2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

MotionNet-PGA: MotionNet with Polar-Guided Attention for Moving Object Segmentation in Scanning Radar

Ren-Yi Yuan, Chieh-Chih Wang and Wen-Chieh Lin

***Abstract*— Moving object segmentation (MOS) is essential for autonomous driving, enabling robust detection, tracking, and prediction of dynamic agents in complex traffic scenarios. Radar sensors offer notable advantages for long-range sensing, but their lower spatial resolution, measurement noise, and geometric distortions—particularly for distant targets—pose significant challenges for accurate MOS. These limitations are amplified when detecting small objects such as scooters. In this work, we present MotionNet-PGA, a Polar-guided Attention Framework designed specifically for scanning radar-based MOS. Our method builds on the multi-frame motion encoding backbone of MotionNet [1], and introduces a polar-guided attention module to suppress clutter, enhance motion feature representation, and improve segmentation of small and distant targets. For evaluation, we construct and annotate the ITRI Radar moving object segmentation Dataset. Experimental results demonstrate that our method surpasses state-of-the-art baseline, MotionNet, by 2.48% in overall IoU and achieves a 4.08% improvement in small-object segmentation. These results highlight the effectiveness of polar-guided attention in addressing scanning radar-specific challenges.**

I. Introduction

Accurate separation of moving objects from dynamic environments is fundamental to the safety and reliability of autonomous driving and robotic systems. The goal of the Moving Object Segmentation (MOS) is to distinguish dynamic targets from static entities and background in sensor data, providing a motion-aware scene representation for downstream applications.

In this work, we specifically study MOS in scanning radar, where the task is to separate the pixels of moving objects from the static entities and background (e.g., parked vehicles and roadside infrastructure). Knowledge of dynamic entities further benefits a wide range of downstream tasks, such as path planning [2], obstacle avoidance [3], localization [4], and long-term map consistency [5], [6].

Moving object segmentation research has so far progressed mainly in camera and LiDAR domains. Camera-based models [7], [8] rely heavily on high-resolution textures, which radar inherently lacks. LiDAR-based methods [9], [10] have explored projecting inputs into the polar domain to mitigate spatial sparsity, but most still apply convolutional or attention mechanisms in Cartesian space, neglecting radar-specific distortions such as angular compression and clutter-induced noise. Furthermore, studies on temporal feature aggregation in video detection [11], [12] and tiny moving object detection [13], [14] demonstrate the value of motion cues, but their designs remain tightly coupled to camera imagery and do not transfer well to radar's unique characteristics.

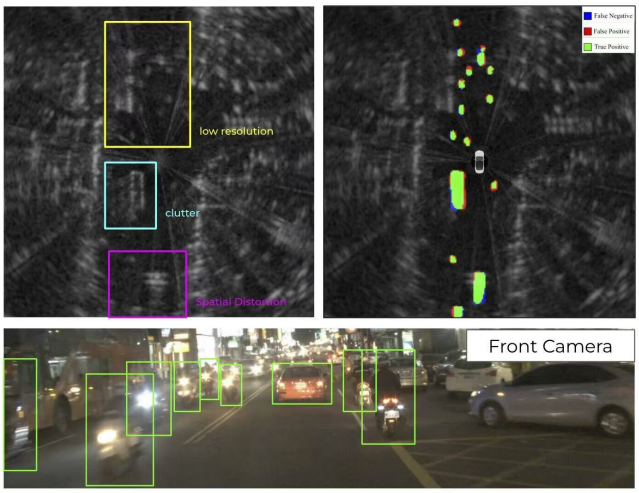


Fig. 1: Challenges of radar-based moving object segmentation. (Left) Scanning radar imagery suffers from low angular resolution, clutter, and spatial distortion. (Right) Our method produces accurate moving object segmentation results with fewer false positives/negatives, even for small and distant objects. (Bottom) A camera view of the same scene for reference.

This gap motivates the exploration of radar as a primary modality. Unlike cameras or LiDAR, scanning radar has attracted increasing attention for its complementary sensing capabilities, and it also introduces unique challenges: low angular resolution, sparse returns, and multipath reflections that induce cluttered and ghost artifacts. Notably, prior radar perception works focus on sparse radar point clouds with irregular distributions, while our studied scanning radar yields denser, sequentially scanned returns with unique geometric properties. This fundamental difference in data characteristics renders existing radar methods inapplicable to our scanning radar MOS task. Existing studies have highlighted these challenges: Srivastav *et al.* identify sparsity and uncertainty as key barriers for deep learning on radar [15]; Zhou *et al.* report that coarse angular resolution makes small or distant objects indistinguishable [16]; and Kopp *et al.* reveal how multipath clutter disrupts motion analysis [17]. As illustrated in Fig. 1, these sensing constraints result in low-resolution targets, clutter-dominated regions, and spatial distortions, underscoring the difficulty of reliable moving object segmentation from radar data.

Ren-Yi Yuan is with the College of Electrical and Computer Engineering, National Yang Ming Chiao Tung University, Hsinchu, Taiwan. y0808.ee11@nycu.edu.tw

Chieh-Chih Wang is with the College of Electrical and Computer Engineering, National Yang Ming Chiao Tung University, and with the Mechanical and Mechatronics Systems Research Laboratories, Industrial Technology Research Institute, Hsinchu, Taiwan. bobwang@ieee.org

Wen-Chieh Lin is with the Institute of Multimedia Engineering, National Yang Ming Chiao Tung University, Hsinchu, Taiwan wclin@cs.nctu.edu.tw

Most prior work on scanning radar has focused on object detection, particularly for vehicles [18]–[21]. Detection methods localize targets at the object level, but they cannot provide the pixel-level motion information required to separate dynamic targets from static clutter. In particular, small objects such as scooters are often missed due to weak reflectivity and sparse radar returns.

