Indoor Multi-View Radar Object Detection via 3D Bounding Box Diffusion

Ryoma Yataka[1,2*], **Pu (Perry) Wang**[2], **Petros Boufounos**[2], **Ryuhei Takahashi**[1]

[1]Information Technology R&D Center (ITC), Mitsubishi Electric Corporation
[2]Mitsubishi Electric Research Laboratories (MERL)

Abstract

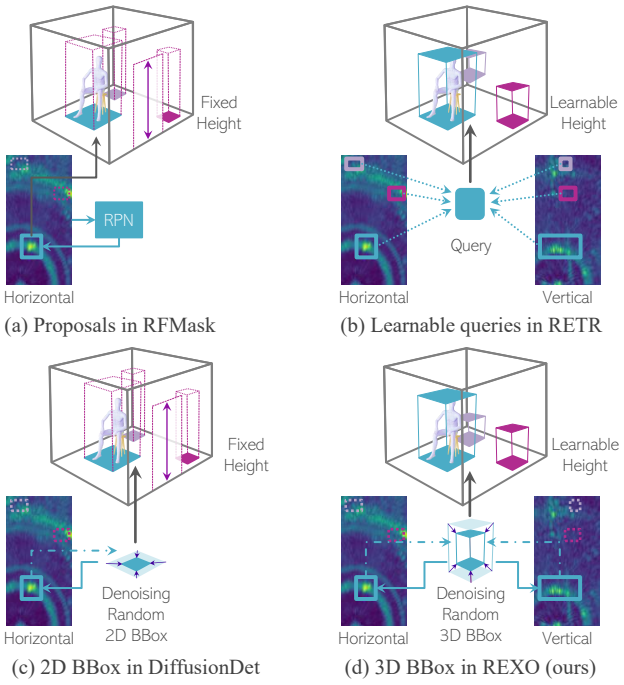
Multi-view indoor radar perception has drawn attention due to its cost-effectiveness and low privacy risks. Existing methods often rely on implicit cross-view radar feature association, such as proposal pairing in RFMask or query-to-feature cross-attention in RETR, which can lead to ambiguous feature matches and degraded detection in complex indoor scenes. To address these limitations, we propose **REXO** (multi-view Radar object dEtection with 3D bounding boX diffusiOn), which lifts the 2D bounding box (BBox) diffusion process of DiffusionDet into the 3D radar space. REXO utilizes these noisy 3D BBoxes to guide an explicit cross-view radar feature association, enhancing the cross-view radar-conditioned denoising process. By accounting for prior knowledge that the person is in contact with the ground, REXO reduces the number of diffusion parameters by determining them from this prior. Evaluated on two open indoor radar datasets, our approach surpasses state-of-the-art methods by a margin of +4.22 AP on the HIBER dataset and +11.02 AP on the MMVR dataset.

Code —
https://github.com/merlresearch/radar-bbox-diffusion
Extended version — https://arxiv.org/abs/2511.17806

1 Introduction

Radar perception has received increasing attention due to its robustness in low-light, adversarial weather, and hazardous conditions (e.g., smoke) (Paek et al. 2022; Yao et al. 2024; Lu et al. 2020; Sun, Petropulu, and Poor 2020; Pandharipande et al. 2023; Skog et al. 2024). Depending on the application, operational specifications, and downstream tasks, radar data can be represented in different forms such as sparse detection points (Zhao et al. 2017; Sengupta et al. 2020; Yang et al. 2023b), reflection heatmaps (Adib et al. 2015; Zhao et al. 2018a; Wu et al. 2023), Doppler signatures, and raw analog-to-digital converter (ADC) data, each with unique characteristics and feature granularity.

For indoor radar perception, single-view and multi-view heatmaps that combine horizontal (depth-horizontal) and vertical (depth-vertical) projections enable object detection, pose estimation, and segmentation on a 2D image plane (Zhao et al. 2018a; Lee et al. 2023; Wu et al. 2023; Rahman et al. 2024). RF-Pose (Zhao et al. 2018a,b) first fused both views with a convolutional autoencoder to regress 2D human keypoints. RFMask (Wu et al. 2023) grafted Faster R-CNN (Ren et al. 2017) onto horizontal heatmaps: its region-proposal network produces horizontal candidates that are paired with fixed-height vertical windows to avoid exhaustive cross-view association (Fig. 1 (a)). More re-



(a) Proposals in RFMask (b) Learnable queries in RETR

(c) 2D BBox in DiffusionDet (d) 3D BBox in REXO (ours)

Figure 1: (a) RFMask (Wu et al. 2023) generates horizontal-view proposals with fixed-height vertical boxes; (b) RETR (Yataka et al. 2024) implicitly links queries to cross-view features via decoder cross-attention; (c) DiffusionDet (Chen et al. 2023b) adapted to horizontal radar allows 2D denoising but needs extra pairing with fixed-height vertical boxes; (d) REXO (**ours**) performs diffusion directly in 3D radar space for simple, explicit cross-view association.

*The work was done as a visiting scientist of MERL from ITC.

cently, the radar detection transformer (RETR) (Yataka et al. 2024) adopted the DETR (Carion et al. 2020). Decoder queries simultaneously attend to both views through cross-attention (Fig. 1 (b)) and are directly regressed to 3D bounding boxes (BBoxes) that are classified as person or background.

On the other hand, image-based object detection has been redefined as a generative denoising process, where random noisy 2D BBoxes are iteratively refined through a diffusion denoising process to yield final clean BBox predictions (Chen et al. 2023b). Referred to as DiffusionDet, it decouples training and inference, and generally surpasses query-based detectors. When ported to horizontal radar heatmaps (Fig. 1 (c)), it denoises 2D boxes but still requires the fixed-height vertical pairing used by RFMask.

We therefore *lift* the diffusion procedure from a 2D plane (image or horizontal radar view) in DiffusionDet to the full 3D radar space, as illustrated in Fig. 1 (d). This simple lifting facilitates cross-view radar feature association and radar-conditioned BBox denoising, while enabling the integration of geometry-aware loss functions and prior constraints on the 3D BBox. Consequently, we introduce the proposed framework as **Radar object dEtection with 3D bounding boX diffusiOn (REXO)** with the following contributions:

1. **2D-to-3D Lifting with Explicit Cross-View Association**: At each diffusion timestep, noisy 3D BBoxes are projected onto every radar view, and RoI-aligned crops supply view-specific features. This BBox-guided association grows *linearly* with the number of views, whereas proposal- or query-based schemes grow quadratically.
2. **Cross-View Radar-Conditioned BBox Detection**: While the cross-view feature association is simplified due to the 2D-to-3D lifting, the denoising process may be more challenging. In turn, the associated radar features are used as conditioning to alleviate the more challenging 3D BBox denoising. To the best of our knowledge, REXO is the first diffusion model in the radar perception field conditioned on multi-view radar.
3. **Ground-Level Constraint**: By using prior knowledge that the person is in contact with the ground, the parameters of the 3D BBox are reduced. Based on this, each noise-free 3D BBox preserves geometric constraints in the image plane to be transformed.

We demonstrate the effectiveness of our contributions through evaluations on two open radar datasets.

