RADAR AS TOKENS: LEARNING UNIFIED RADAR-NATIVE REPRESENTATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Millimeter-wave radar is an increasingly important sensing modality for robust perception under adverse weather (outdoor) and lighting conditions (indoor). However, learning transferable radar representations for diverse indoor/outdoor downstream tasks remains challenging because radar measurements are both noisy and geometrically non-uniform: multipath, sidelobes, and low angular resolution introduce ambiguous returns, while the polar-coordinate representation format induces range-dependent spatial resolution; that is, the further from the antenna, the larger the cross-range resolution cell. Existing radar perception pipelines often reuse LiDAR- or Cartesian-grid representations, leaving downstream models to resolve radar-specific noise and geometry implicitly through task supervision. We introduce (**Rad**ar as **Tok**ens), referred to as RadTok, a radar-native, task-agnostic representation. RadTok is built on two complementary principles. First, it adopts task-agnostic discrete spatial tokenization, inspired by VQ-style representation learning, to compress radar tensors into compact tokens that suppress ubiquitous radar clutter while preserving consistent target structure. Second, it introduces radar-native operations through resolution-aware polar-grid resampling and a learned de-sharpener that models radar's spatially varying degradation. The resulting representation is denoised, resolution-aware, and directly consumed by existing spatial-feature-based perception pipelines without any modification to their architectures. With a single tokenizer design, RadTok consistently improves radar perception across multiple downstream tasks, reducing indoor absolute pose error by 17.8 % on RT-Pose and, on K-Radar, outperforming baseline methods by a margin of $10+$ AP in outdoor 3D detection and 2.2 mIoU in occupancy grid prediction.

1 INTRODUCTION

Millimeter-wave radar has received increasing attention recently as a promising complementary sensing modality alongside the predominant 3D ranging sensors such as RGB-D and LiDAR. This owes to radar's unique advantages such as motion cues from the Doppler effect and robust operation under adverse weather and lighting conditions, which place radar as a strong candidate for reliable perception in computer vision tasks such as 3D detection and human motion understanding.

Despite this potential, learning an effective representation from radar remains highly challenging. Radar measurements break two assumptions baked into typical backbones and heads used for 3D perception. First, spatial bins are assumed to have a uniform size, but radar's physical cell area is proportional to $r \cdot \Delta\theta$ for a fixed range resolution and varies by up to $16\times$ across a single scene at a typical 1° azimuth beamwidth, far exceeding the assumption of uniform bin size. Second, returns are assumed to originate from physical surfaces, but sidelobe and multipath leakage produce ghost echoes that place an object's energy at locations where nothing physical is. Existing radar perception pipelines reuse representations designed for LiDAR or Cartesian grids, leaving the downstream model to learn these radar-specific properties through task supervision alone, which limits accuracy and cross-scene generalization.

In this work, we propose RadTok, a task-agnostic radar representation that explicitly addresses both, illustrated in Fig. 1. To filter inconsistent noise, we adopt a discrete tokenization based on lookup-free quantization, where the bottleneck preserves consistent target structure while suppressing inconsistent ghost echoes. To handle the non-uniform polar geometry, we introduce a resolution-

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

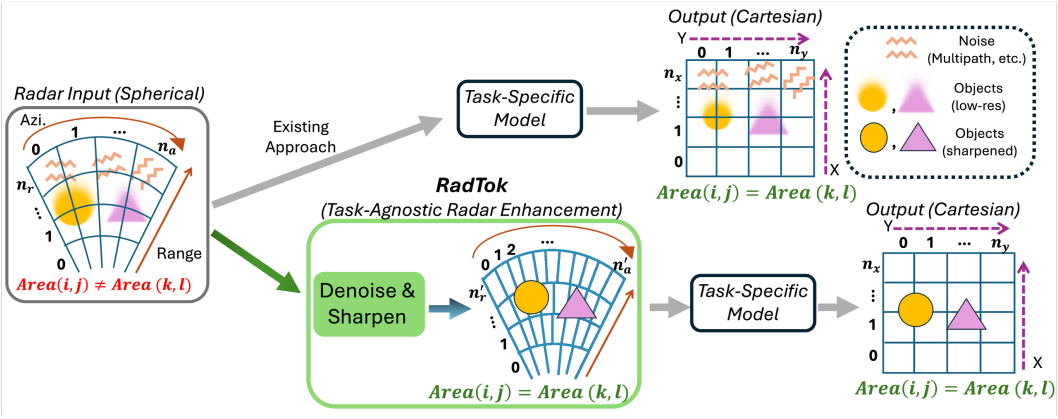

Figure 1: Conceptual comparison of how the radar representation is processed for downstream tasks in existing approaches and RadTok (Ours). RadTok addresses the intrinsic issues of radar input's inconsistent noise and non-uniform resolution for task-specific models, which commonly expect uniform-resolution spatial features as input.

aware spatialization that resamples the azimuth axis with a range-dependent factor, producing a feature map whose physical cell area varies less with range. We supervise the pipeline end-to-end through inverse degradation, where the high-resolution prediction is forward-degraded back to the input shape and matched against the input itself, which requires no paired clean data. The resulting representation can be fed into existing spatial-feature backbones optimized for downstream tasks for radar without any architectural modification. Our contributions are summarized as follows.

1. We identify two intrinsic challenges of radar perception, namely non-uniform spatial resolution that varies up to 16× within a single scene, and ghost echoes from sidelobe and multipath leakage, and address them at the representation level rather than through downstream task supervision.
2. We introduce RadTok, a task-agnostic radar representation with radar-native formulation that combines discrete tokenization for noise filtering with resolution-aware polar-grid spatialization, supervised end-to-end via inverse degradation without paired clean data.
3. Across three diverse downstream tasks, indoor human pose estimation, outdoor 3D object detection, and outdoor occupancy grid prediction, RadTok consistently improves over both task-specific baselines and recent radar representation learning methods.

2 Related Work

Spatial Tokenization. Spatial tokenization converts continuous visual inputs into discrete or compact tokens. Pioneered by vector quantization for image reconstruction van den Oord et al. (2017); Razavi et al. (2019), it has been extended to discrete generative modeling Esser et al. (2021); Chang et al. (2022), transformer-based vision Dosovitskiy et al. (2021); He et al. (2022); Ryoo et al. (2021); Rao et al. (2021), and 3D modalities Yu et al. (2022); Pang et al. (2022); Zhang et al. (2022); Li et al. (2023). We adapt VQ-style tokenization to radar to address the sensor-specific challenges of inconsistent noise and non-uniform resolution.

