RadarMAE: Injecting Physical Inductive Biases into Masked Autoencoders for Advanced Radar Object Detection

Eunchan Kim[1], Jeongwan Shin[1,2], Jae-Ho Choi[1,2]*

[1] DGIST [2] KAIST InnoCORE LLM
{eunchan.kim,jwshin,jhochoi}@dgist.ac.kr

Abstract

Masked Autoencoders (MAEs) have emerged as a powerful self-supervised paradigm in vision, yet their direct application to radar perception remains challenged by a fundamental domain gap. Specifically, while conventional MAEs are designed for RGB images using isotropic grid patching and standard mean squared error (MSE) loss, radar range-azimuth (RA) heatmaps exhibit anisotropic spatial structures and axis-specific physical artifacts, including multipath reflections and sidelobe responses. In this paper, we propose **RadarMAE**, a radar-aware MAE designed to incorporate the physical characteristics of radar signals. We introduce an axial tokenization strategy that accommodates the resolution disparities between the range and azimuth axes while capturing spatially extended radar responses. Furthermore, considering the distinct characteristics of the two axes, we introduce auxiliary distribution-based losses that encourage consistency in range- and azimuth-profile intensity statistics during masked reconstruction. Using only single-frame RA maps, RadarMAE achieves 64.17% AP_{BEV} at IoU 0.3 on K-Radar-20, outperforming vanilla MAE by 7.29 AP points and its architecture-matched scratch baseline by 3.40 AP points. These results support the effectiveness of incorporating radar-specific physical priors into self-supervised representation learning, yielding representations that benefit downstream radar object detection.[1]

1 Introduction

Frequency-modulated continuous-wave (FMCW) radar plays a pivotal role in autonomous driving (AD) systems, primarily due to its reliable perception capabilities even in adverse weather conditions that severely degrade cameras and LiDAR [1–3]. To extract actionable targets from the radar signals, traditional radar detection algorithms have leveraged signal processing techniques such as the constant false alarm rate (CFAR) detector [4, 5]. However, this approach causes severe information sparsification due to its reliance on a heuristic thresholding pipeline, inevitably discarding weak signal reflections and crucial geometric semantics [6–9]. To overcome this limitation, a paradigm shift toward deep learning-based radar object detection has emerged [1, 2, 6, 10, 11], where supervised learning schemes are generally utilized to capture rich spatial contexts and subtle target signatures directly from dense radar heatmaps. Leveraging these advantages, data-driven radar perception methods can preserve richer signal information and achieve superior detection performance compared to conventional CFAR-based pipelines [2, 6, 12–14].

*Corresponding author.

[1] Project site: https://gracenpraise.github.io/RadarMAE/

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

Despite these advancements, fully unlocking the potential of the radar domain remains bottlenecked by the intrinsic limitations of supervised learning, specifically its strict reliance on massive, high-quality annotations corresponding to input radar heatmaps. Compared to camera or LiDAR data, radar heatmaps are difficult to interpret visually. This non-interpretability stems from the superposition of target returns with multipath reflections and background clutter, further aggravated by sidelobe artifacts that manifest as ghost replicas of the true response. Such intertwined physical phenomena obscure clear geometric boundaries, thereby making manual annotation exceedingly expensive and error-prone [3, 15]. To address this bottleneck, self-supervised learning (SSL) provides a promising approach [16–19], reducing reliance on manual annotations by learning robust representations from large-scale unlabeled radar data [3, 20–27].

Among the available SSL frameworks [16–19], masked autoencoders (MAEs) [17] stand out as an exceptionally suitable architecture for object detection in radar heatmaps: its dense patch-level reconstruction forces the model to capture fine-grained local spatial structures, which is crucial for understanding the intricate geometric relationships of object-specific sidelobe and multipath components. Motivated by these representational advantages, recent studies have attempted to apply MAEs to unlabeled radar signals [20, 21, 26–28]. However, directly applying vision-centric MAE designs to radar heatmaps may fail to capture the distinctive characteristics required for effective radar representation learning in autonomous driving. This limitation stems from the fundamental differences between camera images and radar heatmaps, as illustrated in Fig. 1(a) and (c). While optical images capture visually interpretable object boundaries and surface appearances, radar RA heatmaps represent the same targets through electromagnetic reflection responses processed into range and azimuth measurements, typically involving Fourier-based signal processing. Given this fundamental domain gap, directly adopting MAE designs developed for natural images introduces two key limitations when learning representations from radar heatmaps.

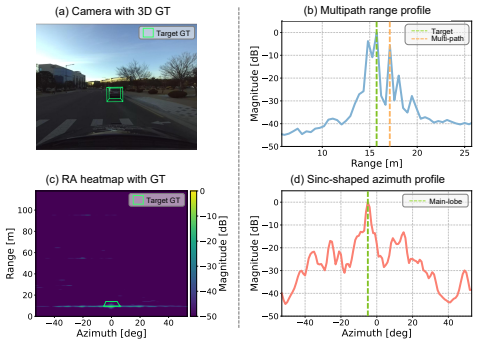

Figure 1: Comparison of camera and radar observations: (a) camera image, (c) corresponding RA heatmap, and radar intensity profiles illustrating (b) range-extended responses and (d) azimuth sidelobes.

The first mismatch lies in the tokenization strategy. Specifically, the standard MAE utilizes square patches, implicitly assuming an isotropic space where the x (*i.e.,* width) and y (*i.e.,* height) axes possess equivalent spatial semantics. In contrast, radar representation, such as range-azimuth (RA) heatmaps, exhibits intrinsically anisotropic properties. Namely, FMCW radar suffers from severe resolution disparity between the range and angle axes. Moreover, target energy disperses along the entire axes in the form of sinc-shaped lobes in polar coordinates (Fig. 3) [9, 29]. Consequently, the spatial footprint of isotropic square patches in the current MAE scheme is inherently incapable of reflecting this asymmetric nature of radar heatmaps as well as fully capturing the sinc-lobe structures.

The second limitation concerns the reconstruction objective. Standard MAE employs a pixel-wise mean squared error (MSE) loss [17], which does not explicitly account for the axis-specific statistical characteristics of radar signals. When masked regions admit multiple plausible reconstructions, MSE-based predictions tend toward the conditional mean, potentially smoothing informative radar responses [30, 31]. In RA heatmaps, multipath reflections can produce delayed and asymmetric intensity profiles along the range axis [32–34], while finite-aperture effects give rise to structured sidelobe patterns along the azimuth axis. These distinct characteristics motivate incorporating axis-specific distributional objectives that encourage the preservation of range- and azimuth-profile statistics during masked reconstruction.

To address these domain-specific challenges, we propose **RadarMAE**, a radar-aware masked autoencoder designed to incorporate inductive biases from radar measurements. Our framework comprises two key components: (1) **axial tokenization**, which represents RA heatmaps as orthogonal, axis-aligned token sequences to accommodate anisotropic resolution and spatially extended radar responses; and (2) **physical distribution-based losses**, which encourage the preservation of axis-specific statistical characteristics associated with multipath reflections and sidelobes during masked reconstruction.

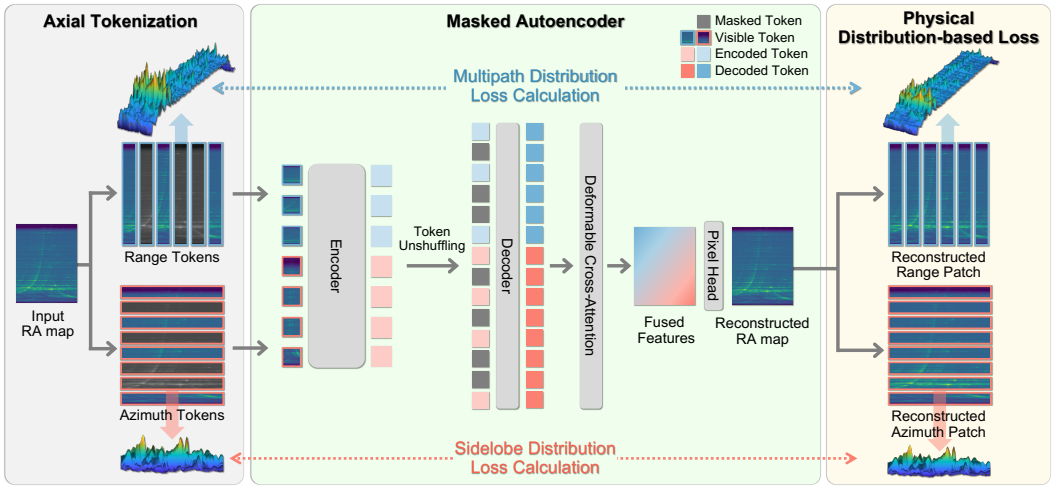

Figure 2: Overview of the RadarMAE pre-training pipeline. RA heatmaps are converted into axial tokens and processed with intensity-weighted masking. Visible tokens are encoded and fused through deformable cross-attention, while reconstruction is guided by pixel-wise and axis-specific physical distribution-based losses.