2 Related Work

Radar-based Object Detection: Learning-based methods have advanced radar detection over traditional model-based approaches (Kay 1998), benefiting from open large-scale radar point cloud datasets like nuScenes (Caesar et al. 2020), Oxford RobotCar (Barnes et al. 2020), and RADIATE (Sheeny et al. 2021). Image-based and point/voxel-based backbones (He et al. 2016; Shi, Li, and Ma 2022) extract semantic features from radar detection points, generate region proposals, and localize objects. High-resolution heatmaps (e.g., K-Radar (Paek et al. 2022), HIBER (Wu et al. 2023), RT-Pose (Ho et al. 2024), MMVR (Rahman et al. 2024)) and raw ADC data (Yang et al. 2023a) have also been leveraged by previously mentioned RF-Pose (Zhao et al. 2018a), RFMask (Wu et al. 2023), and RETR (Yataka et al. 2024). CubeLearn (Zhao et al. 2023) replaces Fourier transforms with learnable modules for an end-to-end radar pipeline, while RAMP-CNN (Gao et al. 2021) enhances range-angle feature extraction via Doppler cues. More recently, diffusion models have been explored for radar applications (Zhang et al. 2024; Luan et al. 2024; Chi et al. 2024; Fan et al. 2024; Wu et al. 2024). Most efforts, e.g., Radar-Diffusion (Zhang et al. 2024; Luan et al. 2024) and DiffRadar (Wu et al. 2024), focus on reconstructing LiDAR-like point clouds from low-resolution radar data, while mmDiff (Fan et al. 2024) estimates and refines pose keypoints from sparse radar points via diffusion process.

Diffusion-based Object Detection: Diffusion models (Song, Meng, and Ermon 2021; Song and Ermon 2019; Rombach et al. 2022; Song et al. 2023) have shown impressive results in tasks such as image and video generation (Ho et al. 2022; Blattmann et al. 2023) and multi-view synthesis (Chen et al. 2023a; Yu et al. 2023). For perception tasks, DiffusionDet (Chen et al. 2023b) first reformulates object detection as a generative denoising process and proposes to model the 2D BBoxes as random parameters in the diffusion process. Diffusion-SS3D (Ho et al. 2023) proposes a diffusion-based detector to enhance the quality of pseudo-labels in semi-supervised 3D object detection by integrating it into a teacher-student framework. CLIFF (Li et al. 2024) further leverages language models to enhance diffusion-based models for open-vocabulary object detection. Diffusion models are also considered for 3D object detection (XU et al. 2024) in the context of LiDAR-Camera fusion (Xiang, Dräger, and Zhang 2023) and other tasks such as pose estimation (Tan et al. 2024) and semantic segmentation (Gu, Chen, and Xu 2024; Amit et al. 2022).

3 Preliminary

Multi-View Radar Heatmaps derive from raw data captured by horizontal and vertical radar arrays, where sampling reflected pulses across each array builds a 3D data cube of ADC samples, pulse samples and array elements (Fig. 2). A 3D FFT transforms each cube into radar spectra along range, Doppler, and angle (azimuth or elevation). Integrating over Doppler yields 2D polar heatmaps (range–azimuth and range–elevation), which are then mapped to Cartesian coordinates. The resulting heatmaps for frame m are denoted $\boldsymbol{Y}_{\texttt{hor}}(m) \in \mathcal{R}^{W\times D}$ and $\boldsymbol{Y}_{\texttt{ver}}(m) \in \mathcal{R}^{H\times D}$. Stacking M consecutive frames gives $\boldsymbol{Y}_{\texttt{hor}} \in \mathcal{R}^{M\times W\times D}$ and $\boldsymbol{Y}_{\texttt{ver}} \in \mathcal{R}^{M\times H\times D}$ for temporal modeling.

Diffusion Models such as the denoising diffusion probabilistic model (DDPM) (Ho, Jain, and Abbeel 2020) and the denoising diffusion implicit model (DDIM) (Song, Meng, and Ermon 2021), define Markovian or non-Markovian forward processes by gradually adding noise to samples $\boldsymbol{x}_0$, e.g., image pixels,

$$q\left(\boldsymbol{x}_t \mid \boldsymbol{x}_0\right) = \mathcal{N}\left(\boldsymbol{x}_t; \sqrt{\bar{\alpha}_t}\boldsymbol{x}_0, \left(1-\bar{\alpha}_t\right)\boldsymbol{I}\right), \qquad (1)$$

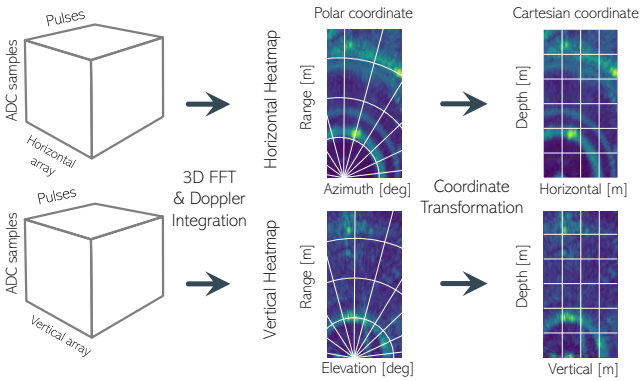


Figure 2: Generation of multi-view heatmaps from raw data.

where $t \in \{0, \dots, T\}$ and $\bar{\alpha}_t = \prod_{s=0}^{t}(1-\beta_s)$ with β_s denoting the noise variance schedule. At time t, $\boldsymbol{x}_t = \sqrt{\bar{\alpha}_t}\boldsymbol{x}_0 + \sqrt{(1-\bar{\alpha}_t)}\boldsymbol{\epsilon}$ with $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$.

During training, a noise prediction network is trained to estimate $\boldsymbol{\epsilon}$ from $\boldsymbol{x}_t$ and time index t by minimizing $\min_\theta \|\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, t) - \boldsymbol{\epsilon}\|^2$, where θ represents the trainable weights. During inference, a random $\boldsymbol{x}_T$ is drawn from the standard Gaussian distribution and iteratively denoised using the trained $\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, t)$ in reverse time: $\boldsymbol{x}_T \to \cdots \to \boldsymbol{x}_t \to \boldsymbol{x}_{t-1} \to \cdots \to \boldsymbol{x}_0$. Sampling strategies such as DDPM (Ho, Jain, and Abbeel 2020)

$$\boldsymbol{x}_{t-1} = \left(\boldsymbol{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, t)\right)/\sqrt{\alpha_t} + \sigma_t\boldsymbol{\epsilon}_t, \quad (2)$$

with $\boldsymbol{\epsilon}_t \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$, and DDIM (Song, Meng, and Ermon 2021) can be used to trade off between quality and speed.

4 REXO: BBox Diffusion in 3D Radar Space

DiffusionDet (Chen et al. 2023b) reformulates object detection as a denoising diffusion process, treating $\boldsymbol{x}_t$ as 2D BBox parameters instead of image pixels. As shown in Fig. 3, we extend this to multi-view radar by lifting $\boldsymbol{x}_t$ to 3D BBoxes in radar coordinates: $\boldsymbol{x}_t = \left\{c_x^t, c_y^t, c_z^t, w^t, h^t, d^t\right\}^\top \in \mathbb{R}^6$, where (c_x^t, c_y^t, c_z^t) defines the center and (w^t, h^t, d^t) the size at time t in the Cartesian $\{\texttt{horizontal}, \texttt{vertical}, \texttt{depth}\}$ space. Conditioned on radar heatmaps $\{\boldsymbol{Y}_{\texttt{hor}}, \boldsymbol{Y}_{\texttt{ver}}\}$, REXO performs 3D BBox diffusion in two phases (Fig. 3): 1) a **forward process** that adds noise to ground-truth (GT) BBoxes $\boldsymbol{x}_0$ to produce random $\boldsymbol{x}_T$ during training, and 2) a **reverse process** that denoises random $\boldsymbol{x}_T$ to estimate noise-free $\hat{\boldsymbol{x}}_0$ during inference. The denoised BBoxes are also projected to the 2D image plane for supervision in both radar and image domains. We describe REXO in four parts: training, inference, ground-level constraint and loss.