Spatial Feature Representations for 3D Perception. Projecting sensor data into structured spatial features $\mathbf{F} \in \mathbb{R}^{H \times W \times C}$ has become standard for 3D perception, pioneered by BEV-based detectors Yang et al. (2018); Lang et al. (2019); Yin et al. (2021) and extended to multi-modal Bai et al. (2022); Liu et al. (2023); Li et al. (2024) and unified detection, segmentation, and occupancy settings Zhang et al. (2023); Li et al. (2025). RadTok produces a tensor of the same spatial layout that these backbones expect, and therefore requires no modification to existing perception heads.

Radar-based Spatial Features. Radar-based methods learn spatial feature maps from range-azimuth or Doppler tensors, voxelized grids, or full 4D radar tensors Wang et al. (2021); Yang et al. (2020); Ouaknine et al. (2021); Rebut et al. (2022); Pan et al. (2024) for tasks such as detection, segmentation, tracking, pose estimation, and occupancy prediction. More recent works utilize the full 4D radar

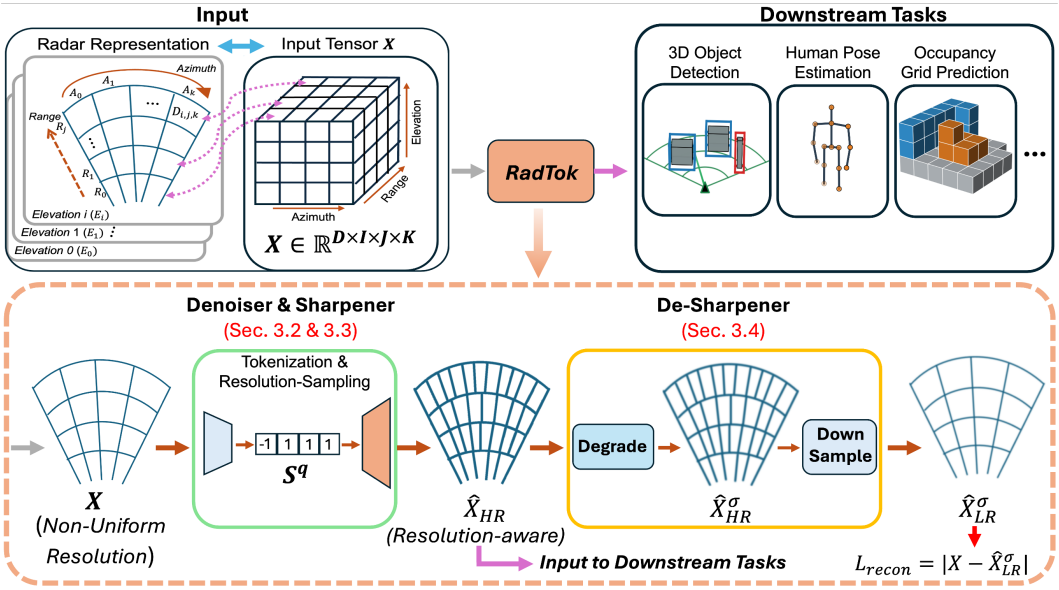

Figure 2: Overall pipeline of the proposed method, RadTok. The radar input tensor $\mathbf{X}$ is fed to *Denoiser* using spatial tokenization to reconstruct only the consistent signals. The output is then resampled and passed through the decoder $\mathcal{D}(\cdot)$, which is used for inference. During training, the reverse process of blurring and downsampling is performed to reconstruct the input $\mathbf{X}$.

tensor since extracting a point cloud discards the radar-native information in the original measurement. For instance, HuPR Lee et al. (2023) and RT-Pose Ho et al. (2024) target 3D keypoint estimation from 4D radar, and RadarOcc Ding et al. (2024) targets occupancy prediction on K-Radar Paek et al. (2022). These works treat the radar tensor as a generic spatial feature map and do not explicitly model radar's non-uniform resolution or ghost echoes. In terms of input representation, we follow these recent works and use 4D radar tensor as our input.

Radar Representation Learning. While sensor fusion with LiDAR or camera has been explored Bang et al. (2024); Zhao et al. (2024); Lai et al. (2026), we focus on the unimodal radar setting. Existing approaches fall into two paradigms. For contrastive learning, Radical Hao et al. (2024) combines cross-modal radar-to-vision and intra-modal contrastive objectives over augmented range-azimuth heatmaps with random beamforming, and a similar BYOL-style scheme has been applied to range-Doppler maps Decourt et al. (2024). For autoencoder pretraining, GRT Huang et al. (2025) trains a Vision Transformer encoder and decoder on raw 4D radar tensors with LiDAR-supervised occupancy, and Occupancy-MAE Min et al. (2024) applies masked autoencoding originally developed for LiDAR voxels. Despite these attempts, none explicitly address radar's non-uniform resolution or ghost echoes, and most are evaluated on a single radar platform and task. We address both at the representation level with a task-agnostic representation that combines discrete tokenization for noise filtering with resolution-aware spatialization, and that generalizes across radar platforms and downstream tasks without modification.

3 Method

3.1 Overview

Given a radar input tensor $\mathbf{X} \in \mathbb{R}^{D \times I \times J \times K}$, where D, I, J, and K are the Doppler, elevation, range, and azimuth dimensions, our goal is to produce a compact, noise-robust, resolution-aware feature map that drops into existing downstream perception models without architectural modification. The method addresses two intrinsic challenges of radar perception. First, the spatial resolution is non-uniform and worsens with range. Second, sidelobe and multipath leakage produce ghost echoes that place an object's energy at locations where nothing physical is.

Our framework consists of three modules, illustrated in Fig. 2. *Radar-aware tokenization* (§3.2) compresses the channel-rich content of $\mathbf{X}$ into discrete tokens whose discrete bottleneck preserves

consistent target structure while suppressing non-reproducible noise. *Resolution-aware spatialization* (§3.3) lifts these tokens onto a polar-grid-aware geometry that reduces the range-dependent variation in physical cell area, and a decoder synthesizes a high-resolution feature map $\hat{\mathbf{X}}_{\mathrm{HR}}$. *Self-supervised training via inverse degradation* (§3.4) supervises the pipeline by forward-degrading $\hat{\mathbf{X}}_{\mathrm{HR}}$ back to $\mathbf{X}$'s shape through a learned blur followed by inverse resampling, and matching it against $\mathbf{X}$ itself. This approach requires no paired ground truth. At inference, $\hat{\mathbf{X}}_{\mathrm{HR}}$ is fed to downstream models for tasks such as 3D detection and pose estimation.

