MITSUBISHI ELECTRIC RESEARCH LABORATORIES
https://www.merl.com

Radar-Conditioned 3D Bounding Box Diffusion for Indoor Human Perception

Yataka, Ryoma; Wang, Pu; Boufounos, Petros T.; Takahashi, Ryuhei

TR2025-154 October 23, 2025

Abstract

Privacy-preserving and cost-effective indoor sensing is vital for embodied agents to collaborate safely with people in dynamic scenes. Multi-view millimeter-wave radar shows great potential for this purpose. However, prevailing methods rely on implicit cross-view association, which this reliance often results in ambiguous feature matches and de- graded performance in cluttered environments. To address these limitations, we propose REXO (multi-view Radar object dEtection with 3D bounding boX diffusiOn), which lifts DiffusionDet's 2D box denoising to the full 3D radar space. Noisy 3D boxes are projected onto all radar views to enable explicit association and radar-conditioned denoising. Evaluated on two open indoor radar datasets, our approach out- performs state-of-the-art methods by +11.02 AP on MMVR and +4.22 AP on HIBER.

IEEE International Conference on Computer Vision (ICCV) Workshop 2025

Mitsubishi Electric Research Laboratories, Inc.
201 Broadway, Cambridge, Massachusetts 02139

Radar-Conditioned 3D Bounding Box Diffusion for Indoor Human Perception

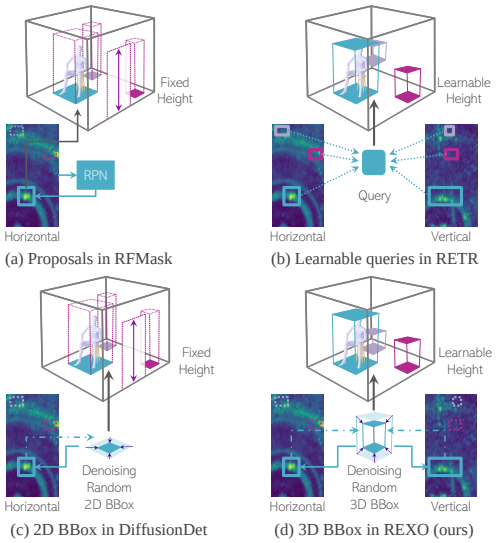
Ryoma Yataka[1,2*], Pu (Perry) Wang[2], Petros Boufounos[2], and Ryuhei Takahashi[1]
[1]*Information Technology R&D Center, Mitsubishi Electric Corporation*, Kanagawa 247-8501, Japan
[2]*Mitsubishi Electric Research Laboratories (MERL)*, Cambridge, MA 02139, USA

Abstract

Privacy-preserving and cost-effective indoor sensing is vital for embodied agents to collaborate safely with people in dynamic scenes. Multi-view millimeter-wave radar shows great potential for this purpose. However, prevailing methods rely on implicit cross-view association, which this reliance often results in ambiguous feature matches and degraded performance in cluttered environments. To address these limitations, we propose ***REXO*** *(multi-view Radar object dEtection with 3D bounding boX diffusiOn), which lifts DiffusionDet's 2D box denoising to the full 3D radar space. Noisy 3D boxes are projected onto all radar views to enable* ***explicit*** *association and radar-conditioned denoising. Evaluated on two open indoor radar datasets, our approach outperforms state-of-the-art methods by* $+11.02$ *AP on MMVR and* $+4.22$ *AP on HIBER.*

1. Introduction

Reliable perception is crucial for embodied agents in indoor settings (homes, factories, clinics), where scene understanding, motion capture, and human-robot collaboration are required. Radar is increasingly used for navigation, manipulation, and safer human-robot interaction because it provides robust awareness in low light, smoke, dust, and even through cardboard and plastic [23, 31]. For example, FuseGrasp [9] fuses radar and camera to grasp transparent objects, exploiting millimeter-wave (mmWave) radar's ability to render transparent materials opaque and robotic-arm radar imaging to recover shapes invisible to RGB-D. On the other hand, radar-only perception (see Appendix A) remains challenging. Multi-view radar methods either pair horizontal proposals with fixed-height vertical ones [37] (Fig. 1 (a)) or use query-based transformers to regress 3D bounding boxes (BBoxes) from both views [40] (Fig. 1 (b)). Image-based object detection has been redefined as a generative denoising process, where a random noisy 2D BBox is iteratively refined through a diffusion denoising process



