Encoding needs scale diversity without processing every channel with an expensive large kernel; fusion needs a global receptive field without replacing spatial detail; and decoding needs output-view adaptation without dense token-to-token attention. This principle also leads to an efficiency-oriented implementation: independent channel groups and view branches are parallelizable, residual paths avoid redundant feature reconstruction, transform operations provide global mixing without quadratic attention, and compact context readout grows linearly with the feature-grid size.

Existing RSS methods employ multi-view fusion, peak-aware convolutions, attention, or explicit temporal relations (Ouaknine et al., 2021a; Dalbah et al., 2024; Zhang et al., 2023; Li et al., 2024;

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

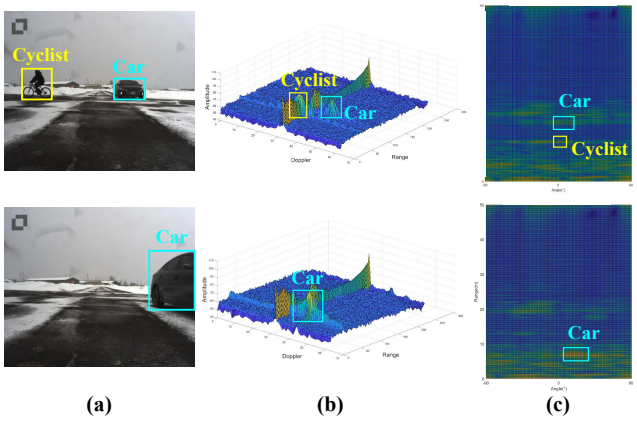

Figure 1: Radar frequency maps. From left to right: (a) synchronized camera images, (b) 3-D visualizations of Range–Doppler (RD) amplitudes, and (c) 2-D Range–Angle (RA) maps. The underlying RAD tensor spans range, angle, and Doppler; its projections contain sparse target responses with view-dependent shapes.

Zhang et al., 2024). These methods demonstrate that radar-aware spatial support, cross-view context, and temporal observations are important, but they also introduce a practical tension between representation strength and inference cost. SAGE-RSS addresses this tension through a stage-aligned encoder–fusion–decoder design. PSE progressively gathers multi-scale support using channel partitioning and an identity route. GLFR couples a global transform branch with a residual local branch. CCAD broadcasts a compact context vector and reconstructs RD and RA masks through parallel square and asymmetric kernels. The resulting architecture remains convolution dominated and exposes parallel execution across views, channel groups, and module branches. The main contributions are summarized as follows:

- We formulate multi-view RSS as a stage-wise sparse-evidence preservation problem, identifying scale-dependent evidence loss in encoding, global–local imbalance in fusion, and view-dependent reconstruction in decoding.
- We develop SAGE-RSS with three stage-aligned components: PSE for efficient progressive support encoding, GLFR for parallel global–local refinement, and CCAD for compact context conditioning and axis-aware reconstruction. Their residual, grouped, transform-based, and linear-complexity operations are selected jointly for accuracy and inference efficiency.
- We evaluate SAGE-RSS on CARRADA and obtain 63.91% RD mIoU, 46.74% RA mIoU, and 55.33% mean-view mIoU. Under the same single-RTX-3090 setting used for the efficiency comparison, SAGE-RSS reaches 33.30 FPS, supporting real-time radar perception and deployment on latency-sensitive edge platforms.

2 Related Work

2.1 Multi-view Radar Semantic Segmentation

RSS-Net (Kaul et al., 2020) introduced weakly supervised multi-class segmentation for FMCW radar, and RAMP-CNN (Gao et al., 2021) combined multi-receptive-field aggregation with temporal convolution. TMVA-Net (Ouaknine et al., 2021a) established a parallel RD/RA/AD architecture with cross-view coherence supervision. Subsequent methods introduce different radar-oriented priors: TransRSS (Zou et al., 2023) and TransRadar (Dalbah et al., 2024) use attention for global context, PeakConv (Zhang et al., 2023) and AdaPKC-Net (Li et al., 2024) adapt receptive fields to radar peaks, and TARSS-Net (Zhang et al., 2024) models target-history relations. SAGE-RSS complements these directions by organizing multi-scale support, global–local fusion, and axis-aware reconstruction along the encoder–fusion–decoder pipeline while retaining efficient parallel computation.

2.2 Frequency-Domain Feature Learning

Frequency-domain operators provide global receptive fields and model structures that may be difficult to capture with local spatial filtering alone. FFC (Chi et al., 2020) performs convolution in the Fourier domain, GFNet (Rao et al., 2021) learns global frequency filters, and AFNO (Guibas et al., 2022) introduces adaptive Fourier token mixing. Related studies further demonstrate the value of spectral representations in image reconstruction and dense prediction (Suvorov et al., 2022; Jiang et al., 2021; Fujieda et al., 2018; Li et al., 2023). For RSS, frequency-domain learning is particularly relevant because Doppler represents radial velocity in the frequency domain, with moving targets typically forming sparse peaks that differ from clutter and other interference.