Moving object segmentation (MOS) addresses these shortcomings by providing pixel-level motion separation, which can recover fine-grained targets and motion cues that detection overlooks. However, segmentation on radar itself remains challenging: small objects are still difficult to capture because of low angular resolution, sparse backscatter, and high sensitivity to noise. This highlights that small-object difficulty is not only a limitation of detection but also a modality-inherent challenge for radar segmentation. To date, MOS on scanning radar has rarely been explored. While radar has been studied for detection, tracking, and localization, no prior work has addressed segmentation on this modality, leaving a clear and important research gap.

Motivated by these gaps, we propose a radar-native moving object segmentation framework. Our solution introduces a polar-guided attention architecture that explicitly accounts for radar noise, sparsity, and geometric distortion while leveraging spatial priors inherent to the polar domain.

To realize this, we build on a spatio-temporal convolutional (STC) backbone inspired by MotionNet [1], which highlights the importance of temporal encoding for moving object segmentation. Based on this backbone, we propose MotionNet-PGA, which adapts MotionNet to scanning radar and further augments it with a novel Polar-guided Attention module. In this design, motion features are extracted in the Cartesian domain to preserve temporal continuity, while polar-domain appearance cues provide lightweight guidance to enhance geometric fidelity. This formulation explicitly aligns feature learning with radar geometry, thereby suppressing clutter and improving the segmentation of small and distant moving objects.

The key contributions of our paper are summarized as follows:

- We propose a radar-native moving object segmentation framework that builds on the Spatio-Temporal Convolution backbone from MotionNet and extends it with a Polar-guided Attention module. This design explicitly aligns motion and appearance features with radar geometry, improving robustness against clutter, mitigating long-range distortion, and enhancing the segmentation of small moving objects.
- We introduce the first annotated moving object segmentation (MOS) dataset for scanning radar, derived from the ITRI Radar Dataset. This dataset establishes a benchmark for radar-based motion perception and reproducible evaluation.We plan to make the annotations publicly available to facilitate future research. On this dataset, our method achieves an overall IoU of 68.36%, surpassing the strongest LiDAR-based baseline (MotionNet) by +2.48%, with class-specific improvements of +2.34% for cars and +4.08% for scooters.

II. Related Work

Recent studies have explored scanning radar for detection, tracking, and localization [23], [24], yet challenges such as low angular resolution and spatial distortion remain. Importantly, no prior work has focused on moving object segmentation with radar. We therefore review related approaches from LiDAR and cameras, discuss their limitations when transferred to scanning radar, and highlight the insights that motivate our method.

A. LiDAR-based moving object segmentation

Recent progress in LiDAR-based moving object segmentation predominantly depends on utilizing the comprehensive 3D geometric structures afforded by point clouds. These methodologies commonly involve projecting point clouds into structured representations, such as Bird's Eye View (BEV) and Range View, or computing residual images to effectively capture object motion.

LiDAR-based moving object segmentation (MOS) methods leverage dense 3D geometric structures from point clouds to capture motion cues. A common strategy is to project point clouds into structured representations such as Bird's Eye View (BEV) or Range View, or to compute residual images between consecutive frames. Wu et al. [1] proposed a spatio-temporal pyramid network in BEV, jointly performing perception and motion prediction with tailored loss functions to ensure spatial–temporal consistency. Sun et al. [10] adopted a dual-branch design to extract spatial features from range images and temporal features from point clouds, fused via motion-guided attention. Zhou et al. [9] computed height differences in polar BEV columns to obtain motion cues, then integrated them with appearance features through co-attention. Cheng et al. [25] fused BEV, range, and residual views via multi-view attention to improve segmentation robustness. While effective, these approaches assume high-resolution, low-noise 3D point clouds and distinct geometric changes between frames—conditions not met in 2D radar data, which suffer from low angular resolution, severe noise, and projection distortions. Moreover, while some methods (e.g., MV-MOS [25]) use polar coordinates during preprocessing to mitigate sparsity at long range, polar priors are typically discarded in subsequent feature modeling, with convolution and attention applied in Cartesian space. Polar representation is also validated for LiDAR semantic segmentation [26] and radar occupancy prediction [27], but these works focus on static perception and lack temporal motion fusion for dynamic MOS.

B. Camera-based moving object segmentation

Camera-based MOS methods have achieved notable progress in video object segmentation but generally assume high-resolution, texture-rich imagery, limiting their applicability to radar data. Wang et al. [28] proposed VisTR, the first end-to-end video instance segmentation model that formulates object tracking as a set prediction problem via