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

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

Figure 1. (a) RFMask [37] generates horizontal-view proposals with fixed height; (b) RETR [40] implicitly links queries to cross-view features; (c) DiffusionDet [7] needs pairing with fixed-height vertical BBoxes; (d) REXO (**ours**) performs diffusion directly in 3D radar space for simple, explicit cross-view association.

to yield a final clean BBox [7] and this approach generally surpasses query-based detectors. When ported to horizontal radar heatmaps (Fig. 1 (c)), it denoises 2D BBoxes 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, a noisy 3D BBox is 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.

*The work was done at MERL as a visiting scientist.

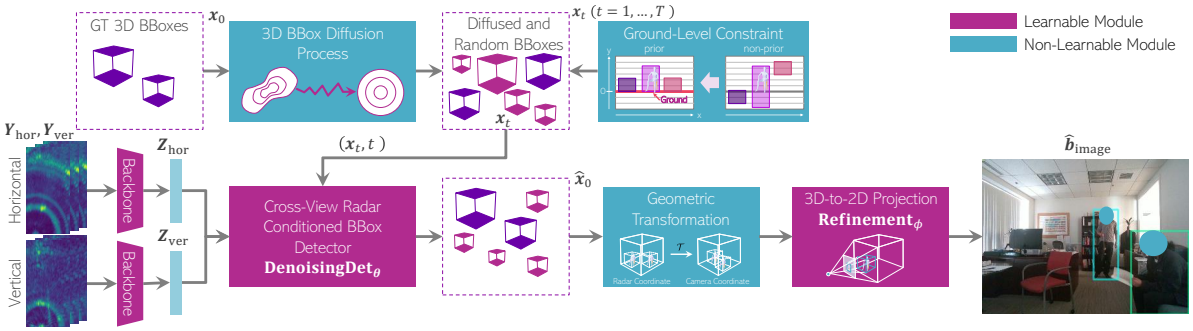


Figure 2. **REXO training:** 1) A backbone extracts horizontal/vertical radar features; 2) Ground-truth 3D BBox $\boldsymbol{x}_0$ are diffused to noisy $\boldsymbol{x}_t$; 3) $\boldsymbol{x}_t$ is grounded using a ground-level constraint; 4) $\texttt{DenoisingDet}_\theta$ projects $\boldsymbol{x}_t$ 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 BBox $\hat{\boldsymbol{b}}_{\texttt{image}}$.

2. **3D BBox Denoising**: 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 multi-view radar perception field.

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

2. REXO: BBox Diffusion in 3D Radar Space

DiffusionDet [7] reformulates object detection as a denoising diffusion process [17, 32], treating $\boldsymbol{x}_t$ as 2D BBox parameters instead of image pixels. We extend this to multi-view radar by lifting $\boldsymbol{x}_t$ to 3D BBox in radar coordinates system. Conditioned on radar heatmaps (see in Fig. 2), REXO performs 3D BBox diffusion in two phases: 1) a **forward process** that adds noise to ground-truth (GT) BBox $\boldsymbol{x}_0$ 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 BBox is also projected to the 2D image plane for supervision in both radar and image domains. We describe REXO in two parts: training and inference.

2.1. Training

Backbone: As illustrated in Fig. 2, we first generate two radar heatmaps (horizontal $\boldsymbol{Y}_{\texttt{hor}} \in \mathcal{R}^{M\times W\times D}$ and vertical $\boldsymbol{Y}_{\texttt{ver}} \in \mathcal{R}^{M\times H\times D}$ where M, W, H and D denote the number of consecutive frames, width, height and depth, respectively. More details are described in Appendix B) derived from raw data captured by horizontal and vertical radar arrays. Taking the two radar heatmaps $\boldsymbol{Y}_{\texttt{hor}} \in \mathcal{R}^{M\times W\times D}$ and $\boldsymbol{Y}_{\texttt{ver}} \in \mathcal{R}^{M\times H\times D}$ as inputs, a shared backbone network (e.g., ResNet [14]) generates horizontal-view and vertical-view radar feature maps: $\boldsymbol{Z}_{\texttt{hor}} = \texttt{backbone}\,(\boldsymbol{Y}_{\texttt{hor}})$ and $\boldsymbol{Z}_{\texttt{ver}} = \texttt{backbone}\,(\boldsymbol{Y}_{\texttt{ver}})$.