2.3 Large-Kernel and Multi-Scale Convolutions

Large-kernel convolution expands the effective receptive field and improves context modeling. RepLKNet (Ding et al., 2022), SLaK (Liu et al., 2023), and UniRepLKNet (Ding et al., 2024) show that large-kernel ConvNets can be competitive with attention-based architectures while avoiding the quadratic cost of dense attention. UniConvNet (Wang & Xi, 2025) proposes progressive receptive field aggregation with moderate kernels, providing a compromise between local detail and broad context. For radar maps, broader context helps separate true targets from sidelobes, multipath artifacts, and fragmented reflections. Inspired by these developments, SAGE-RSS combines progressive kernels in PSE with parallel square and asymmetric paths in CCAD to capture multi-scale and axis-dependent radar structures.

3 Method

3.1 Overall RSS Framework

Built on a multi-view encoder–decoder architecture(Ouaknine et al., 2021a), SAGE-RSS receives five-frame RD, RA, and AD inputs and predicts semantic masks for RD and RA (Fig. 2). Let the three sequences be $\mathcal{X}^v \in \mathbb{R}^{B\times 1\times T\times H_v\times W_v}$ for $v \in \{RD, RA, AD\}$. View-specific 3-D encoders aggregate the input sequence and map each view to a two-dimensional feature hierarchy. PSE operates before cross-view alignment to collect multi-scale target support within each projection. Features sharing the target output axes are then aligned, concatenated, and projected into $\mathbf{Z}^{RD}$ and $\mathbf{Z}^{RA}$. GLFR refines each aligned representation by combining global transform-space interaction with local spatial preservation. CCAD receives the refined features together with encoder skip information and reconstructs the final RD or RA logits.

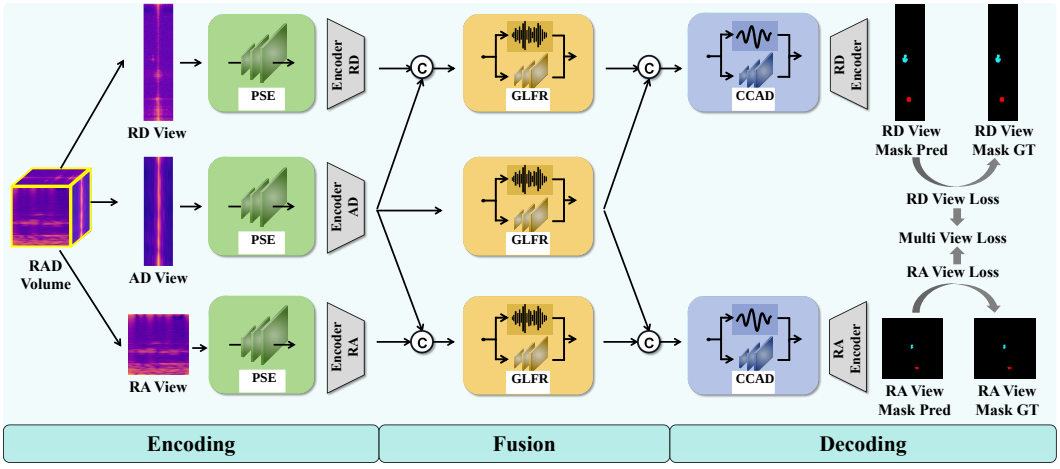

Figure 2: Overview of SAGE-RSS. Five-frame RD, RA, and AD inputs are processed by parallel encoders. PSE gathers multi-scale support while retaining an identity group; GLFR balances transform-space context with a residual spatial path; and CCAD conditions axis-aware RD and RA reconstruction on compact context extracted after sequence aggregation.

The design follows one insight: the context required by RSS changes with the representation stage. Encoding must recognize targets with different spatial extents, fusion must integrate distant and cross-view evidence without removing local peaks, and decoding must recover output-view structures from compact features. SAGE-RSS assigns one module to each responsibility and implements them with efficient operators. View branches, channel groups, and dual-path refiners can execute in parallel; residual connections reuse existing features; transform operations provide global interaction without quadratic attention; and compact context broadcasting keeps decoder complexity linear in the feature-grid size. This stage-aligned design jointly improves representation quality and preserves real-time inference.

3.2 PSE: Progressive Support Encoder

PSE is motivated by the observation that radar targets occupy markedly different support sizes: compact pedestrian or cyclist responses require fine local evidence, whereas vehicle responses and fragmented echoes benefit from wider context. Applying a large kernel to every channel is unnecessary and expensive. PSE instead partitions channels into four groups, processes three groups with progressively enlarged receptive fields, and reserves one identity group. This design aligns the encoder's function: progressive branches accumulate support across scales, while the identity route preserves high-frequency local evidence (Wang & Xi, 2025). Because the groups are independent at each stage and operate on channel subsets, they reduce the cost of uniformly applying large kernels and are amenable to parallel execution. Given $\mathbf{X} \in \mathbb{R}^{B\times C\times H\times W}$, the progressive aggregation is

$$\begin{aligned}\mathbf{O}_1 &= \mathrm{DWConv}_{7\times7}(\mathbf{X}_1) \odot \mathrm{SE}(\mathbf{X}_1),\\ \mathbf{O}_2 &= \mathrm{Conv}_{9\times9}([\mathbf{O}_1, \mathbf{X}_2]) \odot \mathrm{SE}([\mathbf{O}_1, \mathbf{X}_2]),\\ \mathbf{O}_3 &= \mathrm{Conv}_{11\times11}([\mathbf{O}_1, \mathbf{X}_2, \mathbf{X}_3]) \odot \mathrm{SE}([\mathbf{O}_1, \mathbf{X}_2, \mathbf{X}_3]).\end{aligned} \tag{1}$$