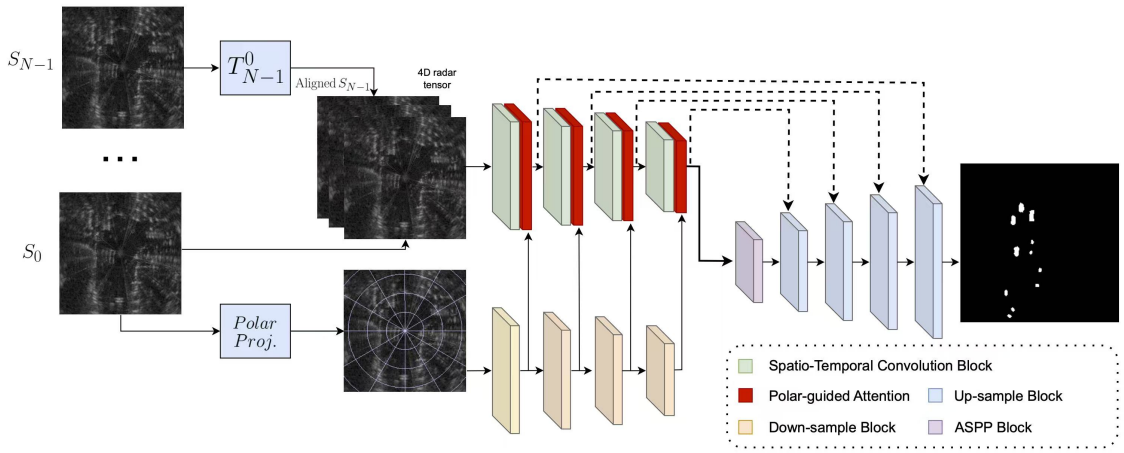


Fig. 2: Overview of MotionNet-PGA. Multiple radar frames $\{S_{n-1}, \ldots, S_0\}$ are aligned to the reference frame S_0 using the transformations T^0_j, and then stacked into a 4D radar tensor. The spatio-temporal convolution (STC Block) encodes motion features in the Cartesian domain, while the polar projection branch converts radar data into range–angle coordinates for polar-domain feature extraction (Down-sample Block). The **Polar-guided Attention Module** (red blocks) fuses motion and appearance features from both domains to enhance small and distant object representation. The decoder (Up-sample Block), together with skip connections and an ASPP (Atrous Spatial Pyramid Pooling [22]) Block, outputs the final moving object segmentation mask.

a Transformer architecture. While effective on RGB videos, it incurs high computational cost and struggles with densely packed small objects. To improve spatial–temporal integration, Cheng et al. [7] introduced SeqFormer, combining multi-scale feature aggregation with temporal context modeling through a Transformer encoder–decoder. However, its attention mechanism relies heavily on visual texture and color contrast, making it less effective for noisy, textureless radar signals. MED-VT [8] further enhanced motion modeling with a multiscale encoder–decoder Transformer to refine object boundaries over time, but its dependence on appearance cues and camera-centric priors reduces performance on radar imagery with weak boundaries and sparse cues. Chen et al. [29] proposed soft-PPM loss and flow difference loss to handle occluded pixels and optical flow uncertainties within a self-supervised motion segmentation framework. DSEC-MOS [30] proposed a moving object segmentation framework based on event camera data, offering high temporal resolution but relying on event-based sensors and optical flow supervision, limiting generalizability to scanning radar systems.

Although both LiDAR- and camera-based methods have achieved strong results in their respective domains, their direct application to radar is hindered by radar-specific challenges such as background clutter, low angular resolution, and projection distortion. Moreover, their attention mechanisms are not tailored to the geometric and signal characteristics of radar data. To address these issues, we propose a polar-guided attention framework that preserves polar-domain representations throughout the network. Unlike LiDAR methods that use polar coordinates only for preprocessing, our approach computes spatial attention directly in the polar domain, maintaining geometric consistency for distant and small targets. A temporal encoding branch further aggregates motion features across multiple frames, avoiding reliance on noisy residual differencing and enhancing robustness in cluttered scenes.

III. Moving Object Segmentation

We propose a radar-based moving object segmentation framework (Fig. 2) that addresses clutter, low angular resolution, and geometric distortions inherent to scanning radar. This framework consists of radar data preprocessing (Sec. III-A), a dual-branch network structure with motion encoding, polar-domain appearance encoding, and attention-based fusion (Sec. III-B), and the implementation details of training and loss design (Sec. III-C).

A. Radar Data Preprocessing

Sequential radar scans suffer from spatial misalignment due to ego motion, which can obscure true dynamics and cause slowly moving objects to appear static. To stabilize the background and highlight motion cues, we estimate the relative pose between consecutive scans using radar odometry [31] and warp earlier scans into the coordinate frame of the current scan.

Let $\mathcal{S} = \{S_j\}_{j=0}^{N-1}$ denote N radar frames, where each pixel $\mathbf{p}_i = [x_i, y_i]^\top$ is in Cartesian coordinates. To align historical frames to the reference S_0, we apply a transformation T^0_j from frame j to frame 0:

$$S_j^{\to 0} = \{\, T^0_j\, \mathbf{p}_i \mid \mathbf{p}_i \in S_j \,\}.$$

Here, T^0_j denotes the mapping that transforms points from frame j into the coordinate system of the reference frame 0. In practice, T^0_j can be obtained by composing the ego-motion transformations between consecutive frames, but for clarity we directly use T^0_j to represent the overall alignment.

Aligned frames are then converted from Cartesian coordinates (x, y) to polar coordinates (r, θ), where

$$r = \sqrt{x^2 + y^2}, \quad \theta = \arctan(y/x).$$

We denote a radar pixel in the polar domain as $\mathbf{p}_i^{\text{polar}} = (r_i, \theta_i)$, which directly corresponds to its Cartesian position

$\mathbf{p}_i = (x_i, y_i)$. Representing radar data in the polar domain preserves the inherent sensing geometry of scanning radar, thereby alleviating distance-dependent distortions and offering more consistent shapes for small or faraway objects. This representation provides a stronger foundation for subsequent feature modeling. Finally, the N ego-motion-compensated polar-domain frames are stacked into a 4D tensor of size $N \times C \times H \times W$ and fed to the encoder, where C, H, and W denote the number of feature channels, the range dimension, and the angle dimension, respectively.

B. Network Structure

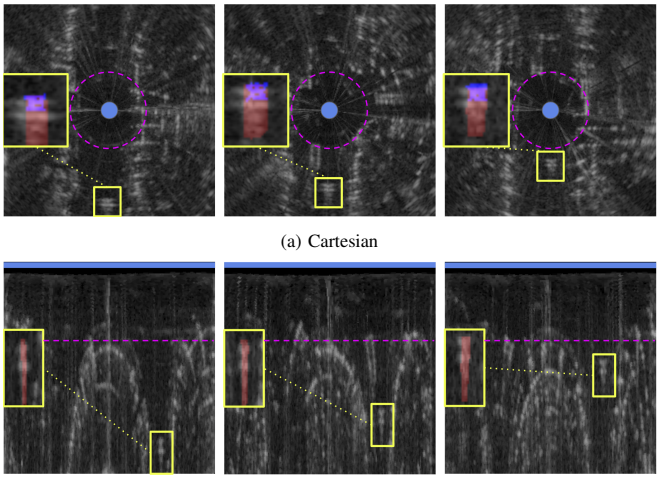
1) Motion Encoder: We adopt a U-Net-like backbone inspired by MotionNet [1], where the core module is the Spatio-Temporal Convolution (STC) block. Each STC block first applies a 2D convolution to extract spatial features from individual frames, followed by a degenerate 3D convolution with kernel size $N \times 1 \times 1$ along the temporal axis to capture motion cues. This design efficiently models temporal dynamics while preserving spatial resolution and robustness to radar-specific noise such as clutter and multipath reflections. STC blocks are stacked hierarchically, with skip connections linking the encoder and decoder.