The main contributions of this work are summarized as follows:

1. We propose axial tokenization, which represents range–azimuth (RA) heatmaps as orthogonal, axis-aligned token sequences. This design explicitly accounts for the anisotropic measurement structure of radar signals while reducing the number of tokens compared with conventional square patch tokenization.
2. We develop a radar-aware masked pre-training strategy that combines intensity-weighted block masking with axis-specific physical distribution-based losses. These components emphasize informative radar responses and encourage the preservation of range- and azimuth-profile statistics during reconstruction.
3. Using only single-frame RA maps, RadarMAE achieves a bird's-eye-view (BEV) AP of 64.17% at IoU 0.3 on K-Radar-20, outperforming vanilla MAE by 7.29 AP points and its architecture-matched scratch baseline by 3.40 AP points.

2 Related Work

Radar Perception with SSL. Unlike camera or LiDAR data, radar signals inherently lack human interpretability, making it exceedingly difficult to construct high-quality, large-scale labeled datasets [3, 15]. To circumvent this critical limitation, recent research has increasingly gravitated toward SSL to exploit vast amounts of unannotated raw radar data [3, 20–24, 26, 27]. Early SSL frameworks in this domain primarily adapted contrastive learning (CL) paradigms, such as SimCLR [16] and MoCo [18], to radar heatmaps [3, 24, 25, 35]. However, CL is fundamentally designed to extract global, instance-level representations; while advantageous for classification, this high-level abstraction obscures fine-grained spatial details and yields suboptimal representations for dense prediction tasks such as object detection, where precise localization is critical [36, 37]. In contrast, MAE [17] formulates dense patch-level reconstruction as a pretext task, compelling the network to recover fine-grained local structures rather than instance-level abstractions [38, 39]. This property aligns naturally with radar heatmap-based detection, where target-specific geometric cues, such as sidelobes and multipath interference, are encoded in delicate spatial patterns that global instance-level contrastive objectives may fail to explicitly preserve [29, 32, 36, 37].

Motivated by the representational advantages of masked modeling, recent work has actively extended MAE to various radar applications. In autonomous driving, frameworks such as Robu-MARC [26] have utilized masked autoencoding to enhance radar-camera fusion and robust 3D perception, while BEV-MAE [27] adopts BEV-guided masking for LiDAR point cloud pre-training and 3D object detection. In the remote sensing domain, models like CROMA [20] and SARMAE [28] have

adopted patch-based masking to extract robust representations from multi-modal and SAR imagery. While these pioneering efforts establish the potential of MAE in the radar domain, their primary focus remains confined to application-level usage rather than fundamentally redesigning the MAE framework itself. Namely, existing radar MAE approaches directly adopt the core mechanics of vision-centric MAE (*i.e.,* tokenization scheme and pixel-wise losses) without radar-specific adaptations. Consequently, the learned representations cannot reflect the anisotropic geometry and dispersion of sinc-shaped sidelobes that fundamentally characterize radar data, leaving substantial room for radar-specific designs.

Injecting Domain-Specific Physical Priors into SSL. Incorporating domain-specific physical priors has become a prevailing strategy for representation learning on signals that inherently deviate from the standard image modality. For instance, in audio and spectrogram analysis, frameworks such as AudioMAE [40] and AST [41] improve representation learning by introducing asymmetric patches tailored to the distinct physical semantics of the time and frequency axes. Similarly, adaptations such as VideoMAE [42] for video and SatMAE [43] for remote sensing incorporate structural priors specific to their respective domains. These successes across diverse modalities collectively demonstrate that injecting the inherent heterogeneity of signal dimensions is crucial for capturing robust features. Despite these advances in related modalities, radar-specific inductive biases remain relatively underexplored in SSL.

3 Preliminary

3.1 Range-Azimuth Representation in FMCW Radar

Physical Basis of the Radar Domain Gap. While camera images capture dense and isotropic spatial reflections in a single snapshot, FMCW radar RA heatmaps are constructed through two orthogonal and physically distinct signal processing pipelines [44, 45]. The range axis is the mapped result of the round-trip time delay of chirp signals reflected from objects. Consequently, the spatial resolution along this axis is fundamentally determined by the sweep bandwidth of the transmitted signal. Conversely, the azimuth axis is derived from the spatial phase differences of the reflected signals received across multiple antenna elements. Accordingly, the angular resolution along this axis is fundamentally determined by the size of the radar's physical or virtual antenna aperture. Originating from fundamentally different physical mechanisms (bandwidth vs. spatial aperture), these two dimensions intrinsically embed highly distinct, dimension-specific physical artifacts into the resulting heatmap, which we explicitly formulate in the following subsection.

3.2 Physical Artifacts in RA Heatmap Distributions

Resolution Disparity. Driven by the aforementioned hardware limitations, the fundamentally different physical mechanisms defining range and azimuth yield severe resolution asymmetry. This structural disparity results in highly anisotropic target profiles across the two dimensions, as can also be observed in Fig. 3.

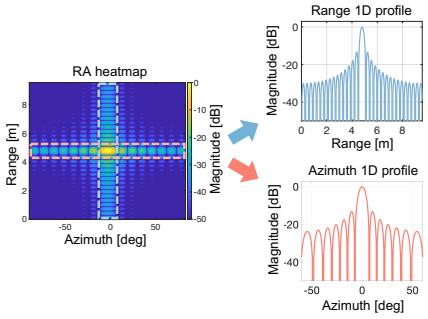

Figure 3: Ideal sinc-shaped point-spread-function (PSF) sidelobes of the FMCW radar RA heatmap.

Multipath Delay. The range dimension exhibits its own distinct artifact: multipath delay, which can distort the range distribution (Fig. 1b) [32–34]. In automotive radar, multipath can arise from indirect reflections involving vehicle surfaces, the ground, and surrounding structures. For example, reflections between a vehicle's undercarriage and the ground surface can produce delayed propagation paths. Because these indirect paths involve longer propagation distances, their responses may appear at farther range bins, contributing to asymmetric and extended range-intensity profiles [32–34]. Although similar multipath effects can also occur in stationary clutter, these target-associated responses may provide additional contextual cues for distinguishing vehicles from surrounding clutter under certain conditions.

Sinc-Shaped Sidelobes. In radar signal processing, applying the fast Fourier transform (FFT) over finite boundaries—restricted by limited bandwidth in range and finite antenna arrays in az-

imuth—mathematically acts as a rectangular window. This inevitably causes spectral leakage in both dimensions, dispersing a target's energy and forming sidelobes governed by $\mathrm{sinc}(x) = \frac{\sin(\pi x)}{\pi x}$ [29, 45, 46]. Similar to multipath effects, these sinc-shaped artifacts form multiple ghost reflections around the target (Fig. 3). Nevertheless, because these sidelobes manifest as deterministic peaks rather than mere random noise, they can aid in understanding the target's underlying properties. Therefore, these inherent sinc characteristics must be explicitly considered when forming effective radar representations.

4 Proposed Methods

To address the limitations of applying standard MAE directly to the radar domain, RadarMAE introduces a radar-aware architecture designed to incorporate the physical characteristics of RA heatmaps. As depicted in Fig. 2, the model is first pre-trained to reconstruct masked radar signals using axial tokenization and radar-aware masking. After pre-training, the decoder is discarded, and a detection head is appended to the pre-trained encoder to fine-tune the network for the downstream task of object detection. The details of each component are as follows.

4.1 Radar-Aware Axial Representation

Axial Tokenization. Standard MAE architectures adopt isotropic square patches, which do not explicitly account for the resolution disparities between the range and azimuth axes or the sinc-shaped reflections extending along these axes. To address this geometric mismatch, we introduce an axial tokenization strategy, which represents the radar signal as orthogonal axial stripes. As illustrated in Fig. 2, by spanning the entire length of each respective axis, these stripes accommodate the resolution disparities while capturing spatially extended radar responses within individual tokens. Prior to linear projection, we apply a lightweight per-stripe early convolution to stabilize local feature extraction without mixing neighboring stripes within each stream. The resulting feature maps ($\tilde{\mathbf{x}}_i^r$ and $\tilde{\mathbf{x}}_j^a$) are then independently projected into range-wise and azimuth-wise tokens:

$$\mathbf{r}_i = \mathbf{W}_r \,\mathrm{vec}(\tilde{\mathbf{x}}_i^r) + \mathbf{b}_r \in \mathbb{R}^d, \tag{1}$$

$$\mathbf{a}_j = \mathbf{W}_a \,\mathrm{vec}(\tilde{\mathbf{x}}_j^a) + \mathbf{b}_a \in \mathbb{R}^d. \tag{2}$$

Here, $\mathrm{vec}(\cdot)$ denotes vectorization, d is the hidden embedding dimension, and $\mathbf{W}_r$, $\mathbf{W}_a$, $\mathbf{b}_r$, and $\mathbf{b}_a$ are learnable projection matrices and bias vectors. Let H and W denote the azimuth and range dimensions, respectively, and P_a and P_r their corresponding tokenization intervals. By decoupling the axes, this formulation reduces the token count from $N_{\mathrm{grid}} = \lceil H/P_a \rceil \lceil W/P_r \rceil$ for conventional grid patching to $N_{\mathrm{axial}} = \lceil H/P_a \rceil + \lceil W/P_r \rceil$, reducing the computational cost of self-attention under the considered configurations.

Physical Positional Encoding. To incorporate radar measurement geometry, we replace conventional index-based embeddings with physical positional encoding (PPE) derived from absolute sensor coordinates. Specifically, PPE maps range coordinates in meters and azimuth coordinates in radians into Fourier features, which are subsequently projected into the hidden dimension through a multilayer perceptron (MLP). After adding PPE and learnable axis-type embeddings to distinguish the two axes, the orthogonal sequences are concatenated into a unified input $\mathbf{Z}_0 = [\mathbf{r}; \mathbf{a}]$. An ablation study comparing PPE with conventional positional embeddings is provided in Appendix G.