The channel gate and residual output are

$$\mathrm{SE}(\mathbf{Z}) = \sigma\big(\mathrm{Proj}(\mathrm{AvgPool}(\mathbf{Z}))\big), \tag{2}$$

$$\mathbf{Y}_{\mathrm{PSE}} = \mathrm{Concat}[\mathbf{O}_1, \mathbf{O}_2, \mathbf{O}_3, \mathbf{X}_4] + \mathbf{X}. \tag{3}$$

The progressive order enables broader branches to reuse features produced at smaller scales instead of extracting every scale from the full input independently. SE-based channel recalibration selects informative responses before concatenation, and the identity group plus the outer residual retains a direct information path. Together, these choices let PSE enlarge effective receptive fields with a compact grouped structure, matching its role of efficient multi-scale evidence extraction in the encoder.

3.3 GLFR: Global–Local Fusion Refiner

GLFR is designed for the fusion stage, where aligned multi-view features require both long-range interaction and local peak preservation. Dense attention offers global context but introduces quadratic token interaction and is less suitable for real-time RSS (Yang et al., 2025). GLFR therefore uses two parallel paths: a transform branch that mixes information over the complete latent grid with efficient FFT/IFFT operators, and a spatial residual branch that retains local target morphology. The two paths are independent until the final 1×1 fusion, enabling parallel execution. For $\mathbf{X} \in \mathbb{R}^{B\times C\times H\times W}$, the global branch is

$$\begin{aligned}\tilde{\mathbf{X}} &= \mathcal{F}(\mathbf{X}) \odot \mathbf{W}_c,\\ \mathbf{Y}_{\mathrm{global}} &= \phi_{\mathrm{freq}}\big(\mathcal{F}^{-1}(\tilde{\mathbf{X}})\big),\end{aligned} \tag{4}$$

where $\mathbf{W}_c \in \mathbb{C}^{1\times C\times1\times1}$ is a learnable complex channel weight broadcast over the Fourier grid and ϕ_{freq} is a 1×1 projection followed by normalization and activation. The FFT gives each output position access to the complete latent feature grid, while the compact channel-wise weight keeps the parameter and multiplication overhead low. The inverse transform returns a spatial feature map that can be fused directly with the local path. Thus, the branch supplies global transform-space calibration using efficient structured operators rather than dense pairwise attention.

The local branch constructs a multi-dilation gate and a detail feature:

$$\begin{aligned}\mathbf{M} &= \mathrm{Conv}_{1\times1}\big(\mathrm{Concat}[\mathrm{DWConv}_{d=1,3,5}(\mathbf{X})]\big),\\ \mathbf{D} &= \mathrm{Conv}_{3\times3}\big(\mathrm{SiLU}(\mathrm{Conv}_{1\times1}(\mathbf{X}))\big),\\ \mathbf{Y}_{\mathrm{local}} &= \mathbf{X} + \mathbf{D} \odot \sigma(\mathbf{M}).\end{aligned} \tag{5}$$

The two paths are fused as

$$\mathbf{Y}_{\mathrm{GLFR}} = \phi_{\mathrm{fuse}}\big(\mathrm{Concat}[\mathbf{Y}_{\mathrm{global}}, \mathbf{Y}_{\mathrm{local}}]\big) + \mathbf{X}, \tag{6}$$

where ϕ_{fuse} denotes 1×1 projection, normalization, and activation. The transform path contributes global context, whereas the gated dilated depthwise convolutions preserve local responses at multiple spatial spreads. Their parallel fusion realizes the global–local design insight with FFT, depthwise convolution, and pointwise projection, all of which support efficient GPU execution.

3.4 CCAD: Context-Conditioned Axis-aware Decoder

CCAD addresses the decoder-stage need for globally conditioned but view-sensitive reconstruction. After the view-specific 3-D encoder aggregates five input frames, the decoder receives a compact 2-D feature map. Rather than applying dense self-attention over all $L = H \times W$ positions, CCAD forms a small set of N_m context anchors, reads them with input-adaptive weights, and broadcasts one global context vector to every position. A parallel axis-aware branch then reconstructs square and directional structures with separable and depthwise convolutions. The two branches can execute concurrently, and the context branch has linear complexity in L for fixed N_m. Let $\mathcal{X} \in \mathbb{R}^{B\times1\times T\times H_0\times W_0}$, with $T = 5$. The encoder and decoder projection produce

$$\begin{aligned} \mathbf{X} &= \mathcal{P}_{\mathrm{dec}}\big(\mathcal{E}_{3D}(\mathcal{X})\big) \in \mathbb{R}^{B\times C\times H\times W}, \\ \mathbf{X}_f &= \mathrm{Flatten}(\mathbf{X}) \in \mathbb{R}^{B\times L\times C}, \qquad L = H \times W. \end{aligned} \tag{7}$$