2) Polar-domain Appearance Encoder: Radar data in Cartesian coordinates often exhibit distance-dependent distortions, particularly for far or angular-edge targets [26], [27]. As shown in Fig. 3, objects in Cartesian space suffer from shape compression and deformation with increasing distance from the ego vehicle, while polar coordinates preserve angular structure and mitigate such distortions.

More importantly, representing radar data in the polar domain is not merely a coordinate transformation, but a modality-consistent design. Since scanning radar inherently acquires measurements in (r, θ) space, retaining this native geometry avoids interpolation artifacts introduced during Cartesian projection and ensures that each bin corresponds to a uniform angular and radial resolution. This leads to a more balanced representation of both nearby and distant objects, preventing small or faraway targets from being excessively compressed.

These polar-domain appearance features, extracted using a standard 2D CNN, provide structurally consistent priors that can reliably guide moving object segmentation. In particular, by reducing geometric bias, the polar representation strengthens the effectiveness of the subsequent attention mechanism, enabling it to highlight genuine moving objects even under severe clutter and long-range distortion.

3) Polar-guided Attention Module: We build upon the motion–appearance attention paradigm commonly used in LiDAR-based MOS [32], but redesign it for the sensing characteristics of scanning radar. In LiDAR, dense 3D geometry provides stable motion cues, making it natural to use motion to guide appearance features. When directly applied to radar, however, this design fails: appearance features are easily dominated by clutter and multipath noise, while motion cues obtained from simple residual differencing—a strategy commonly used in LiDAR-based moving object segmentation—are too unstable to serve as reliable guidance.



(a) Cartesian

(b) Polar

Fig. 3: Comparison between Cartesian and Polar representations of radar data. (a) In Cartesian coordinates, distortion increases with distance from the ego vehicle. (b) Polar coordinates preserve angular structure and mitigate such distortions.

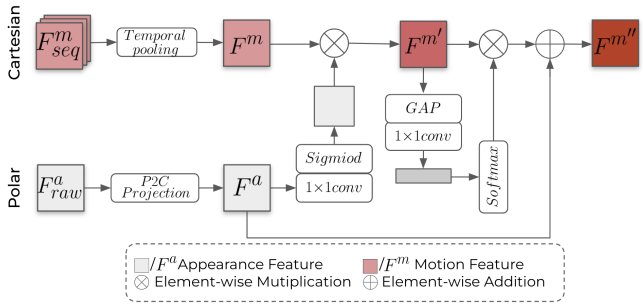


Fig. 4: Structure of the Polar-guided Attention module.

To overcome these limitations, we first replace residual differencing with a multi-frame STC encoder, which produces temporally consistent motion features. More importantly, we invert the guidance direction: rather than motion guiding appearance as in LiDAR, our module treats motion as the backbone and employs polar-domain appearance as lightweight guidance. This formulation is coordinate-aware: motion is extracted in Cartesian coordinates to preserve temporal continuity, while appearance is encoded in the polar domain to retain geometric fidelity for small or distant targets. The polar appearance is then projected back into the Cartesian grid and used to spatially and channel-wise gate the motion backbone, yielding features that are both motion-discriminative and geometry-consistent, as illustrated in Fig. 4.

Formally, let F^m_{seq} denote the sequence of motion features extracted in Cartesian coordinates. After temporal pooling, we obtain the aggregated motion feature $F^m \in \mathbb{R}^{C \times H \times W}$, where C is the number of motion feature channels, and H and W are the height (range dimension) and width (angle dimension) of the Cartesian grid, respectively. Similarly, let F^a_{raw} denote the appearance features encoded in polar (r, θ)

coordinates. To enable interaction between the two domains, F^a_{raw} is projected into the Cartesian grid via a differentiable resampling operator $\Pi_{P\to C}$ (bilinear interpolation):

$$F^a = \Pi_{P\to C}(F^a_{raw}) \in \mathbb{R}^{C_a\times H\times W}, \quad (1)$$

where C_a is the number of appearance feature channels, and H and W follow the same definitions as for F^m. All subsequent operations are performed in Cartesian space to ensure pixelwise alignment between F^a and F^m.

From F^a, we compute a spatial attention mask (gray block) using a 1×1 convolution followed by a sigmoid activation, and apply it to gate the motion features:

$$F'_m = F_m \otimes \sigma(\mathrm{Conv}_{1\times1}(F^a)), \quad (2)$$

where $\otimes$ denotes element-wise multiplication. This spatial gating suppresses clutter-dominated activations in regions where appearance provides weak evidence, while preserving responses where motion and appearance are consistent.

To further refine the motion representation, we apply channel attention. Specifically, a Global Average Pooling (GAP) operation is applied to F^a, followed by a 1×1 convolution and a softmax activation to produce channel-wise weights (gray block). These weights are then applied to rescale the motion features:

$$F''_m = \left(F'_m \otimes \left(\mathrm{Softmax}(\mathrm{Conv}_{1\times1}(\mathrm{GAP}(F^a)))\cdot C\right)\right) + F^a, \quad (3)$$

where $\oplus$ represents element-wise addition for the residual connection.

The residual pathway stabilizes training and injects geometry-preserving cues, producing features that are simultaneously motion-discriminative and geometry-consistent.

The proposed polar-guided attention is lightweight and radar-native. By combining Cartesian motion for temporal continuity with polar appearance for geometric fidelity, it yields features that are both robust to clutter and effective for segmenting small and distant objects.