3.2 Radar-Aware Tokenization

Our tokenizer plays two roles. First, it filters inconsistent noise through a discrete bottleneck. Second, it retains the structure needed for downstream high-resolution recovery. A lightweight convolutional encoder $\mathcal{E}(\cdot)$ maps the radar input to a feature map $\mathbf{S} = \mathcal{E}(\mathbf{X}) \in \mathbb{R}^{H \times W \times C}$, capturing local spatial correlations such as multipath patterns across several range and azimuth bins, and mixing the channel-rich phase and Doppler content into compact features. Here, H, W, and C are the azimuth, range (within a region of interest), and feature dimensions. The encoder folds the input's Doppler (D) and elevation (I) axes into C, leaving only the spatial azimuth and range axes.

Each spatial position (h, w) holds a feature vector $\mathbf{s}_{hw} \in \mathbb{R}^C$, which we discretize via lookup-free quantization (LFQ) Yu et al. (2024). Every element of $\mathbf{s}_{hw}$ is mapped to $\{-1, +1\}$ with the sign function $\mathbf{s}^q_{hw} = \mathrm{sign}(\mathbf{s}_{hw})$, trained with straight-through gradient estimation Bengio et al. (2013). The full quantized tensor is then $\mathbf{S}^q \in \{-1, +1\}^{H \times W \times C}$. Through this discrete bottleneck, signals with consistent spatial structure are preserved by the discrete bottleneck, while transient noise and multipath artifacts are suppressed. More specifically, by mapping locally perturbed returns to the same discrete code, the bottleneck stabilizes the tokens against ghost echoes, sidelobes, and clutter while preserving recurring, physically meaningful structure. The resulting tokens form a compressed prior that the next stage spatializes into high-resolution detail.

3.3 Spatializing Tokens onto a Resolution-Aware Grid

Radar lives on a polar grid whose physical cell area is proportional to $r \cdot \Delta\theta$ for a fixed range resolution, varying with range r and azimuth bin width $\Delta\theta$, while standard spatial backbones treat all cells equally. We compensate by upsampling the azimuth axis with a range-dependent factor $f(r) = \mathrm{clamp}(r/r_{\min}, 1, f_{\max})$, so that the cross-range spacing of output samples becomes approximately constant in physical units, bounded by the maximum factor $f_{\max}$.

More specifically, as shown in Fig. 3, for each range column w at physical distance $r(w)$, the differentiable bilinear operator $\mathrm{Lift}(\cdot)$ places the H input rows at the centered $H \cdot f(r(w))$ output rows, and $\mathrm{ToUniform}(\cdot)$ fills the remaining rows. We denote the composite forward operator as $\mathrm{ToHR}(\cdot) = \mathrm{ToUniform} \circ \mathrm{Lift}$, which yields the resampled tokens $\hat{\mathbf{S}}^q = \mathrm{ToHR}(\mathbf{S}^q) \in \mathbb{R}^{\hat{H} \times W \times C}$ with $\hat{H} = H \cdot f_{\max}$. A lightweight decoder $\mathcal{D}(\cdot)$ then synthesizes the final high-resolution feature map $\hat{\mathbf{X}}_{\mathrm{HR}} = \mathcal{D}(\hat{\mathbf{S}}^q) \in \mathbb{R}^{\hat{H} \times W \times C}$, interpreting the channel-rich token content into spatial structure consistent with the LFQ codebook prior. $\hat{\mathbf{X}}_{\mathrm{HR}}$ is fed to the downstream model at inference, while recovering the underlying clear consistent signal that the raw input lacks requires the additional supervision in §3.4.

3.4 Self-Supervised Training via Inverse Degradation

Radar's non-uniform resolution can be modeled as a forward degradation that maps an underlying clear and consistent signal on the resolution-aware grid to the observed input $\mathbf{X}$. Our core intuition is to flip the problem. Instead of learning to deblur the raw input, we model the forward degradation explicitly and train the pipeline to invert it. Specifically, $\mathbf{X}$ is fed through the encoder, quantizer, and decoder to produce $\hat{\mathbf{X}}_{\mathrm{HR}}$, which is then forward-degraded back to $\mathbf{X}$'s shape and matched against $\mathbf{X}$.

The forward degradation is a learned per-position Gaussian blur with axis-wise widths $\boldsymbol{\sigma}_r, \boldsymbol{\sigma}_a \in \mathbb{R}^{\hat{H} \times W}$, predicted from $\mathbf{X}$ by a small CNN, followed by $\mathrm{ToLR}(\cdot)$, the inverse of the resolution-aware resampling ToHR. The reconstruction loss is the L1 difference between the degraded prediction and

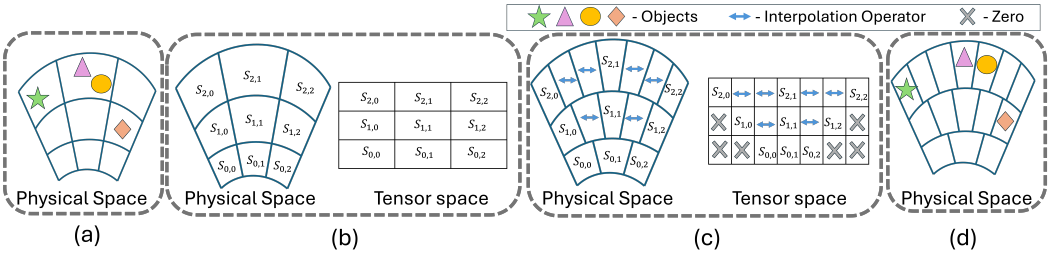

Figure 3: Conceptual diagram demonstrating how resolution-aware sampling works. Due to the non-uniform resolution of the radar input, objects are represented differently depending on their location with respect to the radar, (a). The input is tokenized, (b), for resolution-aware resampling, (c), using differentiable bilinear sampling. The resulting resolution-aware radar tensor $\hat{\mathbf{X}}_{\mathrm{HR}}$ is then fed to the downstream models, (d).