4.2 Intensity-Weighted Block Masking

While axial tokenization is highly effective for capturing extended sinc signatures, its orthogonal intersecting structure introduces a critical issue: the potential for local information leakage where the range and azimuth stripes overlap each other. To mitigate this issue and reduce reliance on such shortcuts, we explicitly break this spatial continuity by occluding contiguous segments rather than isolated patches.

Building upon this, we further introduce an intensity-weighted block masking strategy specifically designed for sparse radar RA heatmaps. While standard random masking is highly effective for densely distributed camera images, applying it to radar RA heatmaps risks entirely erasing rare target signals, causing the network to collapse into trivial background interpolation. Based on the domain

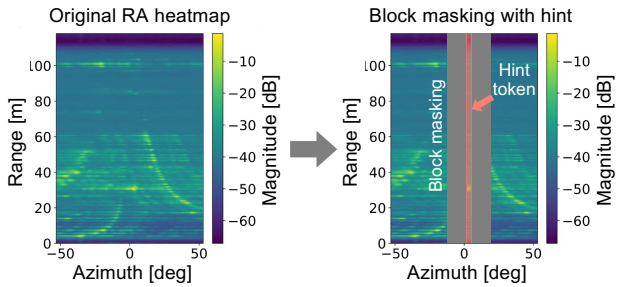

Figure 4: Illustration of intensity-weighted block masking. Contiguous blocks of 5–8 tokens are preferentially selected according to radar intensity. One randomly selected original-measurement token within each selected block remains visible as a hint. Masking is applied independently to the range and azimuth token streams.

prior that high-intensity reflections contain informative target signatures, we prioritize masking these contiguous high-intensity regions to force the network to reconstruct meaningful radar responses rather than background patterns.

Detailed visualizations and ablation analyses of this masking strategy are provided in Appendix F. Finally, to explicitly preserve their distinct physical properties throughout the pipeline, the masked positions are replaced with dedicated learnable mask tokens ($\mathbf{m}_R$, $\mathbf{m}_A$) for the range and azimuth axes, respectively.

4.3 Masked Reconstruction Architecture

Encoder and Decoder. Following the asymmetric MAE paradigm [17], our encoder processes only the visible subset of the unified token sequence $\mathbf{Z}_0$, while a lightweight 8-block decoder reconstructs the signal. Processing the concatenated orthogonal axial stripes through a shared Transformer encoder allows distinct axial inputs to interact via self-attention. This mechanism enables the model to learn cross-axis relationships and geometric dependencies from sparse radar observations. To initiate decoding, we restore the masked positions using the dedicated learnable mask tokens ($\mathbf{m}_R$ and $\mathbf{m}_A$). This asymmetric design encourages the encoder to learn representations from partially observed radar signals.

Deformable Cross-Attention-Based Fusion. To reconstruct the decoupled axial tokens back into a dense 2D patch grid, a geometric mapping is required. We introduce an axial deformable cross-attention mechanism [47]. Let $\mathbf{Q} \in \mathbb{R}^{H_g \times W_g \times d}$ denote a set of learnable 2D spatial queries assigned to a uniform reference grid, where H_g and W_g represent the spatial dimensions. For a query $\mathbf{q}_{ij}$ at reference point p_{ij}, the network predicts continuous 2D sampling offsets Δp_{ijk} and attention weights A_{ijk} for K sampling points. The projected sampling location $p_{ij} + \Delta p_{ijk}$ is subsequently decomposed into orthogonal coordinates x and y. Using these coordinates, the model independently extracts continuous features from the range token bank $\mathbf{V}_r$ and azimuth token bank $\mathbf{V}_a$, which are derived from the outputs of the preceding decoder blocks, via bilinear grid sampling. These sampled features are weighted by A_{ijk}, aggregated along their respective axes, and finally fused via element-wise averaging to yield the 2D feature map. This strategy enables the 2D queries to adaptively integrate information from orthogonal axial tokens.

Reconstruction Objectives. We employ intensity-weighted L1 reconstruction as the primary training objective, complemented by two axis-specific auxiliary losses. While the reconstruction loss minimizes pixel-wise prediction errors, the auxiliary terms encourage the preservation of range- and azimuth-profile intensity statistics. The overall pre-training objective is formulated as

$$\mathcal{L}_{total} = \mathcal{L}_{rec} + \lambda_A \mathcal{L}_A + \lambda_R \mathcal{L}_R. \tag{3}$$

Here, $\mathcal{L}_{rec}$ denotes the base intensity-weighted mean absolute error (L1) loss for stripe reconstruction. The weighting emphasizes stronger measured responses without assuming that all such responses correspond to true targets. Meanwhile, $\mathcal{L}_A$ and $\mathcal{L}_R$ are auxiliary distribution-matching terms for the azimuth and range axes, respectively, with coefficients λ_A and λ_R.

Table 1: Tokenization ablation. For all configurations, MSE loss, random masking, and 500 epochs are applied identically. GFLOPs are measured during pre-training. Best results are shown in bold.

Patch size	# Tokens	$AP_{\mathrm{BEV}}(\%)\uparrow$	Encoder GFLOPs↓	Total GFLOPs↓
16×16	112	49.46	**17.67**	**23.97**
8×8	448	54.91	48.51	75.47
4×16	448	55.23	50.04	77.93
16×4	448	54.59	47.65	74.15
Axial	142	**57.46**	22.22	30.14

Azimuth-Profile Moment Loss. As depicted in Fig. 3, the finite aperture of the antenna array produces structured sidelobe responses along the azimuth axis [45, 48]. These responses can exhibit sharp peaks and asymmetric intensity distributions. To encourage consistency in their statistical characteristics, we match the variance (σ^2), skewness (g_1), and kurtosis (g_2) of the predicted and target intensities for masked azimuth tokens:

$$\begin{aligned}\mathcal{L}_A = &|\sigma^2(y) - \sigma^2(\hat{y})| \\ &+ |g_1(y) - g_1(\hat{y})| \\ &+ \lambda_k |g_2(y) - g_2(\hat{y})|.\end{aligned} \tag{4}$$

Here, y and $\hat{y}$ denote the target and reconstructed intensities, respectively. This moment-matching objective regularizes the distributional characteristics of azimuth responses but does not uniquely constrain the spatial locations, spacing, or periodicity of sidelobe peaks.

Range-Profile Distribution Loss. As formulated in Sec. 3.2, radar waves frequently undergo indirect reflections, such as those from vehicle undercarriages or surrounding structures [32–34]. These delayed reflections can produce asymmetric, heavy-tailed intensity profiles along the range axis. To encourage consistency in the empirical intensity distributions of reconstructed range profiles, we employ a 1D Wasserstein-based distribution loss [49]. For equally weighted scalar intensity samples, the 1D Wasserstein distance can be computed as the mean absolute difference between sorted samples:

$$\mathcal{L}_R = \frac{1}{N_R} \sum_{i=1}^{N_R} \left| \mathrm{sort}(y)_i - \mathrm{sort}(\hat{y})_i \right|. \tag{5}$$

Here, N_R denotes the number of scalar intensity samples from the masked range profiles, while y and $\hat{y}$ represent the target and reconstructed intensities, respectively. The $\mathrm{sort}(\cdot)$ operator is differentiable almost everywhere and has a computational complexity of $\mathcal{O}(N_R \log N_R)$. Because the loss compares sorted intensities rather than their range coordinates, it encourages distributional consistency without explicitly constraining the spatial locations or directions of multipath reflections.

4.4 Fine-Tuning for Object Detection

After pre-training, we discard the decoder and fine-tune the encoder with an anchor-based detection head identical to RTNH [2]. Axial tokens are projected to a BEV feature map through a lightweight stripe-to-BEV mapping, followed by ASPP-Lite [50] for multi-scale context.

5 Experiments

5.1 Datasets and Experimental Settings

K-Radar Dataset. The K-Radar dataset is a large-scale, widely used 4D radar dataset for 3D object detection in autonomous driving, providing 4D radar tensors (4DRT) that contain power measurements along the Doppler, range, azimuth, and elevation dimensions [2]. In this work, we generate RA maps from the provided 3D range-azimuth-elevation (RAE) maps by averaging over the elevation dimension, and use the resulting RA maps as the sole input modality for both RadarMAE pre-training and downstream object detection. We focus on the sedan class within a longitudinal range of $[0, 72]\,\mathrm{m}$ and a lateral range of $[-6.4, 6.4]\,\mathrm{m}$.

Evaluation Metrics. We evaluate detection performance using the intersection-over-union (IoU)-based average precision (AP) under the standard KITTI protocol [51] with 40 recall positions. We

Table 2: Comparison on K-Radar-20. BEV and 3D AP are reported at IoU 0.3/0.5. Methods marked with * use adapted backbones for the corresponding radar input. RadarMAE-Scratch shares the same architecture as RadarMAE but is trained without self-supervised pre-training. Best results are shown in bold.