C. Implementation Details

The proposed network is trained using the Adam optimizer [33] with an initial learning rate of 1.6×10^{-3}. A cosine annealing learning rate scheduler with a warm-up stage is applied to improve convergence stability. Training is conducted on an NVIDIA RTX 4090 GPU with a batch size of 16 for 140 epochs.

Since the dataset exhibits a strong imbalance between foreground (moving) and background (static) pixels, we adopt a weighted cross-entropy loss to address this issue. Let M denote the total number of pixels, C the number of classes, y_i the ground truth, $\hat{y}_i$ the prediction, and w_c the weight assigned to each class. The moving object segmentation loss is defined as:

$$L_{\text{motion}}(y_i,\hat{y}_i) = -\sum_{i=1}^{M}\sum_{c=1}^{C} w_c\, y_i \log(\hat{y}_i). \quad (4)$$

This weighted formulation reduces the bias toward the dominant background class and improves segmentation performance on small moving objects.

Fig. 5: Sensor setup for the data collection car in the ITRI dataset. The green circle highlights the scanning radar (Navtech CIR504-X), and the orange circles indicate LiDARs, from top to bottom: Ouster OS1-128, Velodyne VLS-128, and Baraja Spectrum-Scan. Red circles indicate four cameras.

IV. EXPERIMENTS

In this section, we present the experimental validation of our proposed radar-based moving object segmentation framework. We first describe the dataset and evaluation metrics in Experiment Setups (Sec. IV-A). To ensure fair benchmarking, we then detail the adaptation of several representative LiDAR- and camera-based methods for radar input in Implementation of Competing Methods (Sec. IV-B). Quantitative and qualitative results are reported in Evaluation and Comparisons (Sec. IV-C), where our approach is compared against these baselines. Finally, Ablation Study (Sec. IV-D) investigates the contributions of polar representation, polar-guided attention, and temporal context to the overall performance.

A. Experiment Setups

ITRI Dataset. We conduct experiments on a customized subset of the ITRI radar dataset, collected in dynamic urban environments such as Guangfu Road and the ITRI campus in Hsinchu City. These scenarios feature dense traffic and frequent moving agents, including scooters and vehicles. The sensor suite consists of a Navtech CIR504-X scanning radar operating at 4 Hz, complemented by three LiDAR sensors (Ouster OS1-128, Velodyne VLS-128, and Baraja Spectrum-Scan) and four RGB cameras.

The radar provides a maximum range of 500 m, a horizontal resolution of $1.8°$, and a range resolution of 0.175 m. Each radar frame is projected into a 256×256 Cartesian grid, corresponding to an effective area of 44.8×44.8 m. Fig. 5 illustrates typical urban scenes captured in this dataset.

Manual annotations were created by labeling foreground motion masks in radar frames from sequences containing high dynamic activity, such as busy intersections and campus roads with diverse traffic participants. To ensure sufficient motion diversity, static-dominant scenes were excluded. A total of 1,968 frames were selected, with 1,303 for training, 145 for validation, and 520 for testing. The test set is further divided into three subsets: (1) 257 frames with complex scenes containing multiple object types and occlusions, (2) 142 frames containing only cars, and (3) 121 frames focusing on scooters. All annotated objects are confined to the

TABLE I: Comparison of moving object segmentation IoU (%) across all methods. Our radar-based model achieves the highest accuracy on overall IoU, as well as car and scooter categories.

	Method	IoU (%)	IoU_{Car}	$\text{IoU}_{\text{Scooter}}$
Lidar based	MotionNet (2020)	65.88	70.78	65.19
	MotionBEV (2023)	57.44	52.26	48.32
Camera based	CS-Unet (2022)	63.53	68.15	64.23
	MedVT (2023)	54.36	50.67	49.41
	DSEC-MOS (2024)	52.80	55.90	46.30
Radar based	**Ours**	**68.36**	**73.12**	**69.27**

effective range of 44.8 m. This choice is due to the reliance on LiDAR returns for manual labeling: beyond this distance, the LiDAR point density drops sharply, making it difficult to reliably infer object shapes. Similar observations have also been reported in prior work on LiDAR-based moving object segmentation, e.g., SemanticKITTI-MOS [34].

This dataset provides a challenging benchmark for radar-based moving object segmentation in real-world conditions. The dataset will be released to facilitate further research.

Evaluation Metrics. We evaluate moving object segmentation performance using the Intersection over Union (IoU), defined as

$$\text{IoU} = \frac{TP}{TP + FP + FN}, \tag{5}$$

where TP, FP, and FN denote the numbers of true positives, false positives, and false negatives, respectively.

Following the objective of foreground moving object segmentation, IoU is computed exclusively for the moving object category. Static background pixels, which dominate radar scans and are trivially predicted, are excluded from evaluation. This ensures that the metric reflects the model's ability to detect and localize dynamic entities such as vehicles and scooters.

B. Implementation of the Competing Methods

We benchmark our approach against 5 representative methods originally developed for LiDAR, camera, each adapted for radar input. MotionNet [1], originally designed for LiDAR BEV inputs, jointly performs detection and motion prediction on a per-cell basis; in our adaptation, the BEV input is replaced with raw radar images, and only the state-estimation head is retained to classify moving and static cells. We adapted MotionBEV [9] to radar input by replacing its LiDAR residual images with residual radar maps in the motion branch, and by feeding raw radar frames into the appearance branch. CS-UNet [35], a U-Net with cross-frame transformers, is adapted by stacking three consecutive radar frames as three input channels while preserving the original architecture. MEDVT [8], a video transformer with multi-scale attention, does not rely on optical flow and thus can directly accept radar frames as input to its transformer encoder.

Finally, DSEC-MOS [30], originally designed for event cameras with event-driven priors, is adapted to radar by substituting radar masks for event masks and by omitting event-specific prior inputs. The main encoder–decoder and fusion design are kept intact, enabling a fair evaluation of its visual transformation pipeline on radar data.