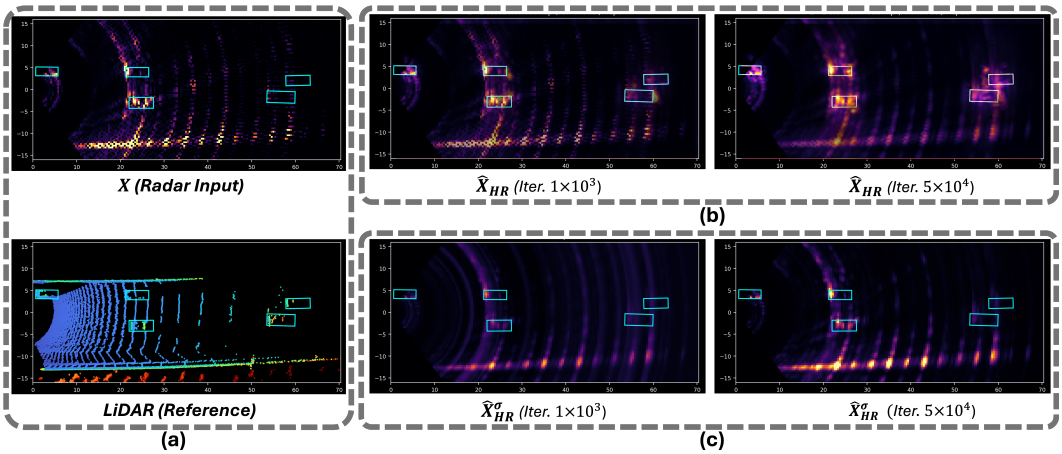

Figure 4: Illustration of how the training process improves the radar representation in BEV on the K-Radar dataset. (a) Input $\mathbf{X}$ and LiDAR for visual reference, where the cyan bounding boxes indicate the ground-truth object locations. (b) The denoised and sharpened representation $\hat{\mathbf{X}}_{\mathrm{HR}}$. As training progresses, the signals become amplified (bright) around the object regions, while inconsistent noisy signals are suppressed (dark). At the same time, the degraded output progressively becomes similar to $\mathbf{X}$, illustrating the effect of the inverse-degradation pipeline.

the original input, given by

$$\mathcal{L}_{\mathrm{recon}} = \left\| \mathrm{ToLR}\big(\mathrm{Blur}_{\boldsymbol{\sigma}_r,\boldsymbol{\sigma}_a}(\hat{\mathbf{X}}_{\mathrm{HR}})\big) - \mathbf{X} \right\|_1. \tag{1}$$

At inference, the blur and ToLR stages are skipped and $\hat{\mathbf{X}}_{\mathrm{HR}}$ is used directly.

We additionally regularize the discrete bottleneck following Yu et al. (2024). The entropy loss $\mathcal{L}_{\mathrm{ent}} = \mathbb{E}[H(q(\mathbf{S}))] - H[\mathbb{E}(q(\mathbf{S}))]$ encourages codebook utilization, while the commitment loss $\mathcal{L}_{\mathrm{commit}}$ stabilizes the encoder and quantizer interface. The full tokenization loss is

$$\mathcal{L}_{\mathrm{token}} = \mathcal{L}_{\mathrm{recon}} + \alpha\,\mathcal{L}_{\mathrm{commit}} + \gamma\,\mathcal{L}_{\mathrm{ent}}. \tag{2}$$

3.5 Downstream Integration

The tokenized, resolution-aware representation serves as a drop-in replacement for existing radar feature extractors optimized for different downstream tasks. Before downstream processing, we adapt $\hat{\mathbf{X}}_{\mathrm{HR}}$ to the input dimensions expected by each backbone, leaving the task backbone and head unchanged. Therefore, the resulting feature map is directly fed to the downstream model $\mathcal{F}(\cdot)$ to obtain task outputs $\hat{y}$, such as 3D bounding boxes or body keypoints. The overall objective combines the self-supervised tokenization loss with the task-specific supervision as $\mathcal{L} = \mathcal{L}_{\mathrm{task}} + \lambda_{\mathrm{tok}}\,\mathcal{L}_{\mathrm{token}}$, where λ_{tok} balances the two. Since $\mathcal{L}_{\mathrm{token}}$ is fully self-supervised, the representation can be trained jointly with any downstream objective or pre-trained independently and transferred to new tasks.

Fig. 4 shows the qualitative progress of the training with $\mathcal{L}$. We feed the decoded map rather than $\mathbf{S}^q$ because published radar pipelines are designed and optimized for radar heatmaps.

4 EXPERIMENTS

We evaluate on three downstream tasks spanning indoor and outdoor environments, namely (i) *indoor human pose estimation*, (ii) *outdoor 3D object detection*, and (iii) *outdoor occupancy grid prediction*. The tasks differ in radar hardware, spatial resolution, scene characteristics, and downstream architectures, providing a comprehensive test of our approach.

4.1 DATASETS AND EVALUATION METRICS

RT-Pose Ho et al. (2024) (Indoor, Pose Estimation). An indoor dataset captured with a 77 GHz mmWave radar that produces 4D tensors, covering a region of approximately 10 m $\times$ 7.2 m at 4.5 cm range and 15.7 cm azimuth resolution. Ground-truth 3D poses (15 body keypoints) come from a motion capture system. We evaluate on the *full dataset* (111 train and 27 test sequences). Following Ho et al. (2024), we report MPJPE and ABS (absolute MPJPE) with lower being better.

K-Radar Paek et al. (2022) (Outdoor, 3D Detection). An outdoor driving dataset with a 77 GHz FMCW radar producing 4D tensors over a 72 m detection range at 0.4 m grid resolution, an order of magnitude larger than the indoor setting. We evaluate Sedan and Bus/Truck classes, reporting 3D AP at IoU 0.3 under the KITTI protocol. We use the revised (v2) labels and an ROI of $x \in [0, 72]$ m, $y \in [-16, 16]$ m, $z \in [-2, 7.6]$ m with azimuth $\pm 53^\circ$, identical for all methods. Both datasets are read as dense 4D tensors.

K-Radar Paek et al. (2022) (Outdoor, Occupancy Prediction). Following RadarOcc Ding et al. (2024), we additionally predict binary 3D occupancy (foreground vs. background) on K-Radar, and report mean IoU (mIoU), foreground IoU (FG), and background IoU (BG).

4.2 COMPARING METHODS

We compare against *task-specific baselines* (no representation learning) and *radar representation learning methods*. For all representation learning methods, we use the same downstream backbone (HRNet3D for RT-Pose, AnchorHeadSingle for K-Radar), and each method is adapted to the same input dimensions expected by the downstream backbone, ensuring an architecture-controlled comparison where only the input representation differs.

Task-specific baselines. For RT-Pose, we compare against HRNet3D Ho et al. (2024). For K-Radar, we compare against RTN Paek et al. (2022), and RadarOcc Ding et al. (2024) for the object detection and occupancy prediction, respectively. These methods process the raw radar input directly without a learned representation stage. For the K-Radar's object detection task, we utilize RTN instead of RTNH because RTNH keeps only the top 10 % of power measurements as a sparse point set, discarding the radar native information, which potentially limits the long-term improvement of the performance, as suggested by recent studies Ho et al. (2024); Ding et al. (2024).

Radical Hao et al. (2024) (Contrastive). A MoCo-based contrastive method originally designed for range-azimuth heatmaps with random beamforming augmentation. Since our 4D tensor lacks raw antenna channels, we adapt it to the Doppler domain by generating two augmented views via Doppler roll and channel dropout. A lightweight decoder is appended to map encoder features back to the input shape for downstream processing.