Initialization of $\boldsymbol{x}_0$ and Forward Process to $\boldsymbol{x}_t$ with Ground-Level Constraint: 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 BBox $\boldsymbol{x}_{\texttt{rand}} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_6)$ if $N_{\texttt{train}} > N_{\texttt{GT}}$. The diffused 3D BBox $\boldsymbol{x}_t$ at time t can be generated as

$$\boldsymbol{x}_t = \sqrt{\bar{\alpha}_t}\boldsymbol{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}, \tag{1}$$

where $\boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_6)$, and $\bar{\alpha}_t$ denotes the noise variance schedule. Since the BBox is 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 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: $\boldsymbol{x}_t = \{c_x^t, h^t/2, c_z^t, w^t, h^t, d^t\}^\top$ (see the Ground-Level Constraint in Fig. 2).

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

$$\boldsymbol{x}_{t,\texttt{hor}} = \{c_x^t, c_z^t, w^t, d^t\}^\top, \boldsymbol{x}_{t,\texttt{ver}} = \{c_y^t, c_z^t, h^t, d^t\}^\top, \tag{2}$$

and then crops out the cross-view 2D radar features: $\boldsymbol{Z}^{\texttt{crop}}_{\texttt{hor/ver}} = \texttt{RoIAlign}(\boldsymbol{Z}_{\texttt{hor/ver}}, \boldsymbol{x}_{t,\texttt{hor/ver}}) \in \mathbb{R}^{C\times r\times r}$ via a standard ROIAlign operation [15], 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

$$\boldsymbol{Z}^{\texttt{crop}}_{\texttt{radar}} = \{\boldsymbol{Z}^{\texttt{crop}}_{\texttt{hor}}, \boldsymbol{Z}^{\texttt{crop}}_{\texttt{ver}}\} \in \mathbb{R}^{C\times r\times 2r}, \tag{3}$$

each corresponding to a noisy 3D BBox $\boldsymbol{x}_t$. Conditioned on $\boldsymbol{Z}^{\texttt{crop}}_{\texttt{radar}}$, a $\texttt{DenoisingDet}_\theta$ with learnable weights θ is

Table 1. Evaluation on 4 data splits of the MMVR dataset and WALK of the HIBER dataset.

Method	MMVR:P1S1			MMVR:P1S2			MMVR:P2S1			MMVR:P2S2			HIBER:WALK		
	AP	AP_{50}	AP_{75}	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	17.77	52.46	6.78
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	16.58	48.10	6.53
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	14.45	47.33	4.25
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	22.09	59.83	10.99
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	25.33	62.55	15.83

trained to estimate the BBox $\hat{\boldsymbol{x}}_0$ and the class scores $\hat{\boldsymbol{p}}$ as

$$\{\hat{\boldsymbol{x}}_0, \hat{\boldsymbol{p}}\} = \texttt{DenoisingDet}_\theta\left(\boldsymbol{x}_t, t, \boldsymbol{Z}^{\texttt{crop}}_{\texttt{radar}}\right), \quad (4)$$

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.

3D-to-2D Projection with Learnable Refinement: REXO further projects $\hat{\boldsymbol{x}}_0$ in (4) into the 2D image plane. 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 $\boldsymbol{x}^i_{\texttt{camera}} = \boldsymbol{R}\hat{\boldsymbol{x}}^i_{\texttt{radar}} + \boldsymbol{v}$, where $\hat{\boldsymbol{x}}^i_{\texttt{radar}}$ is the i-th corner of $\hat{\boldsymbol{x}}_{\texttt{radar}}$, $\boldsymbol{R}$ and $\boldsymbol{v}$ are the calibrated 3D rotation matrix and translation vector: Each 3D corner $\boldsymbol{x}^i_{\texttt{camera}}$ is projected to the image plane through the calibrated pinhole model:

$$\boldsymbol{b}_{\texttt{init}} = \left\{\bar{c}_x, \bar{c}_y, \bar{w}, \bar{h}\right\}^\top = \texttt{proj}_{\texttt{init}}\left(\boldsymbol{x}_{\texttt{camera}}\right). \quad (5)$$