4.1 Training

We describe REXO training, as illustrated in Fig. 4.

Backbone: Taking the two radar heatmaps $\boldsymbol{Y}_{\texttt{hor}}$ and $\boldsymbol{Y}_{\texttt{ver}}$ as inputs, a backbone network (e.g., ResNet (He et al. 2016)) generates horizontal-view and vertical-view radar feature maps separately: $\boldsymbol{Z}_{\texttt{hor}} = \texttt{backbone}(\boldsymbol{Y}_{\texttt{hor}})$ and $\boldsymbol{Z}_{\texttt{ver}} = \texttt{backbone}(\boldsymbol{Y}_{\texttt{ver}})$, where learnable parameters in the backbone are shared across both views. Each feature map is generated as L multi-scale feature maps in $\mathbb{R}^{C\times\frac{W}{sl}\times\frac{D}{sl}}$ or $\mathbb{R}^{C\times\frac{H}{sl}\times\frac{D}{sl}}$ by using feature pyramid network (Lin et al. 2017) where C, s and $l \in \{1, \cdots, L\}$ represent the number of channels, downsampling ratio over the spatial dimension and the pyramid level, respectively.

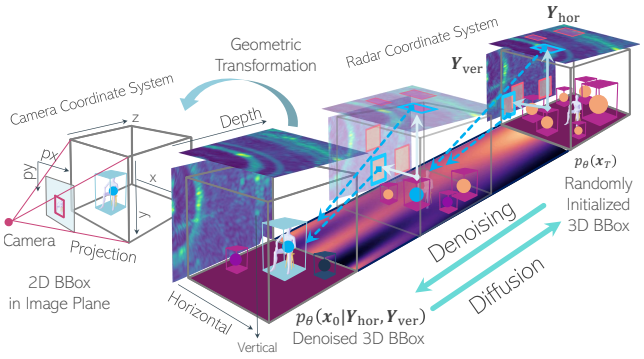


Figure 3: REXO: 1) 3D BBox diffusion process in the radar space; 2) Geometric transformation and 3D-to-2D projection onto the image plane for geometry-aware supervision.

Initialization of $\boldsymbol{x}_0$ and Forward Process to $\boldsymbol{x}_t$: For a given number of BBoxes $N_{\texttt{train}}$ to be detected, $\boldsymbol{x}_0$ is simply initialized by the 3D BBox GT in the radar space $\boldsymbol{x}_{\texttt{radar}} = \{c_x, c_y, c_z, w, h, d\}^\top \in \mathbb{R}^6$ and padded with random 3D BBoxes $\boldsymbol{x}_{\texttt{rand}} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_6)$ if $N_{\texttt{train}} > N_{\texttt{GT}}$. The diffused 3D BBox $\boldsymbol{x}_t$ at time t can be generated as

$$\boldsymbol{x}_t = \sqrt{\bar{\alpha}_t}\boldsymbol{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}, \quad (3)$$

where $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_6)$ and $\bar{\alpha}_t$ is defined in Section 3.

Cross-View Radar-Conditioned BBox Detector $\texttt{DenoisingDet}_\theta$ includes explicit cross-view feature association and radar-conditioned 3D BBox detector. 1) Explicit cross-view feature association: Given the noisy 3D BBox $\boldsymbol{x}_t$ in (3), the $\boldsymbol{x}_t$-guided cross-view feature association first projects $\boldsymbol{x}_t$ onto the two radar views, resulting in two 2D BBoxes (purple solid lines of Fig. 1 (d)),

$$\boldsymbol{x}_{t,\texttt{hor}} = \{c_x^t, c_z^t, w^t, d^t\}^\top, \boldsymbol{x}_{t,\texttt{ver}} = \{c_y^t, c_z^t, h^t, d^t\}^\top, \quad (4)$$

and then crops out the cross-view 2D radar features

$$\begin{aligned}\boldsymbol{Z}_{\texttt{hor}}^{\texttt{crop}} &= \texttt{RoIAlign}(\boldsymbol{Z}_{\texttt{hor}}, \boldsymbol{x}_{t,\texttt{hor}}) \in \mathbb{R}^{C\times r\times r},\\ \boldsymbol{Z}_{\texttt{ver}}^{\texttt{crop}} &= \texttt{RoIAlign}(\boldsymbol{Z}_{\texttt{ver}}, \boldsymbol{x}_{t,\texttt{ver}}) \in \mathbb{R}^{C\times r\times r}, \quad (5)\end{aligned}$$

via a standard ROIAlign operation (He et al. 2017), where r denotes a fixed spatial resolution, e.g., $r = 7$. At time t, this process yields $N_{\texttt{train}}$ pairs of associated radar features

$$\boldsymbol{Z}_{\texttt{radar}}^{\texttt{crop}} = \{\boldsymbol{Z}_{\texttt{hor}}^{\texttt{crop}}, \boldsymbol{Z}_{\texttt{ver}}^{\texttt{crop}}\} \in \mathbb{R}^{C\times r\times 2r}, \quad (6)$$

each corresponding to a noisy 3D BBox $\boldsymbol{x}_t$.

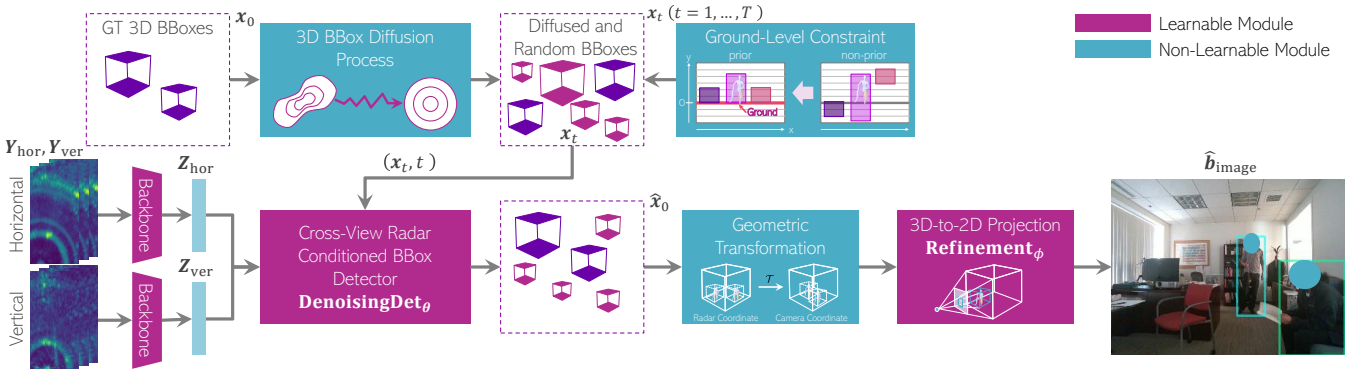


Figure 4: **REXO training:** 1) A shared backbone extracts horizontal/vertical radar features $\{\boldsymbol{Z}_{\texttt{hor}}, \boldsymbol{Z}_{\texttt{ver}}\}$; 2) Ground-truth 3D BBoxes $\boldsymbol{x}_0$ are diffused to noisy $\boldsymbol{x}_t$; 3) $\boldsymbol{x}_t$ is grounded using a ground-level constraint; 4) $\texttt{DenoisingDet}_\theta$ projects $\boldsymbol{x}_t$ onto both views and uses the aligned features to recover $\hat{\boldsymbol{x}}_0$; 5) A radar-to-camera transform and 3D-to-2D projection yield image BBoxes $\hat{\boldsymbol{b}}_{\texttt{image}}$, enabling geometry-aware supervision in radar space and image plane.