Method	Backbone	SSL	Input	$AP_{\text{BEV}}(\%)\uparrow$		$AP_{\text{3D}}(\%)\uparrow$	
				AP@0.3	AP@0.5	AP@0.3	AP@0.5
EchoFusion [7]	*ResNet-50	✗	2D (RA)	54.45	42.28	51.75	24.65
RTNH [2]	3D-SCB	✗	3D (RAE)	61.38	46.47	53.05	17.98
ViT-Scratch [52]	ViT-B	✗	2D (RA)	59.49	44.16	55.24	27.64
MAE [17]	ViT-B	✓	2D (RA)	56.88	42.86	52.97	26.47
RadarMAE-Scratch	*ViT-B	✗	2D (RA)	60.77	45.30	55.24	27.45
RadarMAE	*ViT-B	✓	2D (RA)	**64.17**	**48.44**	**58.85**	**30.93**

report the AP for both BEV predictions (AP_{BEV}) and 3D bounding box predictions (AP_{3D}) on the sedan class, at IoU thresholds of 0.3 and 0.5 to assess the spatial localization accuracy of the model.

5.2 Effect of Axial Tokenization in MAE

To analyze the effect of the tokenization scheme in MAE, we first conducted a series of experiments with varying patch sizes, as shown in Table 1. Decreasing the isotropic square patch size from 16×16 to 8×8 yields a substantial improvement in AP_{BEV}. This indicates that finer resolution can better capture the physical signatures of sparse radar signals, albeit at the cost of an increased computational burden. To accommodate the distinct resolution disparities across the RA map axes and to encompass wider sidelobe regions within a single token, we applied asymmetric 4×16 patches, observing a further slight performance increase. Motivated by these findings, we introduced our proposed axial tokenization, which significantly boosts the performance to the highest AP_{BEV} of 57.46%. This confirms that axial tokenization is highly meaningful for effectively recognizing axis-specific resolutions and preserving the spatial continuity of extended sinc-shaped sidelobes without fragmentation. Furthermore, decoupling the spatial axes drastically reduces the total number of tokens from 448 in the 8×8 configuration to 142. As a result, RadarMAE requires less than half the pre-training computational cost compared to the asymmetric MAE variant, while delivering superior object detection accuracy.

5.3 Comparison with Baseline Methods

In this setup, we conduct experiments on the K-Radar-20 subset, consisting of the first 20 sequences of K-Radar, to align the dataset coverage with EchoFusion [7], which reports radar-only RA-map detection results on this subset. Additional performance comparisons on the full K-Radar dataset, comprising all 58 sequences, are presented in Table 6 in Appendix D.

Standard MAE vs. RadarMAE. We further investigate the impact of directly applying a standard vision-centric MAE to radar data. As shown in Table 2, the standard MAE baseline consistently underperforms across all spatial evaluation metrics compared to the ViT-B model trained from scratch without SSL. This notable drop indicates that directly applying image-oriented masked modeling to radar perception is ineffective and can even be detrimental, primarily due to the substantial domain gap and distinct physical characteristics of radar signals. In contrast, RadarMAE successfully overcomes this limitation by explicitly incorporating radar-specific physical priors. Despite being built upon a standard ViT architecture, our proposed framework achieves the best overall performance, reaching a peak of 64.17% in AP_{BEV}@0.3

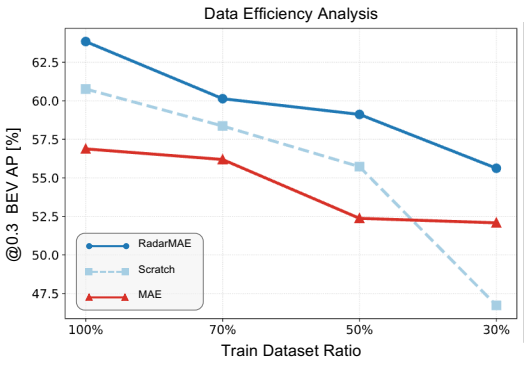

Figure 5: Data efficiency comparison using BEV AP@0.3.

and 30.93% in $AP_{\mathrm{3D}}@0.5$. These results support the effectiveness of incorporating radar-specific physical priors into self-supervised pre-training to learn representations that benefit downstream radar object detection.

Comparison with Other Radar Object Detection Methods. To further evaluate the effectiveness of RadarMAE for radar object detection, we compare it with representative baselines on the K-Radar-20 dataset, as shown in Table 2. To distinguish the effect of self-supervised pre-training from architectural differences, we include RadarMAE-Scratch, which shares the same axial tokenizer, PPE, encoder, stripe-to-BEV mapping, and detection head as RadarMAE but is trained from random initialization. Under this architecture-matched comparison, RadarMAE improves $AP_{\mathrm{BEV}}@0.3$ and $AP_{\mathrm{3D}}@0.3$ by 3.40 and 3.61 points, respectively. These results demonstrate the benefit of radar-aware pre-training beyond the architectural design alone.

Among existing detection methods, our primary baseline is the single-modality 2D RA configuration of EchoFusion [7], which provides a relevant input-matched comparison without camera, LiDAR, or temporal information. Compared to EchoFusion, RadarMAE achieves absolute improvements of 9.72 and 7.10 points in $AP_{\mathrm{BEV}}@0.3$ and $AP_{\mathrm{3D}}@0.3$, respectively. These results demonstrate the effectiveness of the proposed framework for radar object detection using dense RA heatmaps.

We further compare RadarMAE with RTNH [2], a widely used baseline that processes 3D radar RAE tensors. Although both methods employ the same anchor-based detection head, their input representations and encoders differ. Despite relying solely on 2D RA heatmaps without explicit elevation information, RadarMAE achieves higher performance across all reported metrics, including $AP_{\mathrm{3D}}@0.5$ of 30.93% compared with 17.98% for RTNH. This result suggests that radar-aware representation learning can extract informative spatial features for 3D object detection even from a lower-dimensional radar input, highlighting the potential importance of effective radar encoding beyond input dimensionality alone.

5.4 Ablation Studies

Comparison by Training Dataset Ratio. A primary motivation for adopting SSL is to improve data efficiency, which is particularly critical in the radar domain due to the immense difficulty and high cost of manual annotation. To thoroughly evaluate this, we conduct further experiments by varying the available training dataset ratio from 100% down to 30%, as illustrated in Fig. 5. Comparing the standard MAE with the model trained from scratch reveals a notable observation: when labeled data is extremely scarce, specifically at the 30% ratio, the standard MAE provides a distinct pre-training advantage over the scratch baseline. However, as the volume of training data increases, this benefit diminishes, and the standard MAE paradoxically underperforms the scratch model. This inversion implies that misaligned vision-centric priors in standard MAE ultimately hinder the network from fully leveraging larger amounts of radar data. Conversely, RadarMAE exhibits superior data efficiency across all scenarios. By correctly embedding radar-specific physical priors, RadarMAE not only prevents the degradation observed in standard MAE but also consistently maintains a significant performance gap over the scratch baseline across all data availability regimes.

Table 3: Ablation study on reconstruction loss formulations. All metrics are evaluated at IoU 0.3.

Loss	$AP_{\mathrm{BEV}}(\%)\uparrow$
$L2$	57.66
$L1$	62.85
$L1_w$	63.17
$\mathcal{L}_{total}$	64.17

Effect of Physical Distribution-Based Loss. To evaluate the effect of our proposed loss function, we compare different reconstruction losses in Table 3. The standard L2 loss (*i.e.,* MSE) yields the lowest performance (57.66% AP_{BEV}). Switching to an L1 objective substantially improves performance to 62.85%, suggesting that the choice of reconstruction objective is important for radar representation learning. Applying an intensity-weighted strategy ($L1_w$) further improves AP_{BEV} to 63.17%. Finally, incorporating the axis-specific physical distribution-based losses into the reconstruction objective ($\mathcal{L}_{\mathrm{total}}$) achieves 64.17% AP_{BEV}, providing an additional gain of 1.00 AP point over intensity-weighted L1. These auxiliary losses encourage consistency in the intensity statistics of range and azimuth profiles rather than directly constraining the spatial locations of multipath reflections or sidelobe peaks. Overall, these results indicate that the primary reconstruction objective accounts for a substantial portion of the performance improvement, while the physical distribution-based losses provide additional benefits as auxiliary regularizers.

Table 4: Free-space segmentation performance on RADIal. FFT-RadNet uses ADC-derived radar input and is included as a task-level reference. Best performance is shown in bold.

Method	Input	mIoU (%)↑
MAE [17]	RA map	73.03
RadarMAE-Scratch	RA map	74.85
FFT-RadNet [53]	ADC-derived	74.62
RadarMAE	RA map	**75.38**

5.5 Evaluation on RADIal Free-Space Segmentation

To evaluate the applicability of RadarMAE beyond K-Radar object detection, we conduct free-space segmentation experiments on RADIal [53], which differs in radar hardware, acquisition settings, and downstream task. While K-Radar uses the RETINA-4ST radar, RADIal employs a high-definition radar with 12 transmitters, 16 receivers, and 192 virtual antennas. For this evaluation, RadarMAE is independently pre-trained on RADIal RA maps and subsequently fine-tuned for free-space segmentation, without transferring pre-trained weights from K-Radar. This setting allows us to examine whether radar-aware self-supervised pre-training remains beneficial under a different sensor configuration and downstream task. As shown in Table 4, RadarMAE achieves the highest mIoU among the evaluated methods at 75.38%, outperforming vanilla MAE by 2.35 points and its architecture-matched scratch baseline by 0.53 points. The latter comparison helps distinguish the benefit of self-supervised pre-training from that of the architectural design. We also include FFT-RadNet as a task-level reference rather than an input-matched baseline, since it processes ADC-derived range-Doppler observations instead of RA maps. These results provide additional evidence of the applicability of radar-aware self-supervised pre-training to a different radar acquisition setting and downstream segmentation task.