The context branch predicts N_m weights from global average features, projects N_m anchor tokens, and broadcasts their weighted readout:

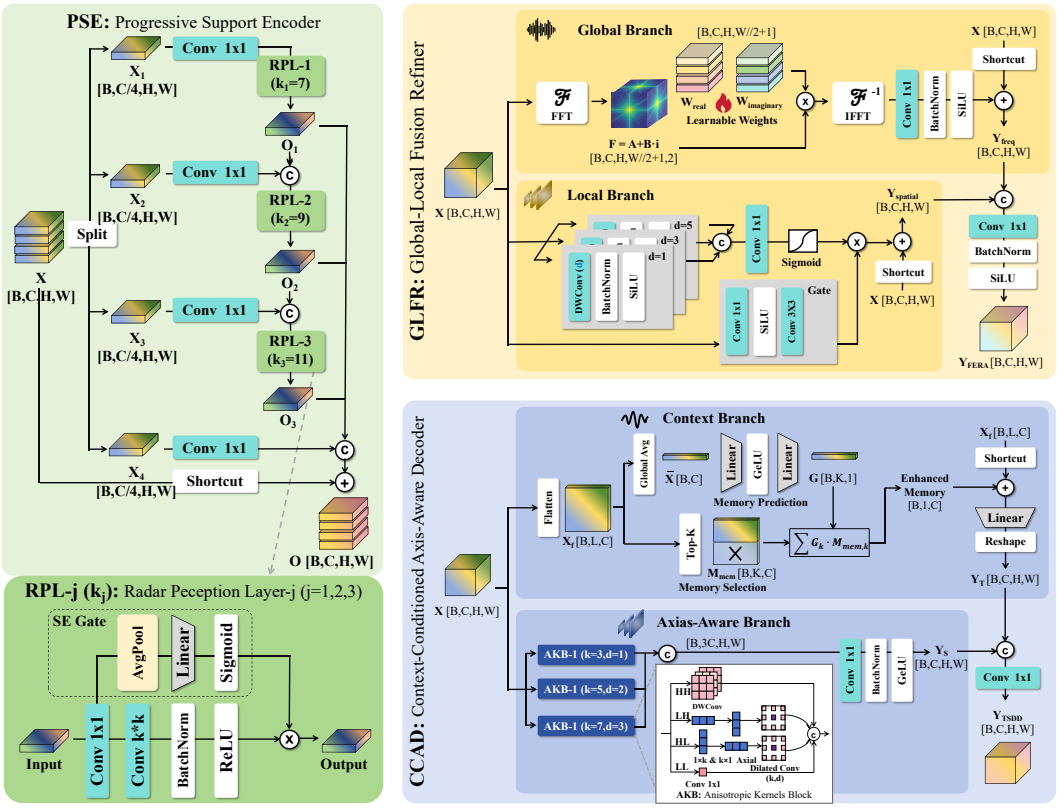

Figure 3: The architecture diagrams of proposed modules: (1) PSE splits channels among progressive receptive fields and an identity group. (2) GLFR combines transform-space channel calibration with residual spatial enhancement. (3) CCAD combines a compact context branch with multi-kernel axis-aware reconstruction.

$$
\begin{aligned}
\boldsymbol{\alpha} &= \mathrm{Softmax}\left(\mathrm{WeightProj}\left(\frac{1}{L}\sum_{l=1}^{L}\mathbf{X}_{f,l}\right)\right),\\
\mathbf{M}_{\mathrm{ctx}} &= \mathrm{MemProj}(\mathbf{X}_f[:,:N_m,:]),\\
\mathbf{E} &= \sum_{m=1}^{N_m}\alpha_m\mathbf{M}_{\mathrm{ctx},m},\\
\mathbf{Y}_C &= \mathrm{reshape}\left(\mathrm{OutProj}(\mathbf{X}_f + \mathbf{1}_L\mathbf{E})\right).
\end{aligned}
\tag{8}
$$

The global mean feature predicts the importance of the N_m compact anchors, and their weighted sum provides an input-conditioned scene descriptor. Broadcasting this descriptor avoids constructing an $L \times L$ attention matrix, while the pointwise output projection combines global context with every local token. The context branch therefore supplies a compact global prior to mask reconstruction with memory and computation that grow linearly with the feature-grid size.

The axis-aware branch uses three multi-kernel blocks with $k \in \{3,5,7\}$ and square-path dilation $d_k \in \{1,2,3\}$:

$$
\begin{aligned}
\mathbf{F}_k &= \mathrm{Conv}_{1\times1}\big(\mathrm{Concat}[LL(\mathbf{X}), LH(\mathbf{X}), HL(\mathbf{X}), HH(\mathbf{X})]\big) + \mathrm{Proj}(\mathbf{X}),\\
LL(\mathbf{X}) &= \mathrm{Conv}_{1\times1}(\mathbf{X}),\\
LH(\mathbf{X}) &= \mathrm{DilConv}_{3\times3,d=5}\big(\mathrm{Conv}_{k\times1}(\mathrm{Conv}_{1\times k}(\mathbf{X}))\big),\\
HL(\mathbf{X}) &= \mathrm{DilConv}_{3\times3,d=5}\big(\mathrm{Conv}_{1\times k}(\mathrm{Conv}_{k\times1}(\mathbf{X}))\big),\\
HH(\mathbf{X}) &= \mathrm{DWConv}_{k\times k,d_k}(\mathbf{X}).
\end{aligned}
\tag{9}
$$