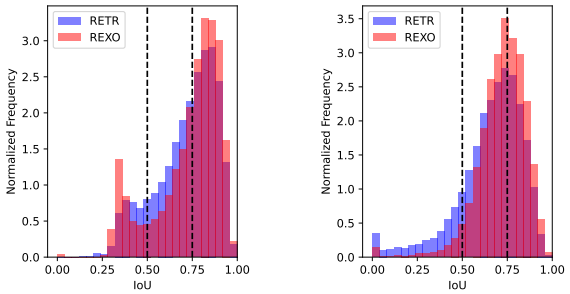
Since $\boldsymbol{b}_{\texttt{init}}$ systematically overshoots the ground-truth extent (see Appendix C), we attach a refinement module with learnable parameter $\boldsymbol{\phi}$ to obtain the offset:

$$\Delta\boldsymbol{b} = \left\{\Delta\bar{x}, \Delta\bar{y}, \Delta\bar{w}, \Delta\bar{h}\right\}^\top = \texttt{Refinement}_{\boldsymbol{\phi}}\left(\boldsymbol{f}\right), \quad (6)$$

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 [17] and `Predictor` denotes the time-dependent predictor [7] with the radar feature and the embedding. Applying these offsets produces the final image-plane box $\widehat{\boldsymbol{b}}_{\texttt{image}}$, achieving tighter alignment without sacrificing geometric consistency.

$$\widehat{\boldsymbol{b}}_{\texttt{image}} = \{\bar{c}_x + \bar{w}\Delta\bar{x}, \bar{c}_y + \bar{h}\Delta\bar{y}, e^{\Delta\bar{w}}\bar{w}, e^{\Delta\bar{h}}\bar{h}\}^\top. \quad (7)$$

Loss: To ensure consistency between the radar and image plane representations, we adopt a simplified scheme of the Tri-plane loss [40] that directly calculates the loss of 3D BBox. REXO employs the Hungarian match cost [21] with a loss function computed in both the 3D and 2D spaces: $\mathcal{L}^{\texttt{GA}}_{\texttt{box}} = \lambda_{\texttt{3D}}\mathcal{L}^{\texttt{3D}}_{\texttt{box}}\left(\boldsymbol{x}_{\texttt{radar}}, \widehat{\boldsymbol{x}}_{\texttt{radar}}\right) + \lambda_{\texttt{2D}}\mathcal{L}^{\texttt{2D}}_{\texttt{box}}(\boldsymbol{b}_{\texttt{image}}, \widehat{\boldsymbol{b}}_{\texttt{image}})$, 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 $\mathcal{L}_{\texttt{GIoU}}$ [27] and the ℓ_1 loss $\mathcal{L}_{\texttt{L}_1}$, and the coefficients λ balance the relative contribution of each loss term.



(a) P1S1 (b) P1S2

Figure 3. AP breakdowns with IoU histograms on MMVR.

2.2. 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 $t = 1$. With $\boldsymbol{x}_t$ and radar features $\{\boldsymbol{Z}_{\texttt{hor}}, \boldsymbol{Z}_{\texttt{ver}}\}$, the trained $\texttt{DenoisingDet}_\theta$ in (4) predicts $\hat{\boldsymbol{x}}_0$, giving

$$p_\theta\left(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t, \boldsymbol{Z}_{\texttt{hor}}, \boldsymbol{Z}_{\texttt{ver}}\right) = \mathcal{N}(\sqrt{\alpha_{t-1}}\boldsymbol{x}_0 + \gamma\boldsymbol{\epsilon}^{(t)}_\theta, \sigma_t^2\boldsymbol{I}_6),$$
$$\boldsymbol{x}_{t-1} = \sqrt{\alpha_{t-1}}\hat{\boldsymbol{x}}_0 + \sqrt{1 - \alpha_{t-1} - \sigma_t^2}\cdot\boldsymbol{\epsilon}^{(t)}_\theta + \sigma_t\boldsymbol{\epsilon}_t, \quad (8)$$

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. 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 and the 3D-to-2D projection. Boxes whose class scores exceed a threshold are output as detections.

3. Experiments

We demonstrate the effectiveness of REXO through evaluations on two open high-resolution radar datasets.