6 Conclusion

In this paper, we proposed **RadarMAE**, a radar-aware MAE framework designed to incorporate the physical characteristics of FMCW radar signals. While conventional MAEs rely on vision-oriented isotropic patching and pixel-wise reconstruction, RA heatmaps exhibit anisotropic spatial structures and axis-specific physical artifacts that motivate radar-aware modeling. To bridge this domain gap, RadarMAE reformulates masked representation learning around the physical measurement structure of radar signals. At its core, axial tokenization represents RA heatmaps as orthogonal range- and azimuth-aligned sequences, accommodating resolution disparities while capturing spatially extended radar responses. Complementing this representation, intensity-weighted block masking emphasizes informative radar responses, while axis-specific auxiliary distribution-based losses encourage consistency in range- and azimuth-profile intensity statistics during reconstruction. Experiments on K-Radar-20 using only single-frame RA maps demonstrate that RadarMAE outperforms both conventional MAE and its architecture-matched scratch baseline, supporting the effectiveness of radar-aware self-supervised pre-training for downstream object detection. Overall, these findings highlight the value of incorporating radar-specific physical inductive biases into self-supervised learning to obtain effective representations for downstream radar perception.

Acknowledgments and Disclosure of Funding

This work was supported by the InnoCORE program of the Ministry of Science and ICT (26-InnoCORE-01 and N10250156), the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. RS-2026-25525987, Digital Columbus Project), the IITP grant funded by the Korea government (MSIT) (No. RS-2025-02219277, AI Star Fellowship Support Project (DGIST)), and the Artificial Intelligence Industrial Convergence Cluster Development Project funded by the Ministry of Science and ICT (MSIT, Korea) and Gwangju Metropolitan City.

References

[1] Marcel Sheeny, Emanuele De Pellegrin, Saptarshi Mukherjee, Alireza Ahrabian, Sen Wang, and Andrew Wallace. Radiate: A radar dataset for automotive perception in bad weather. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1–7. IEEE, 2021.

[2] Dong-Hee Paek, Seung-Hyun Kong, and Kevin Tirta Wijaya. K-radar: 4d radar object detection for autonomous driving in various weather conditions. *Advances in Neural Information Processing Systems*, 35:3819–3829, 2022.

[3] Yiduo Hao, Sohrab Madani, Junfeng Guan, Mohammed Alloulah, Saurabh Gupta, and Haitham Hassanieh. Bootstrapping autonomous driving radars with self-supervised learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15012–15023, 2024.

[4] Harold M. Finn. Adaptive detection mode with threshold control as a function of spatially sampled clutter level estimates. 1968. URL `https://api.semanticscholar.org/CorpusID:208092087`.

[5] Hermann Rohling. Radar cfar thresholding in clutter and multiple target situations. *IEEE Transactions on Aerospace and Electronic Systems*, AES-19(4):608–621, 1983. doi: 10.1109/TAES.1983.309350.

[6] Xu Dong, Pengluo Wang, Pengyue Zhang, and Langechuan Liu. Probabilistic oriented object detection in automotive radar. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2020.

[7] Yang Liu, Feng Wang, Naiyan Wang, and Zhaoxiang Zhang. Echoes beyond points: Unleashing the power of raw radar data in multi-modality fusion. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL `https://openreview.net/forum?id=LZzsn51DPr`.

[8] Yuval Haitman and Oded Bialer. Boostrad: Enhancing object detection by boosting radar reflections. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1638–1647, 2024.

[9] Akarsh Prabhakara, Tao Jin, Arnav Das, Gantavya Bhatt, Lilly Kumari, Elahe Soltanaghai, Jeff Bilmes, Swarun Kumar, and Anthony Rowe. High resolution point clouds from mmwave radar. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 4135–4142, 2023. doi: 10.1109/ICRA48891.2023.10161429.

[10] Sohrab Madani, Jayden Guan, Waleed Ahmed, Saurabh Gupta, and Haitham Hassanieh. Radatron: Accurate detection using multi-resolution cascaded mimo radar. In *European Conference on Computer Vision*, pages 160–178. Springer, 2022.

[11] Mohammadreza Mostajabi, Ching Ming Wang, Darsh Ranjan, and Gilbert Hsyu. High-resolution radar dataset for semi-supervised learning of dynamic objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 100–101, 2020.

[12] Jialong Wu, Mirko Meuter, Markus Schoeler, and Matthias Rottmann. Sparseradnet: Sparse perception neural network on subsampled radar data. In *European Conference on Computer Vision*, pages 52–69. Springer, 2024.

[13] Daniel Brodeski, Igal Bilik, and Raja Giryes. Deep radar detector. In *2019 IEEE Radar Conference (RadarConf)*, pages 1–6. IEEE, 2019.

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

[15] Yi Zhou, Lulu Liu, Haocheng Zhao, Miguel López-Benítez, Limin Yu, and Yutao Yue. Towards deep radar perception for autonomous driving: Datasets, methods, and challenges. *Sensors*, 22(11):4208, 2022.

[16] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org, 2020.

[17] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16000–16009, June 2022.

[18] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020.

[19] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9650–9660, October 2021.

[20] Anthony Fuller, Koreen Millard, and James Green. Croma: Remote sensing representations with contrastive radar-optical masked autoencoders. *Advances in Neural Information Processing Systems*, 36:5506–5538, 2023.

[21] Yi Wang, Hugo Hernández Hernández, Conrad M Albrecht, and Xiao Xiang Zhu. Feature guided masked autoencoder for self-supervised learning in remote sensing. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 18:321–336, 2024.

[22] Yi Wang, Conrad M Albrecht, and Xiao Xiang Zhu. Self-supervised vision transformers for joint sar-optical representation learning. In *IGARSS 2022-2022 IEEE International Geoscience and Remote Sensing Symposium*, pages 139–142. IEEE, 2022.

[23] Hugo Chan-To-Hing and Bharadwaj Veeravalli. Fus-MAE: A cross-attention-based data fusion approach for masked autoencoders in remote sensing. In *IGARSS 2024-2024 IEEE International Geoscience and Remote Sensing Symposium*, pages 6953–6958. IEEE, 2024.

[24] Ruiyuan Song, Dongheng Zhang, Zhi Wu, Cong Yu, Chunyang Xie, Shuai Yang, Yang Hu, and Yan Chen. RF-URL: Unsupervised representation learning for rf sensing. In *Proceedings of the 28th Annual International Conference on Mobile Computing And Networking*, pages 282–295, 2022.

[25] Tianhong Li, Lijie Fan, Yuan Yuan, and Dina Katabi. Unsupervised learning for human sensing using radio signals. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3288–3297, 2022.

[26] Emmanuel Byamugisha, Sheriff Murtala, Danish Sharif, Ingyu Lee, and Gyu Sang Choi. Robu-MARC: A masked autoencoder-aided camera-radar network for robust 3d perception under sensor corruption.

[27] Zhiwei Lin, Yongtao Wang, Shengxiang Qi, Nan Dong, and Ming-Hsuan Yang. BEV-MAE: Bird's eye view masked autoencoders for point cloud pre-training in autonomous driving scenarios. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 3531–3539, 2024.

[28] Danxu Liu, Di Wang, Hebaixu Wang, Haoyang Chen, Wentao Jiang, Yilin Cheng, Haonan Guo, Wei Cui, and Jing Zhang. SARMAE: Masked autoencoder for sar representation learning. *arXiv preprint arXiv:2512.16635*, 2025.

[29] Akarsh Prabhakara, Diana Zhang, Chao Li, Sirajum Munir, Aswin C Sankaranarayanan, Anthony Rowe, and Swarun Kumar. Exploring mmwave radar and camera fusion for high-resolution and long-range depth imaging. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 3995–4002. IEEE, 2022.

[30] Christopher M Bishop and Nasser M Nasrabadi. *Pattern recognition and machine learning*, volume 4. Springer, 2006.

[31] Michael Mathieu, Camille Couprie, and Yann LeCun. Deep multi-scale video prediction beyond mean square error. *arXiv preprint arXiv:1511.05440*, 2015.

[32] Le Zheng, Jiamin Long, Marco Lops, Fan Liu, Xueyao Hu, and Chuanhao Zhao. Detection of ghost targets for automotive radar in the presence of multipath. *IEEE Transactions on Signal Processing*, 72:2204–2220, 2024.

[33] Johannes Kopp, Dominik Kellner, Aldi Piroli, and Klaus Dietmayer. Fast rule-based clutter detection in automotive radar data. In *2021 IEEE International Intelligent Transportation Systems Conference (ITSC)*, pages 3010–3017. IEEE, 2021.

[34] Alexander Kamann, Patrick Held, Florian Perras, Patrick Zaumseil, Thomas Brandmeier, and Ulrich T Schwarz. Automotive radar multipath propagation in uncertain environments. In *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*, pages 859–864. IEEE, 2018.

[35] Matthew Gadd, Daniele De Martini, and Paul Newman. Contrastive learning for unsupervised radar place recognition. In *2021 20th International Conference on Advanced Robotics (ICAR)*, pages 344–349. IEEE, 2021.