The branch output and the final fusion are

$$
\mathbf{Y}_A = \mathrm{GELU}\big(\mathrm{BN}(\mathrm{Conv}_{1\times1}(\mathrm{Concat}[\mathbf{F}_3,\mathbf{F}_5,\mathbf{F}_7]))\big) + \mathbf{X}, \tag{10}
$$

$$
\mathbf{Y}_{\mathrm{CCAD}} = \mathrm{Conv}_{1\times1}(\mathrm{Concat}[\mathbf{Y}_C,\mathbf{Y}_A]). \tag{11}
$$

The square path captures compact local support, while the two ordered asymmetric paths aggregate horizontal and vertical structures before the shared dilated convolution. Running the $k \in \{3,5,7\}$ blocks in parallel supplies multiple reconstruction scales without recurrent processing or dense attention. Their outputs are compressed by a 1×1 projection and fused with the compact context branch, aligning the decoder's function with the design of efficient view-sensitive mask reconstruction (Xie et al., 2025).

3.5 Multi-view RSS Loss

SAGE-RSS produces semantic segmentation masks from RD and RA views. The learning objective must handle both class imbalance and cross-view coherence, since foreground radar targets are sparse and the RD/RA projections describe the same physical objects from different dimensions. Combining region-level supervision with view-consistency regularization encourages the model to learn accurate masks in each view while maintaining physically compatible predictions. These predictions are compared with their corresponding ground truth using a multi-task loss. Following the common RSS setting (Ouaknine et al., 2021a), the loss combines weighted cross-entropy, soft Dice loss, and cross-view consistency:

$$
\mathcal{L}_{\mathrm{wCE}}(\mathbf{y},\mathbf{p}) = -\frac{1}{K}\sum_{k=1}^{K} w_k \sum_{(m,n)\in\Omega}\mathbf{y}[m,n,k]\log\mathbf{p}[m,n,k], \tag{12}
$$

$$
\mathcal{L}_{\mathrm{SDice}} = \frac{1}{K}\sum_{k=1}^{K}\left[1 - \frac{2\sum_{(m,n)}\mathbf{y}[m,n,k]\mathbf{p}[m,n,k]}{\sum_{(m,n)}\mathbf{y}^2[m,n,k] + \mathbf{p}^2[m,n,k]}\right], \tag{13}
$$

$$
\mathcal{L}_{\mathrm{CoL}}(\mathbf{p}^{RD},\mathbf{p}^{RA}) = \|\hat{\mathbf{p}}^{RD} - \hat{\mathbf{p}}^{RA}\|_F^2. \tag{14}
$$

$$
\begin{aligned}
\mathcal{L} =& \lambda_{\mathrm{wCE}}\left(\mathcal{L}_{\mathrm{wCE}}^{RD} + \mathcal{L}_{\mathrm{wCE}}^{RA}\right) + \lambda_{\mathrm{SDice}}\left(\mathcal{L}_{\mathrm{SDice}}^{RD} + \mathcal{L}_{\mathrm{SDice}}^{RA}\right)\\
&+ \lambda_{\mathrm{CoL}}\mathcal{L}_{\mathrm{CoL}}.
\end{aligned}
\tag{15}
$$

Table 1: Comparison on CARRADA. Mean-view metrics are the arithmetic means of RD and RA. Best results are **bold**; second-best results are underlined.

Framework	RD		RA		Mean view		Params @Frames
	mIoU	mDice	mIoU	mDice	mIoU	mDice	
FCN (Long et al., 2015)	54.70%	66.30%	34.50%	40.90%	44.60%	53.60%	134.3M@3
U-Net (Ronneberger et al., 2015)	55.40%	68.00%	32.80%	38.20%	44.10%	53.10%	17.3M@3
DeepLabv3+ (Chen et al., 2018)	50.80%	61.60%	32.70%	38.30%	41.75%	49.95%	59.3M@3
RSS-Net (Kaul et al., 2020)	32.10%	36.90%	32.10%	37.80%	32.10%	37.35%	10.1M@3
RAMP-CNN (Gao et al., 2021)	56.60%	68.50%	27.90%	30.50%	42.25%	49.50%	106.4M@9
TMVA-Net (Ouaknine et al., 2021a)	56.10%	68.00%	37.70%	46.20%	46.90%	57.10%	5.6M@5
TransRadar (Dalbah et al., 2024)	57.20%	69.10%	39.90%	49.50%	48.55%	59.30%	4.9M@5
T-RODNet (Jiang et al., 2022)	–	–	43.50%	53.60%	–	–	162.0M@4
PKCIn-Net (Zhang et al., 2023)	60.70%	72.60%	43.10%	53.70%	51.90%	63.15%	6.3M@5
AdaPKC-Net0 (Li et al., 2024)	61.20%	73.10%	44.10%	55.10%	52.65%	64.10%	6.3M@5
AdaPKC-Net1 (Li et al., 2024)	61.50%	73.60%	43.60%	54.50%	52.55%	64.05%	6.3M@5
AdaPKC-Net2 (Li et al., 2024)	62.10%	74.00%	44.30%	55.50%	53.20%	64.75%	6.3M@5
SAGE-RSS (ours)	**63.91**%	**75.72**%	**46.74**%	**58.22**%	**55.33**%	**66.97**%	14.5M@5

Here, $\mathbf{y}$ is the one-hot ground truth tensor, $\mathbf{p}$ is the predicted probability tensor, K is the number of classes, and w_k is the class-balancing weight. $\hat{\mathbf{p}}^{RD}$ and $\hat{\mathbf{p}}^{RA}$ are aggregated maps obtained along the non-shared Doppler or angle dimension. This objective encourages view-specific learning while improving cross-view consistency.