Datasets: *MMVR* [26] 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: P1: Open Foreground) with 107.9K frames in an open-foreground room with a single subject; and P2: Cluttered Space with 237.9K frames in 5 cluttered rooms with single and multiple subjects.

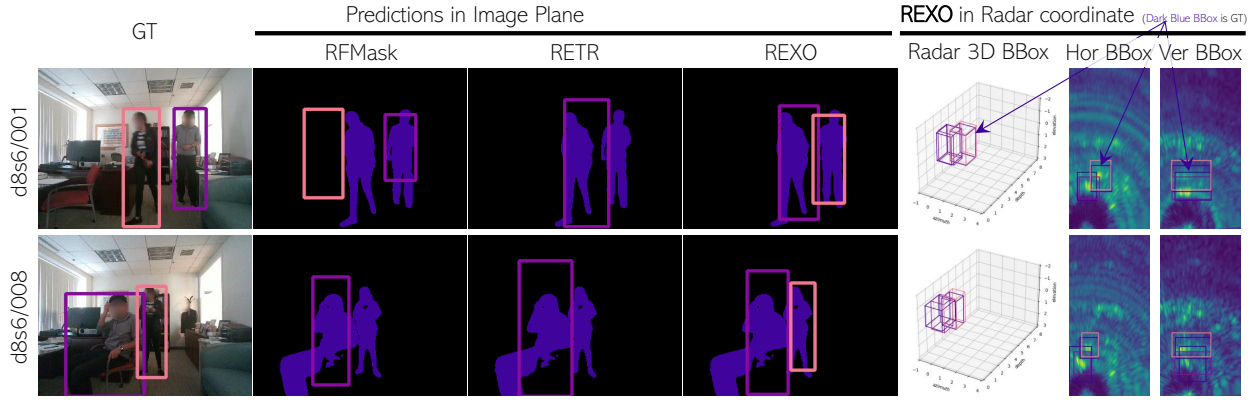


Figure 4. Visualization of unseen frames in P2S2 of MMVR: The left column shows the radar heatmaps, followed by the second column displaying predicted/GT 3D BBoxes in the radar space. Corresponding image-plane 2D BBox predictions are shown in the middle column for two baselines (RFMask and RETR) and 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.

Under each protocol, two data splits are defined to evaluate radar perception performance: S1: a random data split and S2: a cross-session, unseen split. *HIBER* [37], 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.

Implementation: We consider RFMask [37], DETR [5] and RETR [40] as baseline methods. Additionally, we evaluate a 3D extension of RFMask (RFMask3D; see Appendix D), that takes the two radar views as inputs for BBox prediction. Hyperparameter settings are provided in Appendix E.

Metrics: We evaluate performance using average precision (AP) at two IoU thresholds of 0.5 ($\texttt{AP}_{50}$) and 0.75 ($\texttt{AP}_{75}$), along with the mean AP (AP) computed over thresholds in the range of $[0.5 : 0.05 : 0.95]$. For detailed metric definitions, refer to Appendix F.

Result of 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 baseline 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 $\texttt{AP}_{50}$. To understand these differences, we break down the AP into IoU histograms for (a) P1S1 and (b) P1S2, as illustrated in Fig. 3, 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. It is seen that in Fig. 3a, the excess of RETR over REXO (blue areas) over the IoU interval $[0.5, 0.75]$ is greater than that of REXO over RETR (pink areas) over the interval $[0.75, 1.0]$, explaining RETR's higher $\texttt{AP}_{50}$ under P1S1. Meanwhile, REXO has better $\texttt{AP}_{75}$ as it provides more high-quality predictions with IoU above 0.75.

Result of HIBER: Table 1 presents the results evaluated on the "WALK" data split of the HIBER dataset. As well as MMVR cases, REXO outperforms RETR across all evaluation metrics with an AP of 25.33, surpassing RETR's AP at 22.09. REXO attains $\texttt{AP}_{50}$ of 62.55 and $\texttt{AP}_{75}$ of 15.83, demonstrating strong performance in both low- and high-IoU BBox performance evaluations.

Visualization: Fig. 4 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. More challenging examples are provided in Appendix G.