2) Radar-conditioned 3D BBox detector: Conditioned on $\boldsymbol{Z}^{\texttt{crop}}_{\texttt{radar}}$, a BBox detector with learnable weights θ is trained to estimate the BBox $\hat{\boldsymbol{x}}_0$ and the class scores $\hat{\boldsymbol{p}}$ as

$$\{\hat{\boldsymbol{x}}_0, \hat{\boldsymbol{p}}\} = \texttt{BBoxDet}_\theta\left(\boldsymbol{Z}^{\texttt{crop}}_{\texttt{radar}}, t\right), \tag{7}$$

where t specifies the timestep embedding. In our indoor setting, we use a two-class softmax over $\{$*person, background*$\}$. The class-head can extend to C classes (including background) by using a C-way softmax with cross-entropy.

Grouping all the steps from (4) to (7) results in the $\texttt{DenoisingDet}_\theta$ module of Fig. 4:

$$\{\hat{\boldsymbol{x}}_0, \hat{\boldsymbol{p}}\} = \texttt{DenoisingDet}_\theta\left(\boldsymbol{x}_t, t, \boldsymbol{Z}_{\texttt{hor}}, \boldsymbol{Z}_{\texttt{ver}}\right), \tag{8}$$

where all trainable weights θ are inherited from the BBox detector $\texttt{BBoxDet}_\theta$.

3D-to-2D Projection with Learnable Refinement. REXO further projects $\hat{\boldsymbol{x}}_0$ in (8) into the 2D image plane through the 3D camera coordinate system via a calibrated geometric transformation $\mathcal{T}$. By setting $\hat{\boldsymbol{x}}_{\texttt{radar}} = \hat{\boldsymbol{x}}_0$, we convert each of the 8 corners of the corresponding 3D BBox $\hat{\boldsymbol{x}}_{\texttt{radar}}$ using

$$\boldsymbol{x}^i_{\texttt{camera}} = \boldsymbol{R}\hat{\boldsymbol{x}}^i_{\texttt{radar}} + \boldsymbol{v}, \quad i = 1, 2, \cdots, 8, \tag{9}$$

where $\hat{\boldsymbol{x}}^i_{\texttt{radar}}$ is the i-th corner of $\hat{\boldsymbol{x}}_{\texttt{radar}}$, $\boldsymbol{R}$ is the calibrated 3D rotation matrix, and $\boldsymbol{v}$ is the calibrated translation vector. Each 3D corner $\boldsymbol{x}^i_{\texttt{camera}}$ is projected to the image plane through the calibrated pinhole model, and the extrema of the eight projected points yield the initial box $\boldsymbol{b}_{\texttt{init}}$

$$\boldsymbol{b}_{\texttt{init}} = \left\{\bar{c}_x, \bar{c}_y, \bar{w}, \bar{h}\right\}^\top = \texttt{proj}_{\texttt{init}}\left(\boldsymbol{x}_{\texttt{camera}}\right). \tag{10}$$

Since $\boldsymbol{b}_{\texttt{init}}$ systematically overshoots the ground-truth extent, we attach a refinement module with learnable parameter $\boldsymbol{\phi}$ to obtain the offset:

$$\Delta\boldsymbol{b} = \left\{\Delta\bar{x}, \Delta\bar{y}, \Delta\bar{w}, \Delta\bar{h}\right\}^\top = \texttt{Refinement}_{\boldsymbol{\phi}}\left(\boldsymbol{f}\right), \tag{11}$$

where $\boldsymbol{f} = \texttt{Predictor}\left(\texttt{e}_t, \boldsymbol{Z}^{\texttt{crop}}_{\texttt{radar}}\right)$ is the time-dependent feature. $\texttt{e}_t$ denotes the timestep embedding (Ho, Jain, and Abbeel 2020) and $\texttt{Predictor}$ denotes the time-dependent predictor (Chen et al. 2023b) with the radar feature and the embedding. Applying these offsets produces the final image-plane box $\hat{\boldsymbol{b}}_{\texttt{image}}$, achieving tighter alignment without sacrificing geometric consistency.

$$\hat{\boldsymbol{b}}_{\texttt{image}} = \{\bar{c}_x + \bar{w}\Delta\bar{x}, \bar{c}_y + \bar{h}\Delta\bar{y}, e^{\Delta\bar{w}}\bar{w}, e^{\Delta\bar{h}}\bar{h}\}^\top. \tag{12}$$

4.2 Inference

REXO infers objects by reversing the diffusion process. Given a target count N, we sample random 3D boxes $\boldsymbol{x}_T \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_6)$ in the radar coordinate system at $t = T$ and denoise them down to $t = 1$.

Denoising Process in 3D Radar Space: With $\boldsymbol{x}_t$ and radar features $\{\boldsymbol{Z}_{\texttt{hor}}, \boldsymbol{Z}_{\texttt{ver}}\}$, the trained $\texttt{DenoisingDet}_\theta$ in (8) predicts $\hat{\boldsymbol{x}}_0$, giving

$$p_\theta\left(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t, \boldsymbol{Z}_{\texttt{hor}}, \boldsymbol{Z}_{\texttt{ver}}\right) = \mathcal{N}(\sqrt{\alpha_{t-1}}\boldsymbol{x}_0 + \gamma\boldsymbol{\epsilon}^{(t)}_\theta, \sigma_t^2\boldsymbol{I}_6),$$

$$\boldsymbol{x}_{t-1} = \sqrt{\alpha_{t-1}}\hat{\boldsymbol{x}}_0 + \sqrt{1 - \alpha_{t-1} - \sigma_t^2}\cdot\boldsymbol{\epsilon}^{(t)}_\theta + \sigma_t\boldsymbol{\epsilon}_t, \tag{13}$$

where $\boldsymbol{\epsilon}^{(t)}_\theta = \left(\boldsymbol{x}_t - \sqrt{\alpha_t}\hat{\boldsymbol{x}}_0\right)/\sqrt{1 - \alpha_t}$ specifies the direction pointing to the noisy BBox $\boldsymbol{x}_t$ at time t, and $\boldsymbol{\epsilon}_t \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_6)$ represents a random BBox. Note that the denoising step is inherently conditioned on the cross-view radar feature maps via the estimated $\hat{\boldsymbol{x}}_0$ from the $\texttt{DenoisingDet}_\theta$ module.

2D Image Plane BBox Prediction: After the final step, $\boldsymbol{x}_0$ ($= \hat{\boldsymbol{x}}_{\texttt{radar}}$) is converted to image plane boxes $\hat{\boldsymbol{b}}_{\texttt{image}}$ via the radar–to-camera transform in (9) and the 3D-to-2D projection of (12). Boxes whose class scores exceed a threshold are output as detections.