4 Experiments

4.1 Dataset and Metrics

The proposed SAGE-RSS is evaluated on CARRADA (Ouaknine et al., 2021b) and CARRADA-RAC (Zhang et al., 2023). CARRADA contains synchronized camera and FMCW radar data with four classes: pedestrian, cyclist, car, and background. Each sample provides a $256 \times 256 \times 64$ range-angle-Doppler tensor, from which RA and RD maps are derived for segmentation; CARRADA-RAC refines the annotations for improved consistency. Following prior work (Hazra & Santra, 2019; Zhang et al., 2023), we report mIoU, which measures class-wise region agreement, and mDice, which emphasizes foreground overlap, on the RD and RA views. Each Global View metric is the arithmetic average of its RD and RA values.

4.2 Implementation Details

To capture temporal information, MuDoRSS takes five consecutive radar frames as input, following the setting in PeakConv (Zhang et al., 2023). The RA view resolution is 256×256, while AD and RD views use 256×64. The model is implemented in PyTorch and trained on two NVIDIA RTX 3090 GPUs with the Adam optimizer (Kingma & Ba, 2014). The initial learning rate is 0.0001 and decays exponentially during training. Training is conducted for 300 epochs with a batch size of 3. The number of memory slots in TSDD is set to $N_m = 8$. The overall loss is the combination of weighted cross-entropy, Dice loss, and cross-view consistency described in the Loss Section.

4.3 Main Results

Table 1 shows that SAGE-RSS obtains 63.91% RD mIoU and 46.74% RA mIoU on CARRADA. Relative to AdaPKC-Net2, the differences are 1.81 and 2.44 percentage points, yielding a 2.13-point improvement in mean-view mIoU. SAGE-RSS also improves mean-view mDice by 2.22 points. The simultaneous gains on RD and RA indicate that the stage-aligned refinement supports both velocity-related and spatial target representations rather than favoring only one output view.

CARRADA-RAC (Zhang et al., 2023) provides revised annotations for the CARRADA radar sequences. Table 2 reports RD and RA results under this annotation setting. SAGE-RSS reaches

63.5% RD mIoU, 75.4% RD mDice, 50.4% RA mIoU, and 62.4% RA mDice. The corresponding arithmetic means are 57.0% mIoU and 68.9% mDice, exceeding the strongest listed prior mean-view scores by 2.2 and 2.3 percentage points, respectively. These results show that the accuracy advantage remains consistent under the refined labels.

Figure 4 provides a qualitative car example. The SAGE-RSS prediction is better aligned with the displayed ground truth and avoids the fragmented or over-expanded regions in several compared outputs. The result is consistent with the overall design: the encoder collects support over multiple extents, the fusion stage combines global and local evidence, and the decoder reconstructs the target with view-sensitive kernels.

Table 2: Comparison on CARRADA-RAC. Best results are **bold**; second-best results are underlined.

Framework	RD		RA	
	mIoU	mDice	mIoU	mDice
TMVA-Net	59.7%	69.9%	46.6%	57.9%
PKCIn-Net	60.6%	72.4%	47.3%	58.7%
AdaPKC$^{\theta}$-Net	60.7%	72.7%	48.1%	59.6%
AdaPKC$^{\xi}$-Net	61.6%	73.6%	47.9%	59.3%
AdaPKC$^{\xi}$-Net^{FiT}	60.8%	72.7%	48.8%	60.5%
SAGE-RSS (ours)	**63.5%**	**75.4%**	**50.4%**	**62.4%**

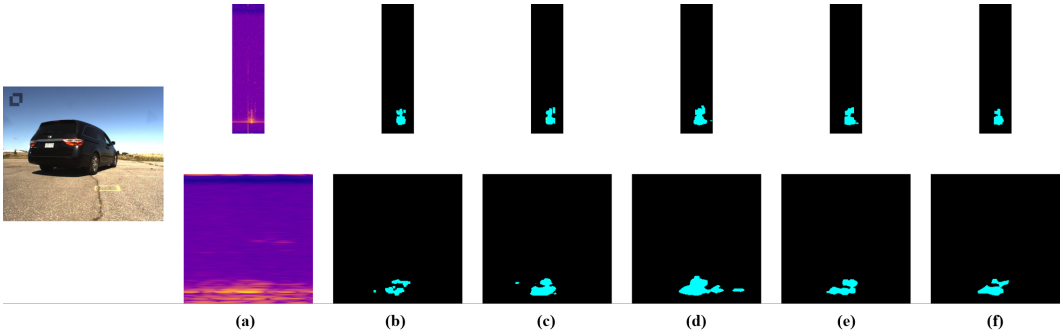

Figure 4: Qualitative radar segmentation results. The leftmost column shows the synchronized RGB camera image. The upper and lower rows depict the RD and RA views, respectively. In each row, (a) original RD/RA map; (b) ground-truth segmentation mask; (c) TransRadar result; (d) TMVA-Net result; (e) AdaPKC$^{\xi}$-Net^{FiT} result; and (f) SAGE-RSS (ours) result. Cyan denotes the car class.