GRT Huang et al. (2025) (Autoencoder). A Vision Transformer encoder and decoder trained to reconstruct the 4D radar tensor. Originally proposed for single-chip radar (8 azimuth $\times$ 2 elevation $\times$ 256 range $\times$ 64 Doppler), we adapt it to our tensors by adjusting the patch size and feed the decoder output to the downstream backbone.

Occupancy-MAE Min et al. (2024) (Masked Autoencoder). A masked autoencoder for 3D voxel grids that randomly masks 3D voxel positions and reconstructs the missing content. Originally developed for LiDAR point clouds, we adapt it to our dense radar tensor by treating the elevation, range, and azimuth axes as the 3D voxel grid, with Doppler bins serving as the per-voxel feature

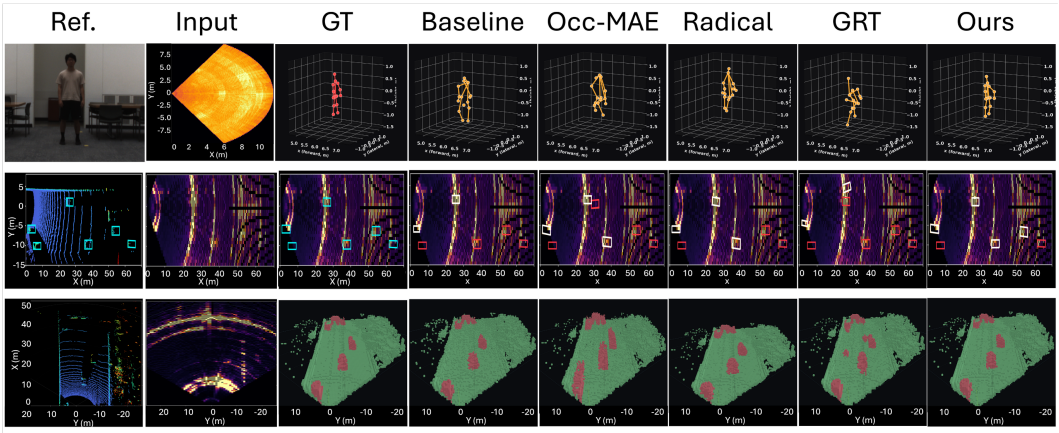

Figure 5: Qualitative comparisons of task baselines, and representation learning-based methods, OCC-MAE, Radical, GRT, and RadTok (Ours) on human pose estimation with RT-Pose dataset (top), 3D object detection (middle), and 3D occupancy grid prediction (bottom) with K-Radar dataset.

channels. Random 3D voxel positions are masked, and the encoder reconstructs their full Doppler spectrum.

4.3 Implementation Details

Tokenizer. For both datasets, we use lookup-free quantization (LFQ) Yu et al. (2024) with a 64-bit binary code per spatial position. The encoder uses depthwise-separable convolutions with 5×5 spatial kernels and 64-dimensional intermediate features. The commitment and entropy loss weights are set to $\alpha = 0.25$ and $\gamma = 0.1$, respectively, and the resampling factor is set to $f_{\max} = 4$.

RT-Pose. Utilizing the published baseline for the dataset, the backbone is HRNet3D Ho et al. (2024) with a CenterHead for heatmap-based keypoint detection, and we use the pixel-unshuffle stem to convert resampled features to backbone resolution. Training uses Adam with lr = 0.0015 (one-cycle schedule), gradient clipping at 35, EMA decay 0.9995 on the full dataset.

K-Radar. Utilizing the published baseline for the dataset, the backbone is a CNN followed by bilinear interpolation to the detection head resolution, with AnchorHeadSingle from Paek et al. (2022) as the detection head. Training uses AdamW with lr = 0.0005, gradient clipping at 10, cosine annealing, and EMA decay 0.999.

Training. Following the compared methods Hao et al. (2024); Huang et al. (2025); Min et al. (2024), RadTok is first pre-trained and then fine-tuned on each downstream task. During pre-training, only the tokenizer is trained with $\mathcal{L}_{\text{token}}$ for 30 epochs, while the downstream backbone and head remain frozen. During fine-tuning, the whole model is initialized from this checkpoint and trained with $\mathcal{L} = \mathcal{L}_{\text{task}} + \lambda_{\text{tok}}\mathcal{L}_{\text{token}}$ for 30 epochs. All experiments use 2×NVIDIA A10 GPUs.

4.4 Main Results

Table 1 presents the main results across all three evaluation settings.

RT-Pose (Indoor, Pose Estimation). Table 1 (left) reports indoor pose estimation on the full RT-Pose dataset. With pre-training, RadTok reaches MPJPE = 111.4 mm and ABS = 198.1 mm, outperforming the HRNet3D baseline (120.7 / 240.9) by 7.7 % and 17.8 % respectively. The larger ABS gap indicates that tokenization disentangles target structure from scene-specific clutter that otherwise biases global position across sequences. Among representation-learning methods, Radical Hao et al. (2024) reaches the second-best MPJPE (112.7) since contrastive invariance filters part of the noise, though it does not model polar geometry and limits fine-grained pose recovery. GRT Huang et al. (2025) reconstructs the full tensor, which preserves structure but fits noise as well as signal, leaving ABS at 414.2. Occ-MAE Min et al. (2024) performs worst (131.7 / 606.9) because masked reconstruction was designed for sparse LiDAR voxels where empty space is uninformative, while every cell of a dense radar tensor carries returns and the masked targets reduce to predicting noise.

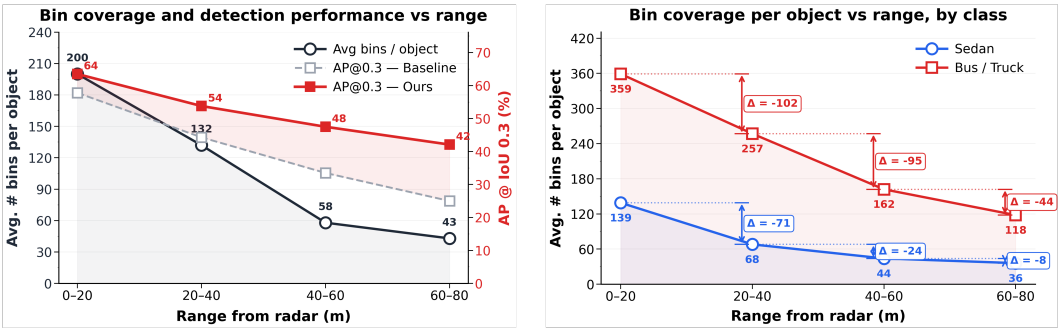

Figure 6: Average number of bins (grid cells) occupied by all objects and the performance in AP@0.3 with respect to distance from radar (a), and average number of bins occupied with respect to specific object categories, such as Sedan and Bus/Truck in K-Radar's object detection task. The average number of occupied bins for objects decreases even for the same object, suggesting a clear problem with non-uniform resolution radar input. The more rapid drop in the average bins for larger vehicles, such as Bus/Truck, compared to Sedan, is associated with the comparatively poor performance of those categories. RadTok (Ours) mitigates the problem through resolution-aware sampling.