For all competing methods, we only adapted the input modality to radar while preserving their original network designs.

C. Evaluation and Comparisons

We compare our approach with state-of-the-art LiDAR- and camera-based MOS methods, as summarized in Table I. Among LiDAR baselines, MotionNet [1] achieves 65.88% IoU, while MotionBEV [9] reaches 57.44%. For camera-based methods, CS-UNet [35] records 63.53%, whereas transformer-based MedVT [8] and DSEC-MOS [30] achieve 54.36% and 52.80%, respectively. Our model attains the best overall performance with 68.36% IoU, surpassing the strongest baseline by +2.48%. The model also achieves the highest IoU for both cars (73.12%, +2.34%) and scooters (69.27%, +4.08%), confirming its robustness across object types.

Qualitative comparisons are shown in Fig. 6. In dense traffic (first row), our method captures small scooters more completely, reducing the fragmented detections observed in MedVT and MotionNet. In close-range scenes (second row), clutter around static vehicles is largely suppressed compared to MotionBEV. In long-range settings (third row), our approach preserves object integrity despite angular distortions, while other baselines tend to break targets into disjoint parts. These results highlight the effectiveness of polar-domain appearance guidance in enhancing motion feature representation for radar-based segmentation.

D. Ablation Study

In this section, we conduct ablation studies to evaluate the contributions of coordinate representation, polar-guided attention, and temporal context length to the overall performance of our model. We first disentangle the effects of attention and coordinate representation. As shown in Table II, using Cartesian motion features without attention achieves 65.88% IoU. Adding attention already provides the major improvement, raising performance to 67.95% (+2.07%), which confirms that attention-guided refinement is the dominant factor. When we further change the appearance branch from Cartesian to Polar, IoU increases to 68.36% (+0.41%). Although the additional gain is smaller, this shows that Polar representation contributes complementary benefits by better preserving the radar's native geometry. Notably, on distant objects, Polar-guided attention provides +0.87% improvement compared to Cartesian-guided, indicating that the geometric fidelity of Polar coordinates is particularly advantageous for long-range segmentation.

To assess the generalizability of the proposed Polar-guided Attention, we integrate it into MotionBEV and CS-UNet. For MotionBEV, which already contains motion and appearance branches with a co-attention module, we retain the original design and insert our Polar-guided Attention module after

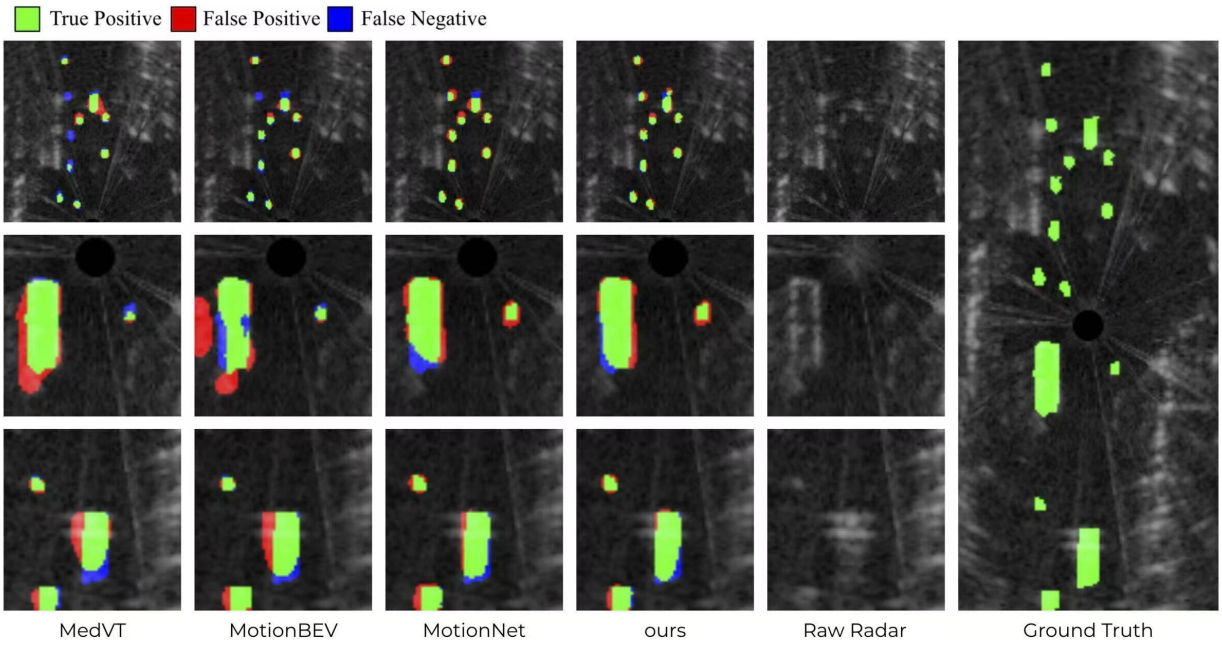


Fig. 6: Qualitative results of different methods on ITRI test set. The results are cropped and zoomed in for better visualization.

TABLE II: Ablation study of attention design. The motion branch is fixed in Cartesian coordinates, while the appearance branch is varied between Cartesian and Polar. Distant-object IoU is evaluated in the far-range region (17.5–44.8 m).

Attention Design	IoU (%)	Distant Objects IoU (%)
No Attention	65.88	65.02
Cartesian-guided	67.95	66.55
Polar-guided (ours)	**68.36**	**67.42**

its attention stage. For CS-UNet, which originally models temporal motion features with a transformer-U-Net, we add an additional appearance branch and apply Polar-guided Attention to fuse it with the motion branch. Table III shows consistent improvements across both backbones. For MotionBEV, IoU increases from 57.44% to 58.23%, with gains in both car and scooter categories. For CS-UNet, IoU improves from 63.53% to 64.59%, including a 1.41% boost for scooters. This confirms that the advantages of Polar-guided Attention extend beyond a specific architecture.