4.4 Ablation Study

Table 3 evaluates the three stage-aligned components. PSE raises mean-view mIoU from 51.10% to 51.94%, with a larger change on RA, which is consistent with its multi-scale spatial-support function. GLFR produces the largest individual gain and reaches 53.18% mean-view mIoU, supporting the use of parallel global and local refinement at the fusion stage. CCAD improves RD mIoU to 61.92% and raises RA mDice to 52.01%, showing the value of compact context conditioning and multi-kernel reconstruction. PSE+CCAD reaches 45.46% RA mIoU, while the full model obtains the best value on every reported metric. These results show that applying the three operations at their corresponding representation stages provides the strongest overall RD/RA performance.

4.5 Efficiency and Deployment

We profile all models under the same experimental setting following PeakConv (Zhang et al., 2023), using one NVIDIA RTX 3090. Each throughput result is averaged over no fewer than three inference runs, with five runs used where available. As shown in Table 4, SAGE-RSS processes 33.30

Table 3: Component ablation on CARRADA.

PSE	GLFR	CCAD	RD		RA	
			mIoU	mDice	mIoU	mDice
			61.16%	72.91%	41.04%	50.49%
✓			61.39%	73.35%	42.48%	52.48%
	✓		61.88%	73.79%	44.48%	55.13%
		✓	61.92%	73.71%	41.78%	52.01%
✓	✓		62.17%	73.98%	43.58%	53.77%
✓		✓	61.76%	73.55%	45.46%	56.34%
✓	✓	✓	**63.91**%	**75.72**%	**46.74**%	**58.22**%

Table 4: Efficiency comparison under the same RSS setting on one NVIDIA RTX 3090. Reported throughput is averaged over 5 repeated inference runs.

Model	Params (M)	MACs (G)	FPS
TMVA-Net	5.63	102.31	68.23
AdaPKC$^{\theta}$-Net	6.32	109.83	6.67
AdaPKC$^{\xi}$-Net	6.32	110.81	6.65
SAGE-RSS	14.56	193.17	33.30

frames per second, about five times the throughput of the two AdaPKC variants, while retaining the highest segmentation accuracy in Table 1. Although SAGE-RSS uses 14.56M parameters and 193.17G MACs, its grouped encoder paths, parallel global–local fusion, compact context readout, and convolution-dominated decoder map efficiently to GPU execution. The measured throughput exceeds the frame rate required by common automotive radar streams, enabling online processing with per-frame latency of approximately 30 ms. This accuracy–throughput balance supports real-time deployment and provides a practical basis for further compression on embedded or edge accelerators.

5 Conclusion

SAGE-RSS formulates multi-view radar semantic segmentation as stage-wise sparse-evidence preservation. PSE uses grouped progressive receptive fields and an identity route for efficient multi-scale encoding; GLFR combines parallel transform-space and spatial paths for global–local fusion; and CCAD couples compact context conditioning with axis-aware multi-kernel reconstruction. The model achieves 55.33% mean-view mIoU on CARRADA. Under the same single-RTX-3090 profiling setting, SAGE-RSS reaches 33.30 FPS, substantially outperforming the measured AdaPKC variants in throughput while delivering higher segmentation accuracy. The parallel branches, compact context aggregation, and convolution-dominated operators make SAGE-RSS suitable for real-time radar perception and provide a practical architecture for further deployment on embedded and edge-side platforms.

6 AI USE STATEMENT

We used AI-based tools to assist with language editing, LATEX typesetting. All research ideas, methods, experiments, analyses, and conclusions are the authors' own, and every AI-assisted passage was checked by the authors

References

Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of*

the European conference on computer vision, pp. 801–818, 2018.

Lu Chi, Borui Jiang, and Yadong Mu. Fast fourier convolution. In *Advances in Neural Information Processing Systems*, volume 33, pp. 4479–4488, 2020.

Yahia Dalbah, Jean Lahoud, and Hisham Cholakkal. 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.

Xiaohan Ding, Xiangyu Zhang, Jungong Han, and Guiguang Ding. Scaling up your kernels to 31x31: Revisiting large kernel design in cnns. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11963–11975, 2022.

Xiaohan Ding, Yiyuan Zhang, Yixiao Ge, Sijie Zhao, Lin Song, Xiangyu Yue, and Ying Shan. Unireplknet: A universal perception large-kernel convnet for audio, video, point cloud, time-series and image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5513–5524, 2024.

Shin Fujieda, Kohei Takayama, and Toshiya Hachisuka. Wavelet convolutional neural networks. *arXiv preprint arXiv:1805.08620*, 2018.