Table 1: Cross-task comparison on three radar benchmarks. Lower is better for MPJPE / ABS (↓). Higher is better for AP / IoU (↑). Best per column in **bold**. Top three per column highlighted as 1st, 2nd, 3rd. All results are averaged over 5 random seeds.

Method	Pre-Train	**Human Pose Est.** (Indoor) MPJPE ↓	ABS ↓	**3D Det.** (Outdoor) Sedan ↑	Bus ↑	**3D Occ.** (Outdoor) mIoU ↑	FG IoU ↑	BG IoU ↑
Task-Specific Method								
Baseline	–	120.7	240.9	44.2	36.1	22.6	15.8	29.4
Representation Learning Method								
Occ-MAE Min et al. (2024)	✗	134.1	638.9	39.1	7.7	22.2	14.6	29.1
	✓	131.7	606.9	45.6	12.0	22.2	14.7	29.1
Radical Hao et al. (2024)	✗	115.0	227.7	45.4	16.3	23.1	16.3	30.3
	✓	112.7	213.4	50.9	24.3	24.2	17.2	**32.0**
GRT Huang et al. (2025)	✗	122.4	455.2	52.6	41.7	22.8	16.0	30.0
	✓	119.9	414.2	54.7	43.8	23.4	16.3	30.5
RadTok (Ours)	✗	113.7	204.3	53.5	45.5	24.6	17.1	31.7
	✓	**111.4**	**198.1**	**54.9**	**46.3**	**24.8**	**17.4**	**32.0**

K-Radar (Outdoor, 3D Object Detection). Table 1 (middle) reports bbox AP on K-Radar for Sedan and Bus/Truck classes. With pre-training, RadTok achieves 54.9 AP for Sedan and 46.3 for Bus/Truck, improving the RTN baseline (44.2 / 36.1) by 10.7 and 10.2 AP respectively. The gap is most pronounced on Bus/Truck, which contains larger objects often observed at long range where non-uniform resolution is most severe (Fig. 6). Radical Hao et al. (2024) reaches only 24.3 AP on Bus/Truck since contrastive views from Doppler perturbations do not address the geometric mismatch between coarse cross-range cells at distance and the uniform-cell assumption of the detection backbone. GRT Huang et al. (2025) preserves spatial structure via reconstruction and yields a stronger 43.8 AP on Bus/Truck, but does not denoise, limiting Sedan AP to 54.7. Occ-MAE Min et al. (2024) fails on Bus/Truck (12.0 AP) for the same reason as on indoor pose. RadTok improves both classes simultaneously because tokenization filters incoherent noise (mostly benefiting the smaller Sedan class) and resolution-aware spatialization compensates for the non-uniform polar grid (mostly benefiting Bus/Truck at far range).

K-Radar (Outdoor, Occupancy Grid Prediction). Table 1 (right) reports occupancy prediction following RadarOcc Ding et al. (2024). RadTok with pre-training achieves the best mIoU, FG IoU, and BG IoU (24.8, 17.4, 32.0), improving the task baseline (22.6, 15.8, 29.4) by 2.2, 1.6, and 2.6 respectively. Radical Hao et al. (2024) matches RadTok on BG (32.0) since contrastive invariance helps separate static background from transient noise, but lags by 0.2 on FG, indicating that contrastive views alone do not preserve fine target structure. GRT Huang et al. (2025) and

Table 2: Component ablation on RT-Pose.

LFQ	ToHR	Blur	$\mathcal{L}_{\text{ent}}$	ABS↓
				240.9
✓				220.4
✓	✓			211.3
✓	✓	✓		201.7
✓	✓	✓	✓	**198.1**

Table 3: Resampling factor on Indoor (RT-Pose) and Outdoor (K-Radar Det.).

$f_{\max}$	Indoor↓	Outdoor↑
$1\times$	240.9	44.2
$2\times$	211.1	51.3
$4\times$	**198.1**	54.9
$6\times$	201.6	**55.6**

Table 4: Quantization method on K-Radar Det.

Quantization	Sedan↑	Bus↑
None (AE)	46.6	38.1
VQ-VAE	50.2	44.2
LFQ (32 bit)	54.6	45.7
LFQ (64 bit)	**54.9**	**46.3**

Occ-MAE Min et al. (2024) stay near or below the baseline, suggesting that LiDAR-style masked reconstruction does not disentangle foreground from background more effectively than the raw input.

Across all three tasks, RadTok preserves both fine-grained structure and target-versus-background separability, two complementary properties that none of the compared methods achieves jointly, as also shown qualitatively in Fig. 5. As a result, RadTok is the only evaluated representation-learning method that consistently improves over the task-specific baseline across all three settings.

4.5 Ablation Study

Pipeline components. Table 2 ablates the key components of our pipeline on the RT-Pose full dataset. Each row enables one additional component, isolating its contribution.

Effect of components. Tokenization alone reduces ABS from 240.9 to 220.4 mm, an 8.5 % reduction over the baseline. The discrete bottleneck acts as an information filter, where radar returns with consistent spatial structure produce reproducible codes while transient noise and multipath artifacts that lack reproducibility are suppressed. Adding resampling brings ABS to 211.3, and the learned blur further to 201.7, completing the resolution-aware pipeline that compensates for the polar grid's range-dependent cell area and recovers fine-grained structure the raw input lacks. The entropy term yields the final 198.1 by encouraging balanced codebook utilization.

Effect of pre-training. Pre-training improves RadTok, as shown Table 1, e.g., ABS from 204.3 to 198.1 and Sedan AP from 53.5 to 54.9. The pre-training phase uses the self-supervised degradation objective as a curriculum to learn meaningful quantization codes without task supervision. The fine-tuning phase then adapts them to the task while the tokenization loss keeps them close to this structure. RadTok converges slower than the task baseline in early epochs but reaches better final performance, as pre-training provides a strong initialization before the model is optimized jointly.

Effect of resampling factor. Table 3 studies the resampling factor $f_{\max}$ at four settings. Indoor performance peaks at $4\times$ with ABS = 198.1 mm, while outdoor detection peaks at $6\times$ with Sedan AP = 55.6, since higher factors compensate for more polar-grid distortion and benefit the long-range outdoor setting more than the short-range indoor one. Beyond a point, the upsampled rows interpolate from a fixed input resolution and contribute diminishing returns.