4. Conclusion

We proposed REXO, a multi-view radar object detection method that refines the 3D BBox through a diffusion process. By explicitly guiding cross-view radar feature association, REXO achieves consistent performance improvements on two open indoor radar datasets over a list of strong baselines.

References

[1] Tomer Amit, Tal Shaharbany, Eliya Nachmani, and Lior Wolf. SegDiff: Image segmentation with diffusion probabilistic models, 2022. 7

[2] Dan Barnes, Matthew Gadd, Paul Murcutt, Paul Newman, and Ingmar Posner. The Oxford radar robotcar dataset: A radar extension to the Oxford robotcar dataset. In *International Conference on Robotics and Automation*, pages 6433–6438, 2020. 7

[3] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. 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)*, pages 22563–22575, 2023. 7

[4] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuScenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11621–11631, 2020. 7

[5] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision (ECCV)*, page 213–229, 2020. 4, 8

[6] Hansheng Chen, Jiatao Gu, Anpei Chen, Wei Tian, Zhuowen Tu, Lingjie Liu, and Hao Su. Single-stage diffusion nerf: A unified approach to 3d generation and reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2416–2425, 2023. 7

[7] Shoufa Chen, Peize Sun, Yibing Song, and Ping Luo. DiffusionDet: Diffusion model for object detection. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 19773–19786, 2023. 1, 2, 3, 7

[8] Guoxuan Chi, Zheng Yang, Chenshu Wu, Jingao Xu, Yuchong Gao, Yunhao Liu, and Tony Xiao Han. RF-diffusion: Radio signal generation via time-frequency diffusion, 2024. 7

[9] Hongyu Deng, Tianfan Xue, and He Chen. Fusegrasp: Radar-camera fusion for robotic grasping of transparent objects. *IEEE Transactions on Mobile Computing*, 24(8):7028–7041, 2025. 1

[10] Mark Everingham, Luc Van Gool, Christopher K. I. Williams, John Winn, and Andrew Zisserman. The PASCAL Visual Object Classes (VOC) Challenge. *International Journal of Computer Vision*, 88(2):303–338, 2010. 9

[11] Junqiao Fan, Jianfei Yang, Yuecong Xu, and Lihua Xie. Diffusion model is a good pose estimator from 3D RF-vision. In *Computer Vision – ECCV 2024*, pages 1–18, Cham, 2025. 7

[12] Xiangyu Gao, Guanbin Xing, Sumit Roy, and Hui Liu. RAMP-CNN: A novel neural network for enhanced automotive radar object recognition. *IEEE Sensors Journal*, 21(4): 5119–5132, 2021. 7

[13] Zhangxuan Gu, Haoxing Chen, and Zhuoer Xu. DiffusionInst: Diffusion model for instance segmentation. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2730–2734, 2024. 7

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 2, 7

[15] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask R-CNN. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2980–2988, 2017. 2

[16] Cheng-Ju Ho, Chen-Hsuan Tai, Yen-Yu Lin, Ming-Hsuan Yang, and Yi-Hsuan Tsai. Diffusion-SS3D: Diffusion model for semi-supervised 3D object detection. In *Advances in Neural Information Processing Systems*, pages 49100–49112, 2023. 7

[17] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, pages 6840–6851, 2020. 2, 3

[18] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. In *Advances in Neural Information Processing Systems*, pages 8633–8646, 2022. 7

[19] Jan Hosang, Rodrigo Benenson, Piotr Dollár, and Bernt Schiele. What makes for effective detection proposals? *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(4):814–830, 2016. 9

[20] S. M. Kay. *Fundamentals of Statistical Signal Processing: Detection Theory*. Prentice Hall, 1998. 7

[21] Harold W. Kuhn. The Hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955. 3

[22] Wuyang Li, Xinyu Liu, Jiayi Ma, and Yixuan Yuan. Cliff: Continual latent diffusion for open-vocabulary object detection. In *ECCV*, 2024. 7

[23] Chris Xiaoxuan Lu, Stefano Rosa, Peijun Zhao, Bing Wang, Changhao Chen, John A. Stankovic, Niki Trigoni, and Andrew Markham. See through smoke: robust indoor mapping with low-cost mmWave radar. In *The 18th International Conference on Mobile Systems, Applications, and Services (MobiSys)*, page 14–27, 2020. 1