Xiangyu Gao, Guanbin Xing, Sumit Roy, and Hui Liu. Ramp-cnn: A novel neural network for enhanced automotive radar object recognition. *IEEE Sensors Journal*, 21(4):5119–5132, 2021. doi: 10.1109/JSEN.2020.3036047.

John Guibas, Morteza Mardani, Zongyi Li, Andrew Tao, Anima Anandkumar, and Bryan Catanzaro. Efficient token mixing for transformers via adaptive fourier neural operators. In *International Conference on Learning Representations*, 2022.

Souvik Hazra and Avik Santra. Radar gesture recognition system in presence of interference using self-attention neural network. In *2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA)*, pp. 1409–1414. IEEE, 2019.

Liming Jiang, Bo Dai, Wayne Wu, and Chen Change Loy. Focal frequency loss for image reconstruction and synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 13919–13929, 2021.

Tiezhen Jiang, Long Zhuang, Qi An, Jianhua Wang, Kai Xiao, and Anqi Wang. T-rodnet: Transformer for vehicular millimeter-wave radar object detection. *IEEE Transactions on Instrumentation and Measurement*, 72:1–12, 2022.

Prannay Kaul, Daniele de Martini, Matthew Gadd, and Paul Newman. Rss-net: Weakly-supervised multi-class semantic segmentation with fmcw radar. In *2020 IEEE Intelligent Vehicles Symposium (IV)*, pp. 431–436, 2020. doi: 10.1109/IV47402.2020.9304674.

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

Teng Li, Liwen Zhang, Youcheng Zhang, Pengcheng Pi, Zongqing Lu, Qingmin Liao, and Zhe Ma. Adapkc: Peakconv with adaptive peak receptive field for radar semantic segmentation. *Advances in Neural Information Processing Systems*, 37:136545–136575, 2024.

Yifan Li, Ziqian Liu, Junli Yang, and Haopeng Zhang. Wavelet transform feature enhancement for semantic segmentation of remote sensing images. *Remote Sensing*, 15(24):5644, 2023.

Shiwei Liu, Tianlong Chen, Xiaohan Chen, Xuxi Chen, Qiao Xiao, Boqian Wu, Tommi Kärkkäinen, Mykola Pechenizkiy, Decebal C Mocanu, and Zhangyang Wang. More convnets in the 2020s: Scaling up kernels beyond 51x51 using sparsity. In *International Conference on Learning Representations*, 2023.

Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3431–3440, 2015.

Arthur Ouaknine, Alasdair Newson, Patrick Pérez, Florence Tupin, and Julien Rebut. Multi-view radar semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15671–15680, 2021a.

Arthur Ouaknine, Alasdair Newson, Julien Rebut, Florence Tupin, and Patrick Pérez. Carrada dataset: Camera and automotive radar with range-angle-doppler annotations. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pp. 5068–5075. IEEE, 2021b.

Yongming Rao, Wenliang Zhao, Zheng Zhu, Jiwen Lu, and Jie Zhou. Global filter networks for image classification. In *Advances in Neural Information Processing Systems*, volume 34, pp. 980–993, 2021.

Mark A Richards et al. *Fundamentals of radar signal processing*, volume 1. Mcgraw-hill New York, 2005.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241. Springer, 2015.

Roman Suvorov, Elizaveta Logacheva, Anton Mashikhin, Anastasia Remizova, Arsenii Ashukha, Aleksei Silvestrov, Naejin Kong, Harshith Goka, Kiwoong Park, and Victor Lempitsky. Resolution-robust large mask inpainting with fourier convolutions. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 2149–2159, 2022.

Yizhou Wang, Zhongyu Jiang, Yudong Li, Jenq-Neng Hwang, Guanbin Xing, and Hui Liu. Rodnet: A real-time radar object detection network cross-supervised by camera-radar fused object 3d localization. *IEEE Journal of Selected Topics in Signal Processing*, 15(4):954–967, 2021.

Yuhao Wang and Wei Xi. Uniconvnet: Expanding effective receptive field while maintaining asymptotically gaussian distribution for convnets of any scale. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 20922–20933, 2025.

Fei Xie, Jiahao Nie, Yujin Tang, Wenkang Zhang, and Hongshen Zhao. Mamba-adaptor: State space model adaptor for visual recognition. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 20124–20134, 2025.

Yue Yang, Xiaoyi Feng, Ming Li, Xiangyun Hu, Jiangying Qin, Armin Gruen, Deren Li, and Jianya Gong. Umis-yolo: Underwater multimodal images instance segmentation with yolo. *IEEE Transactions on Geoscience and Remote Sensing*, 2025.

Liwen Zhang, Xinyan Zhang, Youcheng Zhang, Yufei Guo, Yuanpei Chen, Xuhui Huang, and Zhe Ma. 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.

Youcheng Zhang, Liwen Zhang, Pengcheng Pi, Teng Li, Yuanpei Chen, Shi Peng, and Zhe Ma. Tarss-net: Temporal-aware radar semantic segmentation network. *Advances in Neural Information Processing Systems*, 37:4906–4933, 2024.

Hao Zou, Zhen Xie, Jiarong Ou, and Yutao Gao. Transrss: Transformer-based radar semantic segmentation. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 6965–6972, 2023. doi: 10.1109/ICRA48891.2023.10161200.