4.3 Ground-Level Constraint

Since the BBoxes are now explicitly defined in the 3D radar coordinate system, it is natural to incorporate prior knowledge as a constraint into the diffusion process. Unlike DiffusionDet and RETR, we enforce the reduced five 3D parameters by grounding with $h^t/2$, allowing 3D and 2D gradients

Method	P1S1			P1S2			P2S1			P2S2		
	AP	AP_{50}	AP_{75}	AP	AP_{50}	AP_{75}	AP	AP_{50}	AP_{75}	AP	AP_{50}	AP_{75}
RFMask	25.53	67.30	15.86	24.46	66.82	11.22	31.37	61.50	27.48	6.03	22.77	0.88
RFMask3D	34.84	69.57	31.74	30.75	76.48	16.23	39.89	80.38	35.35	12.26	37.01	4.34
DETR	35.64	77.59	28.00	28.51	75.90	13.42	29.53	63.08	25.35	9.29	34.69	2.49
RETR	39.62	80.55	33.84	30.16	78.95	15.17	46.75	83.80	46.06	12.45	41.30	4.96
REXO	39.23	73.46	37.83	36.48	87.02	20.51	48.35	85.89	48.38	23.47	64.41	10.44

Table 1: Evaluation on 4 data splits of the MMVR. The gray hatch represents the best performance for each metric.

to flow jointly and guiding the denoising process under strict geometric constraints. This ensures that objects are correctly positioned on the floor, reflecting realistic spatial relationships (see the Ground-Level Constraint in Fig. 4):

$$\boldsymbol{x}_t = \left\{c_x^t, h^t/2, c_z^t, w^t, h^t, d^t\right\}^\top. \tag{14}$$

Using this constrained $\boldsymbol{x}_t$ as in (3), REXO predicts $N_{\texttt{train}}$ 3D BBoxes $\hat{\boldsymbol{x}}_{\texttt{radar}}$ and 2D BBoxes $\hat{\boldsymbol{b}}_{\texttt{image}}$, while preserving geometric consistency.

4.4 Loss Function

To ensure consistency between the radar and image plane representations, we adopt a simplified scheme of the Triplane loss (Yataka et al. 2024) that directly calculates the loss of 3D BBox. REXO employs the Hungarian match cost (Kuhn 1955) with a geometry-aware loss function $\mathcal{L}_{\texttt{box}}^{\texttt{GA}}$ computed in both the 3D and 2D spaces:

$$\mathcal{L}_{\texttt{box}}^{\texttt{GA}} = \lambda_{\texttt{3D}}\mathcal{L}_{\texttt{box}}^{\texttt{3D}}(\boldsymbol{x}_{\texttt{radar}}, \widehat{\boldsymbol{x}}_{\texttt{radar}}) + \lambda_{\texttt{2D}}\mathcal{L}_{\texttt{box}}^{\texttt{2D}}(\boldsymbol{b}_{\texttt{image}}, \widehat{\boldsymbol{b}}_{\texttt{image}}),$$

where the 3D/2D BBox loss is defined as $\mathcal{L}_{\texttt{box}}^{*}(\boldsymbol{x}, \widehat{\boldsymbol{x}}) = \lambda_{\texttt{GIoU}}\mathcal{L}_{\texttt{GIoU}}(\boldsymbol{x}, \widehat{\boldsymbol{x}}) + \lambda_{\texttt{L}_1}\mathcal{L}_{\texttt{L}_1}(\boldsymbol{x}, \widehat{\boldsymbol{x}})$ representing a weighted combination of the generalized intersection over union (GIoU) loss $\mathcal{L}_{\texttt{GIoU}}$ (Rezatofighi et al. 2019) and the ℓ_1 loss $\mathcal{L}_{\texttt{L}_1}$, and the coefficients λ balance the relative contribution of each loss term. REXO determines the optimal assignment $\sigma_{\texttt{GA}}^*$ by minimizing the matching cost that combines the original classification cost $\mathcal{L}_{\texttt{class}}$ and $\mathcal{L}_{\texttt{box}}^{\texttt{GA}}$.

5 Experiments

We demonstrate the effectiveness of REXO through evaluations on two open radar datasets: HIBER (Wu et al. 2023) and MMVR (Rahman et al. 2024).

5.1 Setup

High-Resolution Indoor Radar Datasets: MMVR includes multi-view radar heatmaps collected from over 25 human subjects across 6 rooms over a span of 9 days. It consists of 345K data frames collected in 2 protocols: 1) Protocol 1 (P1: Open Foreground) with 107.9K frames in an open-foreground room with a single subject; and 2) Protocol 2 (P2: Cluttered Space) with 237.9K frames in 5 cluttered rooms with single and multiple subjects. Under each protocol, two data splits are defined to evaluate radar perception performance: 1) S1: a random data split and 2) S2: a cross-session, unseen split.

Method	WALK		
	AP	AP_{50}	AP_{75}
RFMask	17.77	52.46	6.78
RFMask3D	16.58	48.10	6.53
DETR	14.45	47.33	4.25
RETR	22.09	59.83	10.99
REXO	25.33	62.55	15.83

Table 2: Evaluation on the WALK data split of the HIBER.

HIBER, partially released, includes multi-view radar heatmaps from 10 human subjects in a single room but from different angles with multiple data splits. In our evaluation, we used the "WALK" data split, consisting of 73.5K data frames with one subject walking in the room. For both datasets, annotations such as 3D BBoxes in the radar coordinate system and 2D image plane BBoxes are provided to train the baseline methods and REXO.

Implementation: We consider RFMask (Wu et al. 2023), DETR (Carion et al. 2020) and RETR (Yataka et al. 2024) as baseline methods. Additionally, we evaluate a 3D extension of RFMask, referred to as RFMask3D, that takes the two radar views as inputs for BBox prediction. Since RFMask and DETR originally compute the BBox loss only in the 2D horizontal radar plane and the 2D image plane, respectively, we follow the implementation of RETR and enhance both methods with a unified bi-plane BBox loss. Furthermore, we introduce a DETR variant with a top-K feature selection, allowing it to take features from both horizontal and vertical heatmaps as input. For RETR, we set the number of object queries to 10. To ensure a fair comparison, we also set $N_{\texttt{train}} = 10$ for REXO during training. All methods are evaluated using $M = 4$ consecutive radar frames.

Metrics: We evaluate performance using average precision (AP) at two IoU thresholds of 0.5 (AP_{50}) and 0.75 (AP_{75}), along with the mean AP (AP) computed over thresholds in the range of $[0.5 : 0.05 : 0.95]$.

5.2 Main Results

MMVR: Table 1 presents the results under the four combinations of two protocols and two data splits of the MMVR dataset. REXO demonstrates significant performance improvements in P1S2, P2S1, and P2S2. Notably, in P2S2 where the test radar frames contain an entirely unseen environment during training, REXO outperforms the best base-

λ_{3D}	λ_{2D}	AP
0.00	1.00	0.98
0.50	1.00	4.23
1.00	0.10	15.55
1.00	0.50	19.38
1.00	1.00	23.47

(a) Strong **2D (λ_{2D})/3D (λ_{3D}) supervision** improves performance.

Method	N_{train}	AP
RETR	10	12.45
	20	9.85
	50	8.49
REXO	10	23.47
	20	20.94
	50	19.67

(b) **# of BBoxes for training**. REXO remains stable.

N	REXO	RETR
2	8.87	8.36
10	23.48	12.45
20	23.00	6.57
40	22.32	3.63
60	21.94	2.65
80	21.70	2.16

(c) **# of BBoxes for inference**. REXO sustains its AP as N increases.

Steps	AP
1	23.48
3	24.01
5	24.12
7	24.17
9	24.25
10	24.27

(d) **# of denoising steps**. More steps slightly improve detection.