[24] Kai Luan, Chenghao Shi, Neng Wang, Yuwei Cheng, Huimin Lu, and Xieyuanli Chen. Diffusion-based point cloud super-resolution for mmwave radar data. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 11171–11177, 2024. 7

[25] Dong-Hee Paek et al. K-Radar: 4D radar object detection for autonomous driving in various weather conditions. In *NeurIPS*, pages 3819–3829, 2022. 7

[26] M. Mahbubur Rahman, Ryoma Yataka, Sorachi Kato, Pu Wang, Peizhao Li, Adriano Cardace, and Petros Boufounos. MMVR: Millimeter-wave multi-view radar dataset and benchmark for indoor perception. In *European Conference on Computer Vision (ECCV)*, pages 306–322, 2024. 3, 7

[27] Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. Generalized intersection over union: A metric and a loss for bounding box regression. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 658–666, 2019. 3

[28] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, 2022. 7

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

[30] Guangsheng Shi, Ruifeng Li, and Chao Ma. PillarNet: Real-time and high-performance pillar-based 3D object detection. In *European Conference on Computer Vision (ECCV)*, page 35–52, 2022. 7

[31] Mikael Skog, Oleksandr Kotlyar, Vladimír Kubelka, and Martin Magnusson. Human detection from 4D radar data in low-visibility field conditions. *arXiv:2404.05307*, 2024. 1

[32] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021. 2, 7

[33] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2019.

[34] Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. In *Proceedings of the 40th International Conference on Machine Learning*, pages 32211–32252. PMLR, 2023. 7

[35] Dayi Tan, Hansheng Chen, Wei Tian, and Lu Xiong. DiffusionRegPose: Enhancing multi-person pose estimation using a diffusion-based end-to-end regression approach. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2230–2239, 2024. 7

[36] Jincheng Wu, Ruixu Geng, Yadong Li, Dongheng Zhang, Zhi Lu, Yang Hu, and Yan Chen. Diffradar:high-quality mmWave radar perception with diffusion probabilistic model. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8291–8295, 2024. 7

[37] Zhi Wu, Dongheng Zhang, Chunyang Xie, Cong Yu, Jinbo Chen, Yang Hu, and Yan Chen. RFMask: A simple baseline for human silhouette segmentation with radio signals. *IEEE Transactions on Multimedia*, 25:4730–4741, 2023. 1, 4, 7, 8, 9

[38] Xinhao Xiang, Simon Dräger, and Jiawei Zhang. 3DifFusionDet: Diffusion model for 3D object detection with robust lidar-camera fusion. *ArXiv*, abs/2311.03742, 2023. 7

[39] Bo Yang, Ishan Khatri, Michael Happold, and Chulong Chen. ADCNet: Learning from raw radar data via distillation. *arXiv:2303.11420*, 2023. 7

[40] Ryoma Yataka, Adriano Cardace, Perry Wang, Petros Boufounos, and Ryuhei Takahashi. RETR: Multi-view radar detection transformer for indoor perception. In *Advances in Neural Information Processing Systems*, pages 19839–19869, 2024. 1, 3, 4, 7, 9

[41] Jason J. Yu, Fereshteh Forghani, Konstantinos G. Derpanis, and Marcus A. Brubaker. Long-term photometric consistent novel view synthesis with diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7094–7104, 2023. 7

[42] Ruibin Zhang, Donglai Xue, Yuhan Wang, Ruixu Geng, and Fei Gao. Towards dense and accurate radar perception via efficient cross-modal diffusion model. *IEEE Robotics and Automation Letters*, 9(9):7429–7436, 2024. 7

[43] Mingmin Zhao, Tianhong Li, Mohammad Abu Alsheikh, Yonglong Tian, Hang Zhao, Antonio Torralba, and Dina Katabi. Through-wall human pose estimation using radio signals. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7356–7365, 2018. 7

[44] Peijun Zhao, Chris Xiaoxuan Lu, Bing Wang, Niki Trigoni, and Andrew Markham. Cubelearn: End-to-end learning for human motion recognition from raw mmWave radar signals. *IEEE Internet of Things Journal*, 10(12):10236–10249, 2023. 7