Finally, we study the effect of temporal context length by varying the number of input scans from two to six. As illustrated in Fig. 7, performance improves as more history is included, peaking at five scans with an IoU of 68.36%. Using six scans, however, causes a drop to 66.52%, likely due to accumulated noise and redundancy. This suggests that incorporating sufficient but not excessive temporal context is critical for robust moving object segmentation.

TABLE III: Performance comparison with and without Polar-guided Attention

Method	IoU (%)	$\mathbf{IoU}_{\text{Car}}$	$\mathbf{IoU}_{\text{Scooter}}$
MotionBEV + w/o Polar-guided Attention	57.44	52.26	48.32
MotionBEV + w/ Polar-guided Attention	**58.23**	**55.96**	**51.61**
CS-UNet + w/o Polar-guided Attention	63.53	68.15	64.23
CS-UNet + w/ Polar-guided Attention	**64.59**	**68.21**	**65.64**

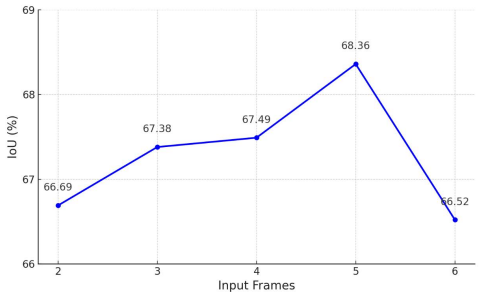


Fig. 7: The moving object segmentation performance vs. the number of input scans N. Tested in ITRI dataset.

V. Conclusion

We propose MotionNet-PGA, a radar-native moving object segmentation framework that builds on a Spatio-Temporal Convolution (STC) backbone from MotionNet, and augments it with a Polar-guided Attention module. This design explicitly aligns motion and appearance features with radar geometry, improving robustness against clutter, mitigating long-range distortion, and enhancing the segmentation of small moving objects. We further introduced the first annotated MOS dataset for scanning radar, derived from the ITRI Radar Dataset, providing a benchmark for reproducible research in this domain. Extensive experiments demonstrate that our approach not only achieves an overall IoU of 68.36%—exceeding the strongest LiDAR-based baseline by +2.48%—but also yields consistent improvements across object categories, including +2.34% for cars and +4.08% for scooters, where the latter highlights the benefit for small moving objects. These results highlight the effectiveness of polar-guided attention and underscore the necessity of exploiting radar-native priors to advance robust motion perception. Looking ahead, we believe this work lays the foundation for future exploration of radar-centric perception, including multimodal

fusion, long-term temporal reasoning, and deployment in real-world autonomous systems.

VI. Acknowledgment

We thank the Autonomous Vehicle Group at MMSL, ITRI, for their assistance with data collection. This work is partially supported by the National Science and Technology Council in Taiwan via NSTC 113-2221-E-A49-148-MY3 and 113-2221-E-A49-159-MY3 and under the "Top Research Centers in Taiwan Key Fields Program" of the Ministry of Education (MOE), Taiwan.

References

[1] P. Wu, S. Chen, and D. Metaxas, "Motionnet: Joint perception and motion prediction for autonomous driving based on bird's eye view maps," 2020. [Online]. Available: https://arxiv.org/abs/2003.06754

[2] R. Kümmerle, M. Ruhnke, B. Steder, C. Stachniss, and W. Burgard, "Autonomous robot navigation in highly populated pedestrian zones," *Journal of Field Robotics*, vol. 32, no. 4, pp. 565–589, 2015.

[3] L. Peters, D. Fridovich-Keil, V. Rubies-Royo, C. J. Tomlin, and C. Stachniss, "Inferring objectives in continuous dynamic games from noise-corrupted partial state observations," 2021. [Online]. Available: https://arxiv.org/abs/2106.03611

[4] X. Chen, A. Milioto, E. Palazzolo, P. Giguere, J. Behley, and C. Stachniss, "Suma++: Efficient lidar-based semantic slam," in *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, Nov. 2019. [Online]. Available: http://dx.doi.org/10.1109/IROS40897.2019.8967704

[5] P. Ruchti and W. Burgard, "Mapping with dynamic-object probabilities calculated from single 3d range scans," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2018, pp. 6331–6336.

[6] I. Vizzo, T. Guadagnino, J. Behley, and C. Stachniss, "Vdbfusion: Flexible and efficient tsdf integration of range sensor data," *Sensors*, vol. 22, no. 3, 2022. [Online]. Available: https://www.mdpi.com/1424-8220/22/3/1296

[7] J. Wu, Y. Jiang, S. Bai, W. Zhang, and X. Bai, "Seqformer: Sequential transformer for video instance segmentation," 2022. [Online]. Available: https://arxiv.org/abs/2112.08275

[8] R. Karim, H. Zhao, R. P. Wildes, and M. Siam, "Med-vt: Multiscale encoder-decoder video transformer with application to object segmentation," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 6323–6333.

[9] B. Zhou, J. Xie, Y. Pan, J. Wu, and C. Lu, "Motionbev: Attention-aware online lidar moving object segmentation with bird's eye view based appearance and motion features," *IEEE Robotics and Automation Letters*, vol. 8, no. 12, p. 8074–8081, Dec. 2023. [Online]. Available: http://dx.doi.org/10.1109/LRA.2023.3325687

[10] J. Sun, Y. Dai, X. Zhang, J. Xu, R. Ai, W. Gu, and X. Chen, "Efficient spatial-temporal information fusion for lidar-based 3d moving object segmentation," 2022. [Online]. Available: https://arxiv.org/abs/2207.02201

[11] Y. Chen, Y. Cao, H. Hu, and L. Wang, "Memory enhanced global-local aggregation for video object detection," in *2020 IEEE/CVF CVPR*, 2020, pp. 10 334–10 343.