Method	AP
DiffusionDet	20.75
REXO (Horizontal)	22.75
REXO (Vertical)	7.18
REXO (Both Views)	23.47

(e) **DiffusionDet vs. single-/multi-view REXO.** Multi-view achieves the best AP.

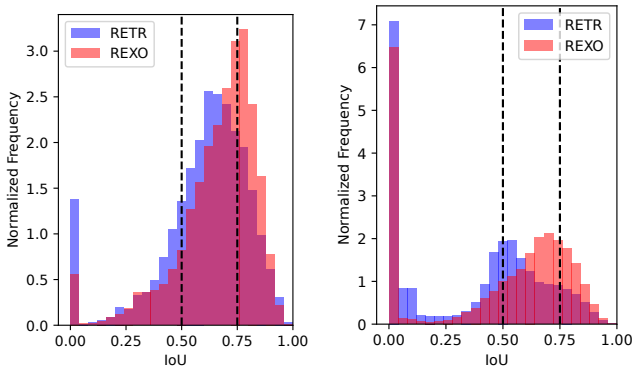
Table 5: Ablation study under P2S2 on MMVR.

Grounding	AP (P2S2)	AP (WALK)
×	22.67	21.11
✓	23.47	25.33

Table 3: The ground-level constraint can improve the detection performance on both datasets.

D [cm]	AP
$D \leq 20$	9.93
$D > 20$	23.47

Table 4: AP drops when depth difference D is less than 20 cm.



(a) Seen Frames under P2S2. (b) Unseen Frames under P2S2.

Figure 5: AP breakdowns with IoU histograms on MMVR.

line RETR by a large margin, boosting AP from 12.45 to 23.47, highlighting its strong generalization capabilities. Surprisingly, under the simplest combination P1S1 where a single subject is recorded in the same room with a random data split, REXO's performance is slightly lower than that of RETR, particularly on the metric AP_{50}.

HIBER: Table 2 presents the results evaluated on the "WALK" data split of the HIBER dataset. For baselines, RFMask, RFMask3D, and DETR show comparable performance, while RETR exhibits the strongest baseline performance. REXO outperforms RETR across all evaluation metrics, demonstrating strong performance in both low- and high-IoU BBox performance evaluations. This ability to consistently outperform the baselines across different IoU thresholds indicates REXO's robustness in capturing object localization with relatively better accuracy.

5.3 Ablation Study

We present ablation studies for REXO under the most challenging "P2S2" of the MMVR dataset.

Effectiveness of Ground-Level Constraint: Table 3 reports the effect of ground-level constraint. In MMVR, the subject stands on the ground or sits in a chair, so the constraint is effective. The table shows that we also evaluated the HIBER dataset as a supplement and observed a significant improvement in performance. It should be noted that constraint is not always accurate when the subject jumps or stands on an obstacle, but it is still effective in terms of stabilizing inference.

2D vs. 3D Supervision Strength: Table 5a compares the various weight parameters λ_{3D} and λ_{2D} in (15). The results highlight the necessity of accounting for the loss of both the 3D BBox and the 2D BBox, and the importance of the prediction accuracy of the 3D BBox in the radar space for the prediction accuracy of the 2D BBox on the image plane. The image plane supervision is essential to train the learnable refinement module. Strong 2D (image plane) and 3D (radar space) supervision yields better performance.

Number of BBoxes in Training: We evaluate the impact of N_{train}, the number of BBoxes for REXO and the number of queries for RETR, on the three AP metrics in Table 5b. It is seen that AP tends to decrease as N_{train} increases for both methods, but this may be due to the number of BBoxes being too large relative to the number of subjects, since the maximum number of subjects in the MMVR is three per frame.

Dynamic Number of BBoxes in Inference: Table 5c evaluates the impact of varying the number of BBoxes during inference. While RETR exhibits a sharp performance decline when the number of queries exceeds 10, REXO experiences a much smaller decrease. This robustness in handling varying numbers of BBoxes during inference is a direct advantage inherited from DiffusionDet.

Number of Iteration Steps: Table 5d presents REXO's performance as the number of iteration steps increases. Increasing the steps from 1 to 10 yields improvements of $+0.78$ in AP, showing consistent gains with more iterations. We also report runtime and FPS on a single NVIDIA RTX 6000: 60 ms (17 FPS) for 1 step, 255 ms (4 FPS) for 5 steps,

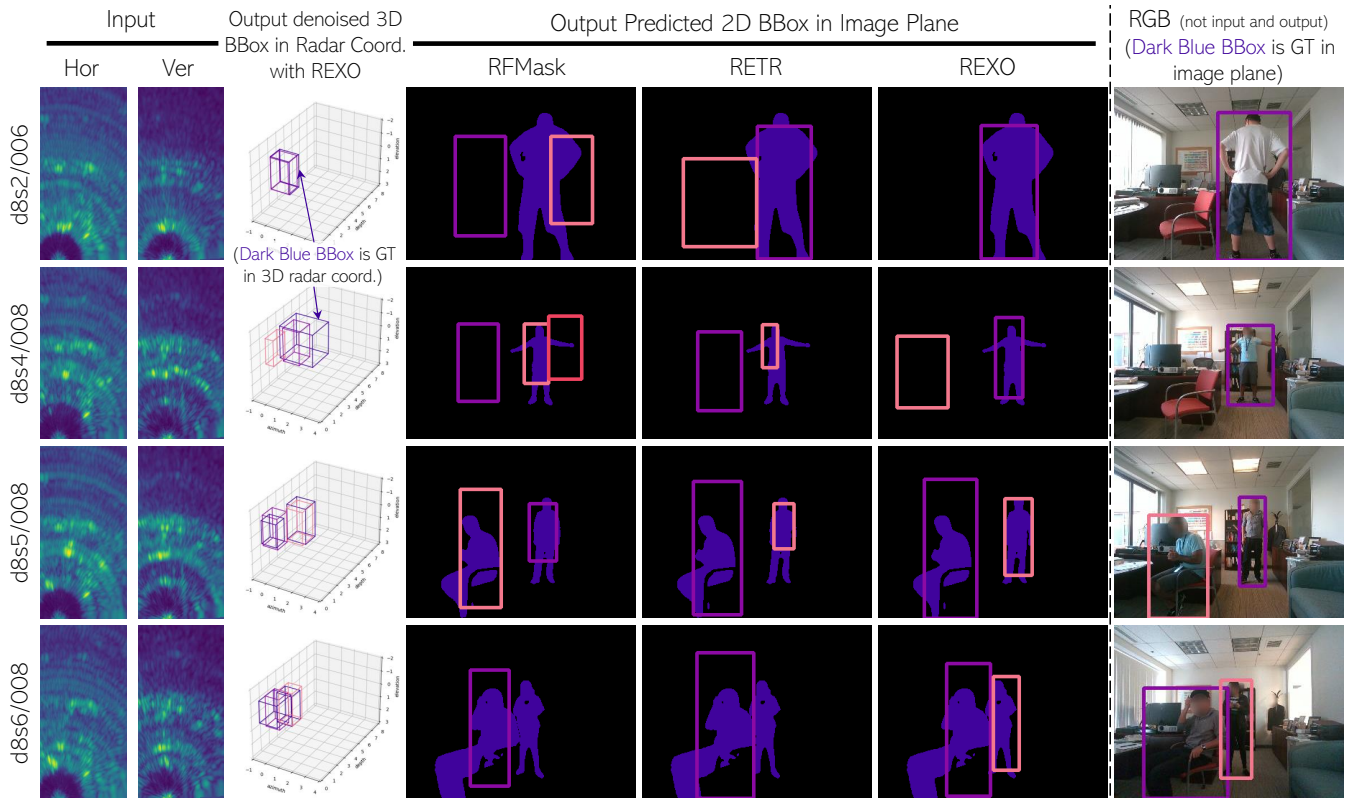


Figure 6: Visualization of unseen frames in P2S2 of MMVR: The left column shows the radar heatmaps, followed by the second column displaying predicted/GT 3D BBoxes in the radar space. Corresponding image-plane 2D BBox predictions are shown in the middle column for two baselines (RFMask and RETR) and REXO, with purple segmentation masks overlaid to illustrate the alignment with human GT. The right column presents the RGB images with GT 2D BBoxes for qualitative check.

and 483 ms (2 FPS) for 10 steps. While more steps improve accuracy, they incur higher latency. Thus, using 5 steps offers a practical trade-off between detection performance and runtime efficiency for indoor human sensing.