Effect of quantization method. Table 4 compares quantization strategies. The plain autoencoder (AE) performs worst (46.6 / 38.1), confirming that discretization is essential for noise filtering. VQ-VAE improves over AE (50.2 / 44.2) but lags LFQ since the codebook can collapse onto a small subset of entries. 64-bit LFQ performs best (54.9 / 46.3), benefiting from a larger effective codebook size without the collapse problem. Additional token and computational-cost results are in A.1.

5 Conclusion

In this paper, we present RadTok, a task-agnostic radar representation that addresses two intrinsic challenges of radar perception, namely non-uniform spatial resolution and inconsistent ghost echoes from sidelobe and multipath leakage. RadTok combines discrete tokenization for noise filtering with resolution-aware spatialization, supervised end-to-end via inverse degradation without paired clean data. Across three diverse downstream tasks, RadTok consistently improves over task-specific baselines and recent radar representation learning methods while remaining compatible with existing perception backbones. The current resampling factor bounds rather than fully equalizes the variation in physical cell area. Extending RadTok to substantially different sensor configurations is left for future work.

AI USE STATEMENT

In this work, we used generative AI tools to retrieve BibTeX entries and draft parts of the paper. We have not used generative AI tools for other tasks indicated in the ICLR 2027 AI Policy for Authors. Specifically, we provided the AI with several titles of relevant works we were aware of and asked it to find the corresponding BibTeX entries for citations. After retrieval, we double-checked that all retrieved BibTeX entries were correct. For drafting, we used the AI tool to draft the related work section first, with instructions for subsections. After drafting, the authors heavily refined it to fit the paper's context.

REFERENCES

Xuyang Bai, Zeyu Hu, Xinge Zhu, Qingqiu Huang, Yilun Chen, Hongbo Fu, and Chiew-Lan Tai. TransFusion: Robust LiDAR-camera fusion for 3D object detection with transformers. In *CVPR*, pp. 1090–1099, 2022.

Geonho Bang, Kwangjin Choi, Jisong Kim, Dongsuk Kum, and Jun Won Choi. RadarDistill: Boosting radar-based object detection performance via knowledge distillation from LiDAR features. In *CVPR*, pp. 15491–15500, 2024.

Yoshua Bengio, Nicholas Léonard, and Aaron Courville. Estimating or propagating gradients through stochastic neurons for conditional computation. *arXiv preprint arXiv:1308.3432*, 2013.

Huiwen Chang, Han Zhang, Lu Jiang, Ce Liu, and William T. Freeman. MaskGIT: Masked generative image transformer. In *CVPR*, pp. 11315–11325, 2022.

Colin Decourt, Rufin VanRullen, Didier Salle, and Thomas Oberlin. Leveraging self-supervised instance contrastive learning for radar object detection. *arXiv preprint arXiv:2402.08427*, 2024.

Fangqiang Ding, Xiangyu Wen, Yunzhou Zhu, Yiming Li, and Chris Xiaoxuan Lu. RadarOcc: Robust 3D occupancy prediction with 4D imaging radar. In *NeurIPS*, 2024.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021.

Patrick Esser, Robin Rombach, and Björn Ommer. Taming transformers for high-resolution image synthesis. In *CVPR*, pp. 12873–12883, 2021.

Yiduo Hao, Sohrab Madani, Junfeng Guan, Mohammed Alloulah, Saurabh Gupta, and Haitham Hassanieh. Bootstrapping autonomous driving radars with self-supervised learning. In *CVPR*, 2024.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *CVPR*, 2022.

Yuan-Hao Ho, Jen-Hao Cheng, Sheng Yao Kuan, Zhongyu Jiang, Wenhao Chai, Hsiang-Wei Huang, Chih-Lung Lin, and Jenq-Neng Hwang. RT-Pose: A 4D radar tensor-based 3D human pose estimation and localization benchmark. In *ECCV*, 2024.

Tianshu Huang, Akarsh Prabhakara, Chuhan Chen, Jay Karhade, Deva Ramanan, Matthew O'Toole, and Anthony Rowe. Towards foundational models for single-chip radar. In *ICCV*, pp. 24655–24665, 2025.

Zengyuan Lai, Jiarui Yang, Songpengcheng Xia, Lizhou Lin, Lan Sun, Renwen Wang, Jianran Liu, Qi Wu, and Ling Pei. RadarLLM: Empowering large language models to understand human motion from millimeter-wave point cloud sequence. In *AAAI*, volume 40, pp. 5791–5799, 2026.

Alex H. Lang, Sourabh Vora, Holger Caesar, Lubing Zhou, Jiong Yang, and Oscar Beijbom. PointPillars: Fast encoders for object detection from point clouds. In *CVPR*, 2019.

Shih-Po Lee, Niraj Prakash Kini, Wen-Hsiao Peng, Ching-Wen Ma, and Jenq-Neng Hwang. HuPR: A benchmark for human pose estimation using millimeter wave radar. In *WACV*, 2023.

Mohan Li, Tao Song, Yanhai Xu, Zhisong Zhou, Guofeng Zhou, and Xulei Liu. A unified-BEV network for joint 3D object detection and map segmentation in complex traffic scenario. SAE Technical Paper 2025-01-7202, SAE International, 2025.

Yiming Li, Zhiding Yu, Christopher Choy, Chaowei Xiao, Jose M. Alvarez, Sanja Fidler, Chen Feng, and Anima Anandkumar. VoxFormer: Sparse voxel transformer for camera-based 3D semantic scene completion. In *CVPR*, 2023.

Zhenxin Li, Shiyi Lan, Jose M. Alvarez, and Zuxuan Wu. BEVNeXt: Reviving dense BEV frameworks for 3D object detection. In *CVPR*, 2024.

Zhijian Liu, Haotian Tang, Alexander Amini, Xinyu Yang, Huizi Mao, Daniela Rus, and Song Han. BEVFusion: Multi-task multi-sensor fusion with unified bird's-eye view representation. In *ICRA*, 2023.

Chen Min, Liang Xiao, Dawei Zhao, Yiming Nie, and Bin Dai. Occupancy-MAE: Self-supervised pre-training large-scale LiDAR point clouds with masked occupancy autoencoders. *IEEE Transactions on Intelligent Vehicles*, 9(7):5150–5162, 2024. doi: 10.1109/TIV.2023.3322409.

Arthur Ouaknine, Alasdair Newson, Julien Rebut, Florence Tupin, and Patrick Pérez. Carrada dataset: Camera and automotive radar with range-angle-doppler annotations. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pp. 5068–5075, 2021. doi: 10.1109/ICPR48806.2021.9413181.