[36] Xinlong Wang, Rufeng Zhang, Chunhua Shen, Tao Kong, and Lei Li. Dense contrastive learning for self-supervised visual pre-training. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3024–3033, 2021.

[37] Zhenda Xie, Yutong Lin, Zheng Zhang, Yue Cao, Stephen Lin, and Han Hu. Propagate yourself: Exploring pixel-level consistency for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16684–16693, 2021.

[38] Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. Simmim: A simple framework for masked image modeling. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9653–9663, 2022.

[39] Chen Wei, Haoqi Fan, Saining Xie, Chao-Yuan Wu, Alan Yuille, and Christoph Feichtenhofer. Masked feature prediction for self-supervised visual pre-training. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14668–14678, 2022.

[40] Po-Yao Huang, Hu Xu, Juncheng Li, Alexei Baevski, Michael Auli, Wojciech Galuba, Florian Metze, and Christoph Feichtenhofer. Masked autoencoders that listen. *Advances in neural information processing systems*, 35:28708–28720, 2022.

[41] Yuan Gong, Yu-An Chung, and James Glass. Ast: Audio spectrogram transformer. *arXiv preprint arXiv:2104.01778*, 2021.

[42] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *Advances in neural information processing systems*, 35: 10078–10093, 2022.

[43] Yezhen Cong, Samar Khanna, Chenlin Meng, Patrick Liu, Erik Rozi, Yutong He, Marshall Burke, David Lobell, and Stefano Ermon. Satmae: Pre-training transformers for temporal and multi-spectral satellite imagery. *Advances in Neural Information Processing Systems*, 35:197–211, 2022.

[44] Sujeet Milind Patole, Murat Torlak, Dan Wang, and Murtaza Ali. Automotive radars: A review of signal processing techniques. *IEEE Signal Processing Magazine*, 34(2):22–35, 2017.

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

[46] KM Muraleedhara Prabhu. *Window functions and their applications in signal processing*. Taylor & Francis, 2014.

[47] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020.

[48] Renyuan Zhang and Siyang Cao. 3d imaging millimeter wave circular synthetic aperture radar. *Sensors*, 17(6):1419, 2017.

[49] Aaditya Ramdas, Nicolás García Trillos, and Marco Cuturi. On wasserstein two-sample testing and related families of nonparametric tests. *Entropy*, 19(2):47, 2017.

[50] Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, et al. Searching for mobilenetv3. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1314–1324, 2019.

[51] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the KITTI vision benchmark suite. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3354–3361. IEEE, 2012.

[52] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[53] Julien Rebut, Arthur Ouaknine, Waqas Malik, and Patrick Pérez. Raw high-definition radar for multi-task learning. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17021–17030, 2022. doi: 10.1109/CVPR52688.2022.01651.

[54] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

A Limitations and Future Works

Despite these promising results, RadarMAE currently focuses on representation learning from 2D RA heatmaps. This choice provides a computationally efficient input representation, but it does not explicitly exploit elevation information. A natural future direction is to extend RadarMAE to higher-dimensional radar representations, such as RAE heatmaps. Such an extension could better capture vertical structures and object geometry that are not fully represented in 2D RA heatmaps, potentially leading to improved detection performance at the cost of increased computational complexity.

Although we additionally evaluate free-space segmentation on RADIal, our evaluation remains limited to two radar datasets and a small number of downstream tasks. Further evaluation on broader sensor configurations and tasks, including object tracking, is needed to establish the generality of the learned representations.

In addition, incorporating Doppler information is another promising direction for improving the discrimination of moving objects and stationary clutter, particularly when they exhibit similar intensity patterns. The axis-specific auxiliary losses constrain aggregate intensity statistics rather than the exact locations of multipath returns or sidelobe peaks. Moreover, intersecting axial stripes and intensity-dependent masking can permit partial information recovery during pre-training. These limitations motivate further study of radar representation learning across broader settings.

B Broader Impacts

This work may improve radar-based perception for autonomous driving, particularly under adverse weather and low-visibility conditions where camera- or LiDAR-based systems can degrade. By using self-supervised pre-training, RadarMAE may also reduce reliance on costly manual annotations for radar heatmaps. At the same time, radar perception is safety-critical. Improved benchmark performance does not guarantee safe deployment, and failures such as missed detections, false positives, or poor generalization could affect downstream driving decisions. Thus, RadarMAE should be viewed as a representation learning framework that requires extensive real-world validation before deployment. More capable radar sensing may also raise privacy and misuse concerns, since radar can reveal the presence, motion, and spatial behavior of people and vehicles. Responsible use should include appropriate data governance and compliance with relevant safety and privacy regulations. Finally, although the proposed 2D RA-based design is relatively efficient, future extensions to higher-dimensional radar inputs may increase computational cost, making energy efficiency and deployment practicality important considerations.

C Implementation Details

Pre-training. We pre-train a ViT-Base encoder with an embedding dimension of 768, 12 Transformer layers, 12 attention heads, and an MLP ratio of 4.0. The input RA maps are generated from K-Radar and resized to 112×258 (azimuth$\times$range) with a single channel. The radar intensity values are converted to the dB scale and normalized using dataset-level statistics. We adopt the proposed axial stripe tokenizer and use physical positional encoding (PPE) based on Fourier features. During pre-training, we apply intensity-weighted block masking with a nominal masking ratio of 0.75, contiguous blocks of 5–8 tokens, and a randomly retained visible-hint token per selected block. The model is optimized using a patch-intensity-weighted ℓ_1 reconstruction loss together with auxiliary azimuth-profile moment and range-profile distribution losses, with both auxiliary loss weights set to 0.1. Pre-training is performed for 800 epochs using AdamW [54] with a base learning rate of 1×10^{-5}, weight decay of 0.05, a 50-epoch warm-up, and a cosine learning-rate schedule.

Fine-tuning. For downstream 3D object detection, we initialize the encoder with the pre-trained RadarMAE weights and attach it to an RTNH-style anchor-based detection pipeline with polar-BEV feature refinement. The region of interest is set to $[0, 72] \times [-6.4, 6.4] \times [-2.0, 7.6]\,\mathrm{m}$, with a voxel size of $0.4\,\mathrm{m}$ and an azimuth field of view of $\pm 53^\circ$. We evaluate the *Sedan* class using K-Radar v1.0 annotations under the KITTI evaluation protocol. The detector is fine-tuned using AdamW with an initial learning rate of 1×10^{-4}, a minimum learning rate of 1×10^{-5}, weight decay of 10^{-3}, batch size 8, and cosine annealing for 11 epochs.

Compute resources. The primary K-Radar benchmark experiments were conducted on a workstation equipped with NVIDIA H200 GPUs (140 GB HBM3e per GPU). The factorial ablation in Appendix I was conducted on a separate GPU server using the same codebase and evaluation protocol. Pre-training of the ViT-Base RadarMAE model for 800 epochs on the full K-Radar pre-training set is performed on a single H200 GPU and takes approximately 9 hours, with a peak GPU memory consumption of about 17 GB. Fine-tuning the RTNH-style detector for 11 epochs on a single H200 GPU takes approximately 0.5 hours per run and uses about 7 GB of GPU memory. The full set of experiments reported in the main paper and the supplementary material, including pre-training ablations across tokenizer, positional encoding, masking, and loss configurations, downstream fine-tuning runs across seeds and ablations, and downstream evaluation, amounts to roughly 200 GPU-hours on a single H200. Beyond the runs reported in the paper, the overall research project required additional compute for preliminary experiments, hyperparameter exploration (e.g., masking ratio, patch size, loss weights, learning-rate schedules), and design iterations on the tokenizer and positional encoding, which we estimate to be approximately 2–3$\times$ the compute of the final reported experiments.

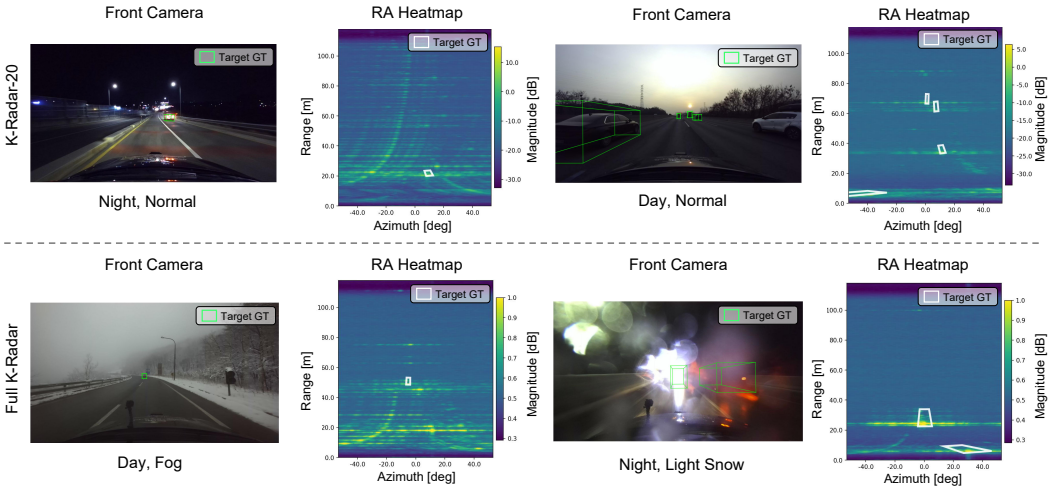

Figure 6: Qualitative examples of front-camera images and corresponding RA heatmaps from K-Radar-20 and the full K-Radar dataset. K-Radar-20 mainly contains normal-weather scenes, whereas the full K-Radar dataset includes more challenging adverse weather conditions such as fog and light snow. Green boxes indicate target ground-truth annotations in the camera view and their corresponding locations in the RA heatmaps.