Additional comparison with DiffusionDet and REXO: Table 5e confirms that our REXO (both views) outperforms the original DiffusionDet with radar heatmaps, which denotes our REXO is more appropriate for our radar settings. It also suggests the horizontal view (i.e., bird's-eye view) is more critical than the vertical view for detection since the vertical view cannot separate the azimuth position.

5.4 Challenging Cases

We provide additional analysis for two scenarios: 1) when two subjects are at a similar depth; and 2) generalization over unseen environments. For 1), when multiple subjects are at the same depth, the reflections from different subjects overlap and form more complex patterns than that for a single subject, potentially leading to failed cross-view feature association and radar-conditioned denoising steps. As confirmed in Table 4, AP drops significantly when the depth difference D is less than 20 cm based on the evaluation over P2S2. For 2), we divide the test radar frames in P2S2 into "Seen" and "Unseen" frames, and analyze their APs using IoU histograms in Fig. 5, where blue and red histograms represent the IoU distributions for RETR and REXO, respectively, and the left and right dotted lines mark the two IoU thresholds at 0.5 and 0.75. For the "Seen" frames, REXO exhibits a better AP as it provides high-quality predictions with IoU around 0.75. For the "Unseen" frames, REXO clearly dominates the IoU range of $[0.75, 1]$, while RETR shows a heavier concentration around an IoU of 0.5.

Fig. 6 further visualizes selected "Unseen" frames from a room never encountered during training in P2S2. It is seen that 2D BBox predictions by REXO align more closely with human segmentation masks (purple pixels) than those of RETR and RFMask. This improvement is potentially due to the explicit cross-view feature association, which strengthens consistency across radar views even in new environments, yielding better generalization.

6 Conclusion

For indoor radar perception, we proposed REXO, a novel multi-view radar object detection method that refines 3D BBoxes through a diffusion process. By explicitly guiding cross-view radar feature association and incorporating ground-level constraint, REXO achieves consistent performance improvements on two open indoor radar datasets over a list of strong baselines.

References

Adib, F.; Hsu, C.-Y.; Mao, H.; Katabi, D.; and Durand, F. 2015. RF-Capture: Capturing a Coarse Human Figure Through a Wall. In *SIGGRAPH Asia*.

Amit, T.; Shaharbany, T.; Nachmani, E.; and Wolf, L. 2022. SegDiff: Image Segmentation with Diffusion Probabilistic Models. arXiv:2112.00390.

Barnes, D.; Gadd, M.; Murcutt, P.; Newman, P.; and Posner, I. 2020. The Oxford radar robotcar dataset: A radar extension to the Oxford robotcar dataset. In *International Conference on Robotics and Automation*, 6433–6438.

Blattmann, A.; Rombach, R.; Ling, H.; Dockhorn, T.; Kim, S. W.; Fidler, S.; and Kreis, K. 2023. Align Your Latents: High-Resolution Video Synthesis With Latent Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 22563–22575.

Caesar, H.; Bankiti, V.; Lang, A. H.; Vora, S.; Liong, V. E.; Xu, Q.; Krishnan, A.; Pan, Y.; Baldan, G.; and Beijbom, O. 2020. nuScenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11621–11631.

Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; and Zagoruyko, S. 2020. End-to-End Object Detection with Transformers. In *European Conference on Computer Vision (ECCV)*, 213–229.

Chen, H.; Gu, J.; Chen, A.; Tian, W.; Tu, Z.; Liu, L.; and Su, H. 2023a. Single-Stage Diffusion NeRF: A Unified Approach to 3D Generation and Reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2416–2425.

Chen, S.; Sun, P.; Song, Y.; and Luo, P. 2023b. DiffusionDet: Diffusion Model for Object Detection. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 19773–19786.

Chi, G.; Yang, Z.; Wu, C.; Xu, J.; Gao, Y.; Liu, Y.; and Han, T. X. 2024. RF-Diffusion: Radio Signal Generation via Time-Frequency Diffusion. arXiv:2404.09140.

Fan, J.; Yang, J.; Xu, Y.; and Xie, L. 2024. Diffusion Model Is a Good Pose Estimator from 3D RF-Vision. In *European Conference on Computer Vision (ECCV)*, 1–18. Cham.

Gao, X.; Xing, G.; Roy, S.; and Liu, H. 2021. RAMP-CNN: A Novel Neural Network for Enhanced Automotive Radar Object Recognition. *IEEE Sensors Journal*, 21(4): 5119–5132.

Gu, Z.; Chen, H.; and Xu, Z. 2024. DiffusionInst: Diffusion Model for Instance Segmentation. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2730–2734.

He, K.; Gkioxari, G.; Dollár, P.; and Girshick, R. 2017. Mask R-CNN. In *IEEE International Conference on Computer Vision (ICCV)*, 2980–2988.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep Residual Learning for Image Recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 770–778.

Ho, C.-J.; Tai, C.-H.; Lin, Y.-Y.; Yang, M.-H.; and Tsai, Y.-H. 2023. Diffusion-SS3D: Diffusion Model for Semi-supervised 3D Object Detection. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, 49100–49112.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising Diffusion Probabilistic Models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, 6840–6851.

Ho, J.; Salimans, T.; Gritsenko, A.; Chan, W.; Norouzi, M.; and Fleet, D. J. 2022. Video Diffusion Models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, 8633–8646.

Ho, Y.-H.; Cheng, J.-H.; Kuan, S. Y.; Jiang, Z.; Chai, W.; Huang, H.-W.; Lin, C.-L.; and Hwang, J.-N. 2024. RT-Pose: A 4D Radar Tensor-based 3D Human Pose Estimation and Localization Benchmark. In *European Conference on Computer Vision (ECCV)*, 107–125. ISBN 978-3-031-73036-8.

Kay, S. M. 1998. *Fundamentals of Statistical Signal Processing: Detection Theory*. Prentice Hall.

Kuhn, H. W. 1955. The Hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2): 83–97.

Lee, S.-P.; Kini, N. P.; Peng, W.-H.; Ma, C.-W.; and Hwang, J.-N. 2023. HuPR: A Benchmark for Human Pose Estimation Using Millimeter Wave Radar. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 5715–5724.

Li, W.; Liu, X.; Ma, J.; and Yuan, Y. 2024. CLIFF: Continual Latent Diffusion for Open-Vocabulary Object Detection. In *European Conference on Computer Vision (ECCV)*.