[12] H. Perreault, G.-A. Bilodeau, N. Saunier, and M. Héritier, "Ffavod: Feature fusion architecture for video object detection," 2021.

[13] Y. Lyu, Z. Liu, H. Li, D. Guo, and Y. Fu, "A real-time and lightweight method for tiny airborne object detection," in *2023 IEEE/CVF CVPRW*, 2023, pp. 3016–3025.

[14] C. W. Corsel, M. van Lier, L. Kampmeijer, N. Boehrer, and E. M. Bakker, "Exploiting temporal context for tiny object detection," in *2023 IEEE/CVF WACVW*, 2023, pp. 1–11.

[15] A. Srivastav and S. Mandal, "Radars for autonomous driving: A review of deep learning methods and challenges," 2023. [Online]. Available: https://arxiv.org/abs/2306.09304

[16] Y. Zhou, L. Liu, H. Zhao, M. López-Benítez, L. Yu, and Y. Yue, "Towards deep radar perception for autonomous driving: Datasets, methods, and challenges," *Sensors*, vol. 22, no. 11, 2022. [Online]. Available: https://www.mdpi.com/1424-8220/22/11/4208

[17] L. Liu, R. Guan, F. Ma, J. Smith, and Y. Yue, "Radar-stda: A high-performance spatial-temporal denoising autoencoder for interference mitigation of fmcw radars," 2023. [Online]. Available: https://arxiv.org/abs/2307.09063

[18] K. Qian, S. Zhu, X. Zhang, and L. E. Li, "Robust multimodal vehicle detection in foggy weather using complementary lidar and radar signals," in *2021 IEEE/CVF CVPR*, 2021, pp. 444–453.

[19] Y.-J. Li, J. Park, M. O'Toole, and K. Kitani, "Modality-agnostic learning for radar-lidar fusion in vehicle detection," in *2022 IEEE/CVF CVPR*, 2022, pp. 908–917.

[20] Y. Yang, J. Liu, T. Huang, Q.-L. Han, G. Ma, and B. Zhu, "Ralibev: Radar and lidar bev fusion learning for anchor box free object detection system," 2023.

[21] P. Li, P. Wang, K. Berntorp, and H. Liu, "Exploiting temporal relations on radar perception for autonomous driving," in *2022 IEEE/CVF CVPR*, 2022, pp. 17 050–17 059.

[22] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," 2018. [Online]. Available: https://arxiv.org/abs/1802.02611

[23] S. Yao, R. Guan, Z. Peng, C. Xu, Y. Shi, W. Ding, E. Gee Lim, Y. Yue, H. Seo, K. Lok Man, J. Ma, X. Zhu, and Y. Yue, "Exploring radar data representations in autonomous driving: A comprehensive review," *IEEE Transactions on Intelligent Transportation Systems*, vol. 26, no. 6, p. 7401–7425, Jun. 2025. [Online]. Available: http://dx.doi.org/10.1109/TITS.2025.3554781

[24] A. Venon, Y. Dupuis, P. Vasseur, and P. Merriaux, "Millimeter wave fmcw radars for perception, recognition and localization in automotive applications: A survey," *IEEE Transactions on Intelligent Vehicles*, vol. 7, no. 3, pp. 533–555, 2022.

[25] J. Cheng, X. Chen, J. Liang, X. Tang, X. Chen, and D. Li, "Mv-mos: Multi-view feature fusion for 3d moving object segmentation," 2024. [Online]. Available: https://arxiv.org/abs/2408.10602

[26] Y. Zhang, Z. Zhou, P. David, X. Yue, Z. Xi, B. Gong, and H. Foroosh, "Polarnet: An improved grid representation for online lidar point clouds semantic segmentation," 2020. [Online]. Available: https://arxiv.org/abs/2003.14032

[27] P.-C. Kung, C.-C. Wang, and W.-C. Lin, "Radar occupancy prediction with lidar supervision while preserving long-range sensing and penetrating capabilities," 2022. [Online]. Available: https://arxiv.org/abs/2112.04282

[28] Y. Wang, Z. Xu, X. Wang, C. Shen, B. Cheng, H. Shen, and H. Xia, "End-to-end video instance segmentation with transformers," 2021. [Online]. Available: https://arxiv.org/abs/2011.14503

[29] C.-Y. Chen, B.-Y. Lai, Y.-S. Huang, W.-C. Lin, and C.-C. Wang, "Self-supervised motion segmentation with confidence-aware loss functions for handling occluded pixels and uncertain optical flow predictions," in *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2024, pp. 9432–9438.

[30] Z. Zhou, Z. Wu, D. P. Paudel, R. Boutteau, F. Yang, L. V. Gool, R. Timofte, and D. Ginhac, "Event-free moving object segmentation from moving ego vehicle," 2024. [Online]. Available: https://arxiv.org/abs/2305.00126

[31] P.-C. Kung, C.-C. Wang, and W.-C. Lin, "A normal distribution transform-based radar odometry designed for scanning and automotive radars," 2023. [Online]. Available: https://arxiv.org/abs/2103.07908

[32] H. Li, G. Chen, G. Li, and Y. Yu, "Motion guided attention for video salient object detection," 2019. [Online]. Available: https://arxiv.org/abs/1909.07061

[33] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *ICLR*, San Diega, CA, USA, 2015.

[34] J. Behley, M. Garbade, A. Milioto, J. Quenzel, S. Behnke, J. Gall, and C. Stachniss, "Towards 3D LiDAR-based semantic scene understanding of 3D point cloud sequences: The SemanticKITTI Dataset," *The International Journal on Robotics Research*, vol. 40, no. 8-9, pp. 959–967, 2021.

[35] Q. Liu, C. Kaul, J. Wang, C. Anagnostopoulos, R. Murray-Smith, and F. Deligianni, "Optimizing vision transformers for medical image segmentation," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, Jun. 2023, p. 1–5. [Online]. Available: http://dx.doi.org/10.1109/ICASSP49357.2023.10096379