D Performance on the Full K-Radar Dataset

Table 5: Comparison between K-Radar-20 and the full K-Radar dataset.

Group	Category	K-Radar-20		Full K-Radar	
		Seq.	Frames	Seq.	Frames
Scale	Total	20	13,008	58	34,994
Time	Day	13	9,008	35	21,920
	Night	7	4,000	23	13,074
Weather	Normal	19	12,781	19	12,781
	Overcast	1	227	2	825
	Rain	0	0	8	4,780
	Sleet	0	0	12	7,040
	Fog	0	0	6	3,578
	Light snow	0	0	4	2,237
	Heavy snow	0	0	7	3,753

To clarify the difference between K-Radar-20 and the full K-Radar dataset, we summarize their scale, time-of-day distribution, and weather diversity in Table 5. K-Radar-20 contains 20 sequences and 13,008 frames, most of which are collected under normal weather conditions. In contrast, the full K-Radar dataset contains 58 sequences and 34,994 frames, covering diverse adverse weather conditions such as rain, sleet, fog, light snow, and heavy snow [2]. Fig. 6 provides qualitative examples showing that the full dataset includes more challenging weather scenes where camera observations are degraded while RA heatmaps remain informative.

We further evaluate RadarMAE on the full K-Radar dataset to verify its scalability under broader data distributions. As shown in Table 6, RadarMAE achieves the best performance across the reported BEV and 3D detection metrics. It outperforms the scratch ViT-B and standard MAE baselines, providing further evidence of the effectiveness of the radar-aware design beyond the K-Radar-20 subset. RadarMAE also surpasses RTNH despite using only 2D RA maps instead of 3D RAE inputs, although these methods use different input representations and architectures.

Table 6: Performance comparison on the full K-Radar dataset. Best results are shown in bold.

Method	Backbone	SSL	Input	$AP_{\mathrm{BEV}}(\%)\uparrow$		$AP_{\mathrm{3D}}(\%)\uparrow$	
				AP@0.3	AP@0.5	AP@0.3	AP@0.5
RTNH [2]	3D-SCB	✗	3D (RAE)	58.04	42.60	49.65	17.87
ViT-B [52]	ViT-B	✗	2D (RA)	58.45	44.18	53.29	28.99
MAE [17]	ViT-B	✓	2D (RA)	55.32	42.85	51.99	28.00
RadarMAE	*ViT-B	✓	2D (RA)	**60.48**	**47.76**	**56.43**	**29.04**

E Qualitative Results

E.1 Object Detection Results

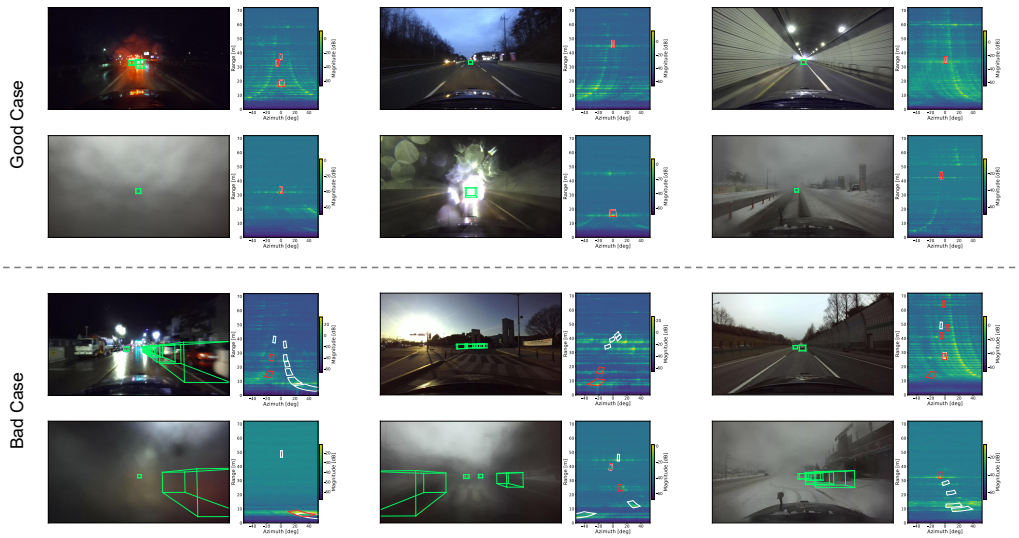

Figure 7: Qualitative detection results of RadarMAE on challenging driving scenes. The top group shows representative good cases, where RadarMAE accurately localizes target vehicles from single-frame RA heatmaps despite adverse visual conditions such as nighttime, fog, tunnels, and snowy roads. The bottom group shows representative bad cases, where weak target responses, severe clutter, or ambiguous radar reflections lead to missed detections or inaccurate localization. In the camera view, green boxes denote ground-truth annotations. In the RA heatmaps, white boxes denote ground-truth target regions, while red boxes denote the predictions of RadarMAE.

E.2 Visualization of Occlusion-Based Sensitivity Maps

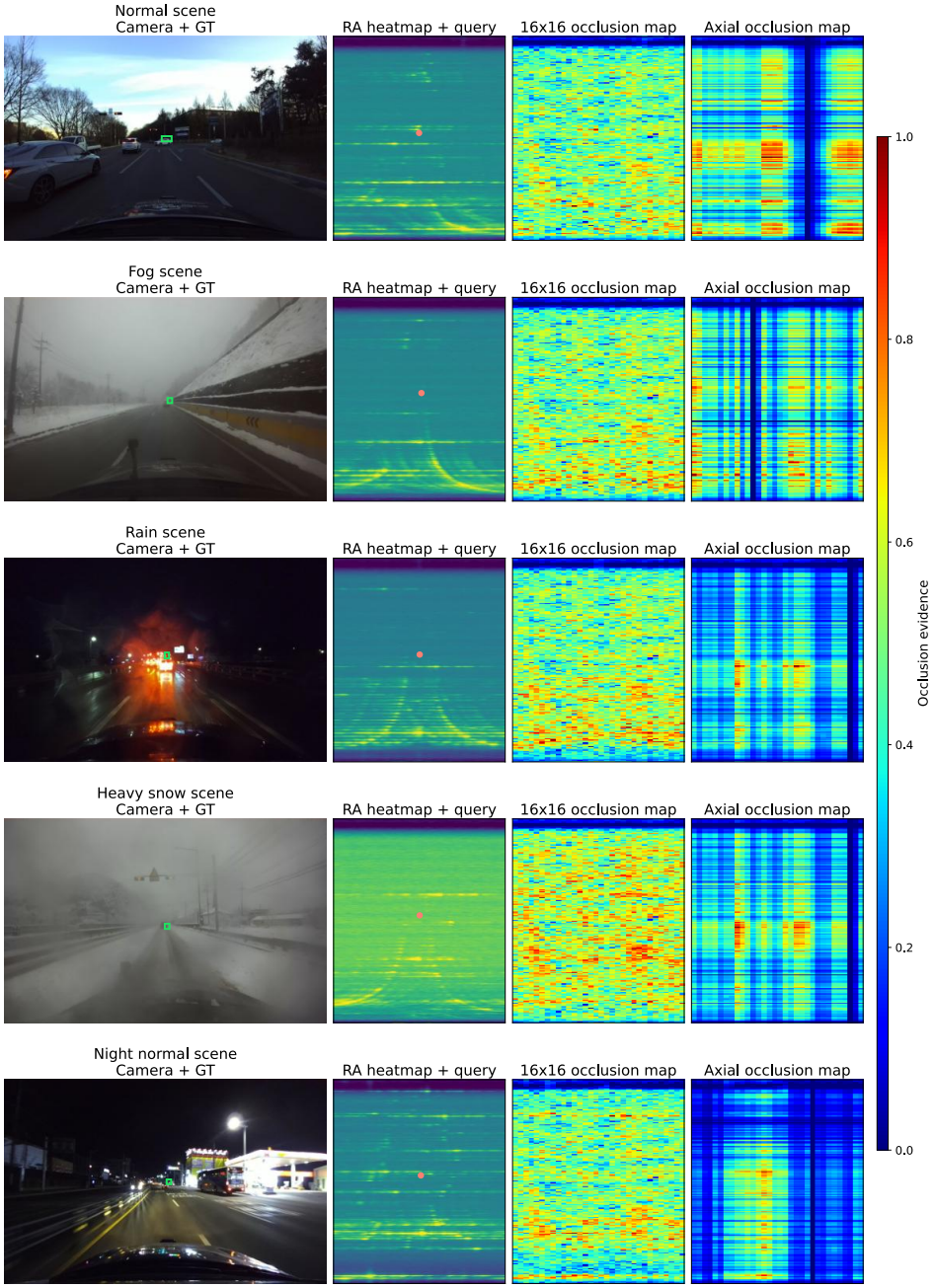

Figure 8: Visualization of occlusion-based attention evidence under diverse driving conditions. Each row shows a front-camera image with the ground-truth target, the corresponding RA heatmap with the query location, the occlusion map obtained from conventional 16×16 square patches, and the occlusion map obtained from the proposed axial tokenization. Compared with the noisy and spatially fragmented responses of square patch occlusion, axial occlusion produces more structured axis-aligned evidence around target-related radar signatures under the evaluated conditions.