Lin, T.-Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; and Belongie, S. 2017. Feature Pyramid Networks for Object Detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 936–944.

Lu, C. X.; Rosa, S.; Zhao, P.; Wang, B.; Chen, C.; Stankovic, J. A.; Trigoni, N.; and Markham, A. 2020. See through smoke: robust indoor mapping with low-cost mmWave radar. In *The 18th International Conference on Mobile Systems, Applications, and Services (MobiSys)*, 14–27.

Luan, K.; Shi, C.; Wang, N.; Cheng, Y.; Lu, H.; and Chen, X. 2024. Diffusion-Based Point Cloud Super-Resolution for mmWave Radar Data. In *IEEE International Conference on Robotics and Automation (ICRA)*, 11171–11177.

Paek, D.-H.; et al. 2022. K-Radar: 4D Radar Object Detection for Autonomous Driving in Various Weather Conditions. In *NeurIPS*, volume 35, 3819–3829.

Pandharipande, A.; Cheng, C.-H.; Dauwels, J.; Gurbuz, S. Z.; Ibanez-Guzman, J.; Li, G.; Piazzoni, A.; Wang, P.; and Santra, A. 2023. Sensing and Machine Learning for Automotive Perception: A Review. *IEEE Sensors Journal*, 23(11): 11097–11115.

Rahman, M. M.; Yataka, R.; Kato, S.; Wang, P.; Li, P.; Cardace, A.; and Boufounos, P. 2024. MMVR: Millimeter-wave Multi-View Radar Dataset and Benchmark for Indoor Perception. In *European Conference on Computer Vision (ECCV)*, 306–322. ISBN 978-3-031-72986-7.

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2017. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(6): 1137–1149.

Rezatofighi, H.; Tsoi, N.; Gwak, J.; Sadeghian, A.; Reid, I.; and Savarese, S. 2019. Generalized Intersection Over Union: A Metric and a Loss for Bounding Box Regression. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 658–666.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-Resolution Image Synthesis With Latent Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 10684–10695.

Sengupta, A.; Jin, F.; Zhang, R.; and Cao, S. 2020. mm-Pose: Real-Time Human Skeletal Posture Estimation Using mmWave Radars and CNNs. *IEEE Sensors Journal*, 20(17): 10032–10044.

Sheeny, M.; De Pellegrin, E.; Mukherjee, S.; Ahrabian, A.; Wang, S.; and Wallace, A. 2021. RADIATE: A Radar Dataset for Automotive Perception in Bad Weather. In *IEEE International Conference on Robotics and Automation (ICRA)*, 1–7.

Shi, G.; Li, R.; and Ma, C. 2022. PillarNet: Real-Time and High-Performance Pillar-Based 3D Object Detection. In *European Conference on Computer Vision (ECCV)*, 35–52.

Skog, M.; Kotlyar, O.; Kubelka, V.; and Magnusson, M. 2024. Human Detection from 4D Radar Data in Low-Visibility Field Conditions. *arXiv:2404.05307*.

Song, J.; Meng, C.; and Ermon, S. 2021. Denoising Diffusion Implicit Models. In *International Conference on Learning Representations (ICLR)*.

Song, Y.; Dhariwal, P.; Chen, M.; and Sutskever, I. 2023. Consistency Models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, 32211–32252.

Song, Y.; and Ermon, S. 2019. Generative Modeling by Estimating Gradients of the Data Distribution. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32.

Sun, S.; Petropulu, A. P.; and Poor, H. V. 2020. MIMO Radar for Advanced Driver-Assistance Systems and Autonomous Driving: Advantages and Challenges. *IEEE Signal Processing Magazine*, 37(4): 98–117.

Tan, D.; Chen, H.; Tian, W.; and Xiong, L. 2024. DiffusionRegPose: Enhancing Multi-Person Pose Estimation Using a Diffusion-Based End-to-End Regression Approach. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2230–2239.

Wu, J.; Geng, R.; Li, Y.; Zhang, D.; Lu, Z.; Hu, Y.; and Chen, Y. 2024. Diffradar:High-Quality mmWave Radar Perception With Diffusion Probabilistic Model. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 8291–8295.

Wu, Z.; Zhang, D.; Xie, C.; Yu, C.; Chen, J.; Hu, Y.; and Chen, Y. 2023. RFMask: A Simple Baseline for Human Silhouette Segmentation With Radio Signals. *IEEE Transactions on Multimedia*, 25: 4730–4741.

Xiang, X.; Dräger, S.; and Zhang, J. 2023. 3DifFusionDet: Diffusion Model for 3D Object Detection with Robust LiDAR-Camera Fusion. *ArXiv*, abs/2311.03742.

XU, C.; Ling, H.; Fidler, S.; and Litany, O. 2024. 3DiffTection: 3D Object Detection with Geometry-Aware Diffusion Features. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 10617–10627.

Yang, B.; Khatri, I.; Happold, M.; and Chen, C. 2023a. ADCNet: Learning from Raw Radar Data via Distillation. *arXiv:2303.11420*.

Yang, J.; Huang, H.; Zhou, Y.; Chen, X.; Xu, Y.; Yuan, S.; Zou, H.; Lu, C. X.; and Xie, L. 2023b. MM-Fi: Multi-Modal Non-Intrusive 4D Human Dataset for Versatile Wireless Sensing. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, 18756–18768.

Yao, S.; Guan, R.; Peng, Z.; Xu, C.; Shi, Y.; Ding, W.; Lim, E. G.; Yue, Y.; Seo, H.; Man, K. L.; Ma, J.; Zhu, X.; and Yue, Y. 2024. Exploring Radar Data Representations in Autonomous Driving: A Comprehensive Review. *arXiv:2312.04861*.

Yataka, R.; Cardace, A.; Wang, P.; Boufounos, P.; and Takahashi, R. 2024. RETR: Multi-View Radar Detection Transformer for Indoor Perception. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, 19839–19869.

Yu, J. J.; Forghani, F.; Derpanis, K. G.; and Brubaker, M. A. 2023. Long-Term Photometric Consistent Novel View Synthesis with Diffusion Models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 7094–7104.

Zhang, R.; Xue, D.; Wang, Y.; Geng, R.; and Gao, F. 2024. Towards Dense and Accurate Radar Perception via Efficient Cross-Modal Diffusion Model. *IEEE Robotics and Automation Letters*, 9(9): 7429–7436.

Zhao, M.; Li, T.; Alsheikh, M. A.; Tian, Y.; Zhao, H.; Torralba, A.; and Katabi, D. 2018a. Through-Wall Human Pose Estimation Using Radio Signals. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 7356–7365.

Zhao, M.; Tian, Y.; Zhao, H.; Alsheikh, M. A.; Li, T.; Hristov, R.; Kabelac, Z.; Katabi, D.; and Torralba, A. 2018b. RF-based 3D skeletons. In *The Conference of the ACM Special Interest Group on Data Communication (SIGCOMM)*, 267–281.

Zhao, M.; Yue, S.; Katabi, D.; Jaakkola, T. S.; and Bianchi, M. T. 2017. Learning Sleep Stages from Radio Signals: A Conditional Adversarial Architecture. In *International Conference on Machine Learning (ICML)*, volume 70, 4100–4109.

Zhao, P.; Lu, C. X.; Wang, B.; Trigoni, N.; and Markham, A. 2023. CubeLearn: End-to-End Learning for Human Motion Recognition From Raw mmWave Radar Signals. *IEEE Internet of Things Journal*, 10(12): 10236–10249.