Dong-Hee Paek, Seung-Hyun Kong, and Kevin Tirta Wijaya. K-Radar: 4D radar object detection for autonomous driving in various weather conditions. In *NeurIPS Datasets and Benchmarks Track*, 2022.

Zhijun Pan, Fangqiang Ding, Hantao Zhong, and Chris Xiaoxuan Lu. RaTrack: Moving object detection and tracking with 4D radar point cloud. In *ICRA*, 2024.

Yatian Pang, Wenxiao Wang, Francis E. H. Tay, Wei Liu, Yonghong Tian, and Li Yuan. Masked autoencoders for point cloud self-supervised learning. In *ECCV*, pp. 604–621, 2022.

Yongming Rao, Wenliang Zhao, Benlin Liu, Jiwen Lu, Jie Zhou, and Cho-Jui Hsieh. DynamicViT: Efficient vision transformers with dynamic token sparsification. In *NeurIPS*, 2021.

Ali Razavi, Aaron van den Oord, and Oriol Vinyals. Generating diverse high-fidelity images with VQ-VAE-2. In *NeurIPS*, 2019.

Julien Rebut, Arthur Ouaknine, Waqas Malik, and Patrick Pérez. Raw high-definition radar for multi-task learning. In *CVPR*, pp. 17021–17030, 2022.

Michael S. Ryoo, AJ Piergiovanni, Anurag Arnab, Mostafa Dehghani, and Anelia Angelova. TokenLearner: Adaptive space-time tokenization for videos. In *NeurIPS*, 2021.

Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. In *NeurIPS*, 2017.

Yizhou Wang, Zhongyu Jiang, Xiangyu Gao, Jenq-Neng Hwang, Guanbin Xing, and Hui Liu. RODNet: Radar object detection using cross-modal supervision. In *WACV*, 2021.

Bin Yang, Wenjie Luo, and Raquel Urtasun. PIXOR: Real-time 3D object detection from point clouds. In *CVPR*, 2018.

Bin Yang, Runsheng Guo, Ming Liang, Sergio Casas, and Raquel Urtasun. RadarNet: Exploiting radar for robust perception of dynamic objects. In *ECCV*, 2020.

Tianwei Yin, Xingyi Zhou, and Philipp Krähenbühl. Center-based 3D object detection and tracking. In *CVPR*, pp. 11784–11793, 2021.

Lijun Yu, José Lezama, Nitesh Bharadwaj Gundavarapu, Luca Versari, Kihyuk Sohn, David Minnen, Yong Cheng, Agrim Gupta, Xiuye Gu, Alexander G. Hauptmann, Boqing Gong, Ming-Hsuan Yang, Irfan Essa, David A. Ross, and Lu Jiang. Language model beats diffusion – tokenizer is key to visual generation. In *ICLR*, 2024.

Xumin Yu, Lulu Tang, Yongming Rao, Tiejun Huang, Jie Zhou, and Jiwen Lu. Point-BERT: Pre-training 3D point cloud transformers with masked point modeling. In *CVPR*, 2022.

Renrui Zhang, Ziyu Guo, Peng Gao, Rongyao Fang, Bin Zhao, Dong Wang, Yu Qiao, and Hongsheng Li. Point-M2AE: Multi-scale masked autoencoders for hierarchical point cloud pre-training. In *NeurIPS*, 2022.

Yunpeng Zhang, Zheng Zhu, and Dalong Du. OccFormer: Dual-path transformer for vision-based 3D semantic occupancy prediction. In *ICCV*, pp. 9433–9443, 2023.

Lingjun Zhao, Jingyu Song, and Katherine A. Skinner. CRKD: Enhanced camera-radar object detection with cross-modality knowledge distillation. In *CVPR*, pp. 15470–15480, 2024.

A Appendix

A.1 Additional Results

Table 5: Tokens as direct input to the detection head on K-Radar detection (Sedan AP, 3 seeds). The task baseline keeps its task-specific backbone.

	Task baseline	Occ-MAE	GRT	RadTok
Mean AP	44.20	40.63	46.77	**47.83**
Range	–	38.7–42.3	46.4–47.0	47.1–48.3

Token as direct input to task head. In the main experiments, the tokens are decoded into a radar feature map so that existing task pipelines can be used without any modification. To test the usability of the tokens themselves, we freeze the pre-trained encoder of each method and train only the same lightweight detection head on its output, with 3 seeds for K-Radar's outdoor object detection task. For RadTok, the input to the head is the token map $\mathbf{S}^q$, and for GRT and Occupancy-MAE, it is the output of their frozen encoders. For all methods, we then add a linear projection that maps each representation to the same embedding dimension before the head. RadTok and GRT are the only two methods above the task baseline (44.20 AP), reaching 47.83 AP and 46.77 AP, despite the fact that the baseline keeps its own task-specific optimized backbone (Table 5). RadTok is ahead of GRT in every seed (47.1–48.3 vs. 46.4–47.0 AP). The absolute AP is lower than the full pipeline, as expected with a shallow head and the absence of the optimized backbone from the task baseline.

Table 6: Computational cost on K-Radar detection (single NVIDIA A10, batch size 1). Added parameters are for RadTok only, relative to the downstream model. Memory and latency are end-to-end.

Configuration	Added params.	Peak memory	Latency	Throughput	vs. 10 FPS
Raw input	–	217 MiB	5.8 ms	171.2 FPS	17.1×
RadTok, $f_{\max}$=1	0.33 M (+3.6%)	279 MiB	11.9 ms	84.2 FPS	8.4×
RadTok, $f_{\max}$=2	0.41 M (+4.6%)	498 MiB	18.4 ms	54.3 FPS	5.4×
RadTok, $f_{\max}$=4	0.59 M (+6.5%)	980 MiB	36.6 ms	27.3 FPS	2.7×
RadTok, $f_{\max}$=6	0.76 M (+8.4%)	1457 MiB	51.5 ms	19.4 FPS	1.9×

Computational cost. Table 6 reports the cost on K-Radar detection, measured on a single NVIDIA A10 at batch size 1. We average 100 times of forward passes. At $f_{\max}$=4, RadTok adds 0.59 M parameters (6.5 % of the downstream model) and runs end-to-end at 36.6 ms per frame (27.3 FPS). This is 2.7× faster than the 10 FPS acquisition rate of the radar. Most of the additional cost in time comes from the upsampled azimuth axis that the downstream model has to process, so $f_{\max}$ controls the trade-off between accuracy and cost. Nevertheless, all settings run faster than the acquisition rate of the radar used for the dataset. However, these results are from a datacenter level GPU (NVIDIA A10), and embedded hardware may need further optimization.