F Weighted Block Masking

The intensity-weighted block masking procedure and visible hint mechanism are illustrated in Fig. 4 in the main paper. As shown in Table 7, weighted block masking outperforms random masking across all reported BEV and 3D detection metrics. In particular, it improves $AP_{\mathrm{BEV}}@0.3$ from 60.95% to 64.17% and $AP_{\mathrm{3D}}@0.5$ from 28.22% to 30.93%. These results support the effectiveness of the proposed masking strategy under the evaluated protocol.

Table 7: Comparison of masking strategies on K-Radar-20. BEV and 3D AP are reported at IoU 0.3/0.5. Best results are shown in bold.

Masking	$AP_{\mathrm{BEV}}(\%)\uparrow$		$AP_{\mathrm{3D}}(\%)\uparrow$	
	$AP@0.3$	$AP@0.5$	$AP@0.3$	$AP@0.5$
Random	60.95	44.98	55.54	28.22
Weighted block	**64.17**	**48.44**	**58.85**	**30.93**

Controlled Mask Selection Comparison. To further isolate the effect of intensity-weighted mask selection, we conduct a separate comparison between uniform random and intensity-weighted selection while keeping the block size, masking ratio, hint-token policy, architecture, and training protocol unchanged. Intensity-weighted selection improves $AP_{\mathrm{BEV}}@0.3$ from 61.79% to 62.53%, corresponding to a gain of 0.74 AP points. This result suggests that intensity-aware selection provides an additional benefit under matched conditions, although intensity-dependent masking may still reveal partial information about the input signal.

G Positional Embedding

We further analyze the effect of positional encoding by comparing conventional positional embedding (PE) with the physical positional encoding (PPE) described in Sec. 4. In this experiment, all components of the final RadarMAE framework are kept fixed, and only the positional encoding strategy is changed. As shown in Table 8, PPE improves both BEV and 3D detection performance under the evaluated protocol, suggesting that measurement-coordinate-based embeddings provide useful spatial cues for downstream localization.

Table 8: Comparison of positional embedding strategies on K-Radar-20. BEV and 3D AP are reported at IoU 0.3/0.5. Best results are shown in bold.

Positional Embedding	$AP_{\mathrm{BEV}}(\%)\uparrow$		$AP_{\mathrm{3D}}(\%)\uparrow$	
	$AP@0.3$	$AP@0.5$	$AP@0.3$	$AP@0.5$
PE	61.23	45.04	56.19	28.81
PPE	**64.17**	**48.44**	**58.85**	**30.93**

H Fine-Tuning Details

Feature Reshaping. After the self-supervised pre-training, we discard the decoder and initialize the encoder with the learned weights to adapt the representations for the downstream detection task. To construct a dense spatial feature map from the axial tokens, we employ a direct stripe-to-BEV mapping that projects the axial token bank onto a 2D BEV grid. During this projection, geometric alignment is achieved by generating polar reference points based on the BEV query coordinates, where learned offsets explicitly compensate for the nonlinear coordinate transformations between the radar polar space and the Cartesian BEV space. Finally, a lightweight ASPP-Lite block [50] is applied to the projected BEV feature map to enrich multi-scale contextual information without heavy computational overhead.

Detection Head. We adopt a standard anchor-based detection head identical to the RTNH baseline [2]. The entire network is fine-tuned end-to-end using a joint detection loss $\mathcal{L}_{\text{det}}$, which is a combination

of focal loss for classification, smooth-L1 loss for bounding box regression, and a directional classification loss.

Table 9: Factorial ablation of RadarMAE components on K-Radar-20. All configurations use the same primary L1 reconstruction objective and evaluation protocol. Checkmarks indicate enabled components, while blank cells indicate disabled components. Best performance is shown in bold.

Axial Tokenization	Block Masking	Physical Loss	AP_{BEV}@0.3 (%)↑
			61.79
		✓	61.93
	✓		62.53
	✓	✓	63.74
✓			64.78
✓		✓	64.53
✓	✓		65.16
✓	✓	✓	**65.61**

I Factorial Ablation of RadarMAE Components

Table 9 presents a controlled $2 \times 2 \times 2$ factorial ablation of axial tokenization, block masking, and physical distribution-based losses. All eight configurations were evaluated on the same GPU server using the same codebase and evaluation protocol. Axial tokenization provides the largest individual improvement, increasing AP_{BEV}@0.3 from 61.79% to 64.78%. With axial tokenization enabled, block masking further improves performance to 65.16%, while adding the physical losses yields 65.61%. However, without block masking, adding the physical losses slightly decreases performance from 64.78% to 64.53%.

The factorial experiments were conducted on a different GPU server from the main benchmark. Accordingly, we interpret the results primarily through comparisons within the factorial experiment rather than directly comparing absolute AP values across the two experimental environments. These results identify axial tokenization as the primary contributor and suggest that the auxiliary losses have a configuration-dependent effect.

J Robustness to Radar Sampling Resolution

To evaluate the sensitivity of RadarMAE to radar sampling resolution, we independently reduce the range or azimuth sampling resolution by a factor of two, following the experiments reported in our rebuttal. As shown in Table 10, RadarMAE exhibits smaller performance degradation than both MAE and RadarMAE-Scratch. Specifically, its BEV AP@0.3 decreases by only 0.40 and 1.23 points under reduced range and azimuth sampling, respectively. These experiments assess robustness to sampling degradation rather than changes in physical antenna aperture or beamwidth.

Table 10: BEV AP@0.3 (%) under reduced range and azimuth sampling on K-Radar-20. Best performance is shown in bold.

Method	Original	Range/2	Azimuth/2
MAE	56.88	48.87	44.36
RadarMAE-Scratch	60.77	58.03	49.25
RadarMAE	**64.17**	**63.77**	**62.94**

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes].

   Justification: The abstract and introduction clearly state the main contributions of Radar-MAE, including axial tokenization, physical distribution-based losses, and downstream radar object detection performance. These claims are supported by the method descriptions in Sec. 4 and the experimental results in Sec. 5.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes].

   Justification: The paper discusses limitations and future directions in Appendix A, including the current focus on RA heatmaps and the need to extend RadarMAE with higher-dimensional radar representations such as elevation and Doppler information.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [N/A].

Justification: The paper does not present formal theoretical results, theorems, or proofs. The mathematical formulations are used to define the proposed architecture and loss functions.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes].

Justification: The paper describes the dataset, input representation, evaluation protocol, model architecture, pre-training objective, fine-tuning procedure, and main hyperparameters in Sec. 4, Sec. 5, and Appendix C. These details provide a reasonable path to reproduce the main experimental results.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes].

   Justification: The experiments are conducted on the publicly available K-Radar dataset, which is properly cited. We provide anonymized code in the supplemental material, including the implementation of RadarMAE and instructions for reproducing the main experimental results.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes].

   Justification: Sec. 5 describes the dataset, evaluation metrics, and comparison settings, while Appendix C provides implementation details such as model size, input resolution, masking ratio, optimizer, learning rates, batch size, and training schedules.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No].

Justification: The paper reports extensive comparisons and ablation studies under the same evaluation protocol, but does not include error bars, confidence intervals, or statistical significance tests. This is mainly due to the high computational cost of repeated self-supervised pre-training and downstream fine-tuning experiments.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes].

   Justification: Section C reports the GPU type (NVIDIA H200, 140 GB), per-experiment wall-clock time and peak GPU memory for both pre-training and fine-tuning, the total estimated GPU-hours for all reported experiments, and discloses that the full research project required roughly 2–3$\times$ this amount of compute for preliminary and failed experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes].

   Justification: The research uses an existing autonomous driving radar dataset and proposes a self-supervised radar representation learning method. To the best of our knowledge, the work conforms to the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.

- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes].

    Justification: The paper includes a dedicated Broader Impacts section discussing potential positive impacts, such as improved radar-based perception under adverse weather and reduced reliance on costly manual annotations. It also discusses potential negative impacts and risks, including safety-critical deployment failures, privacy and misuse concerns, data governance, and computational cost considerations for higher-dimensional radar extensions.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A].

    Justification: The paper does not release a high-risk generative model, scraped dataset, or model with obvious dual-use misuse risks. The method is designed for radar object detection in autonomous driving research.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.

- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes].

    Justification: The paper properly cites the K-Radar dataset and the prior baseline method whose detection head is used in our fine-tuning stage. We explicitly state that the publicly available K-Radar dataset and any reused software components are used in accordance with their official licenses and terms of use.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes].

    Justification: We provide anonymized source code in the supplemental material to support reproduction of the proposed method. The code is accompanied by documentation and instructions for environment setup, data preparation, training, and evaluation.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A].

    Justification: The paper does not involve crowdsourcing experiments or research with human subjects.

    Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A].

    Justification: The paper does not involve human-subject experiments, user studies, or newly collected human-subject data.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A].

    Justification: The proposed method does not use LLMs as part of the core methodology, experiments, or scientific contribution. LLMs were used only for writing assistance, such as grammar correction and wording suggestions, and this did not affect the research method, scientific rigor, or experimental results.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.