A. Related Work

Radar-Only Perception: Learning-based methods have advanced radar detection over traditional model-based approaches [20], benefiting from open large-scale radar point cloud datasets like nuScenes [4], Oxford RobotCar [2], and RADIATE [29]. Image-based and point/voxel-based backbones [14, 30] extract semantic features from radar detection points, generate region proposals, and localize objects. High-resolution heatmaps (e.g., K-Radar [25], HIBER [37], MMVR [26]) and raw ADC data [39] have also been leveraged by previously mentioned RF-Pose [43], RFMask [37], and RETR [40]. CubeLearn [44] replaces Fourier transforms with learnable modules for an end-to-end radar pipeline, while RAMP-CNN [12] enhances range-angle feature extraction via Doppler cues. More recently, diffusion models have been explored for radar applications [8, 11, 24, 36, 42]. Most efforts, e.g., RadarDiffusion [24, 42] and DiffRadar [36], focus on reconstructing LiDAR-like point clouds from low-resolution radar data, while mmDiff [11] estimates and refines pose keypoints from sparse radar points via diffusion process.

Diffusion-based Object Detection: Diffusion models [28, 32–34] have shown impressive results in tasks such as image and video generation [3, 18] and multi-view synthesis [6, 41]. For perception tasks, DiffusionDet [7] 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 [16] 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 [22] further leverages language models to enhance diffusion-based models for open-vocabulary object detection. Diffusion models are also considered for 3D object detection in the context of LiDAR-Camera fusion [38] and other tasks such as pose estimation [35] and semantic segmentation [1, 13].

B. Multi-View Radar Heatmaps

Multi-view radar heatmaps are generated from raw data captured by two radar arrays: a vertical linear array and a horizontal one, as illustrated in Fig. 5. By sampling multiple reflected pulses across the array elements, a 3D raw data cube is constructed for each array, organized along ADC (intra-pulse) samples, pulse (inter-pulse) samples, and array elements. A 3D fast Fourier transform (FFT) converts the data cube into corresponding 3D radar spectra across the range, Doppler velocity, and spatial angle (azimuth for the horizontal array and elevation for the vertical one). To enhance the signal-to-noise ratio (SNR), the 3D radar spectra are integrated along the Doppler domain, generating two 2D radar heatmaps (range-azimuth and range-elevation) in the polar coordinate system. These heatmaps are then transformed into the radar Cartesian coordinate system, where $\boldsymbol{Y}_{\texttt{hor}}(m) \in \mathcal{R}^{W\times D}$ represents the horizontal-depth radar heatmap and $\boldsymbol{Y}_{\texttt{ver}}(m) \in \mathcal{R}^{H\times D}$ the vertical-depth heatmap for the m-th frame. To incorporate temporal information, M consecutive radar frames are grouped together as $\boldsymbol{Y}_{\texttt{hor}} \in \mathcal{R}^{M\times W\times D}$ and vertical $\boldsymbol{Y}_{\texttt{ver}} \in \mathcal{R}^{M\times H\times D}$.

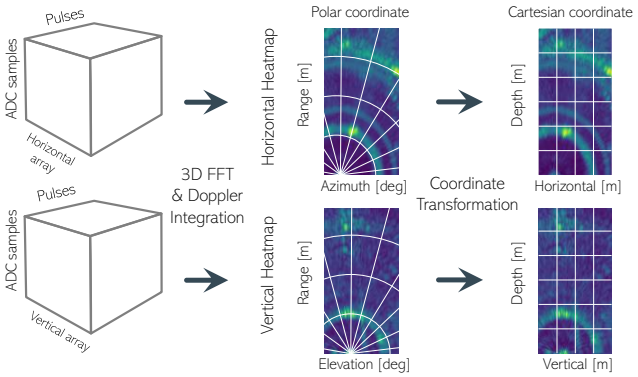


Figure 5. Generation of multi-view heatmaps from raw radar data.

C. Details of 3D-to-2D Projection and Necessity of the Refinement Module

We present the detailed explanations for 3D-to-2D projection and necessity of the refinement module. Given a 3D BBox which consists of its eight vertices

$$\{\boldsymbol{x}^i_{\texttt{camera}} \in \mathbb{R}^3 \mid i = 1, \ldots, 8\}, \tag{9}$$

where each $\boldsymbol{x}^i_{\texttt{camera}}$ is expressed in the 3D camera coordinate system, our goal is to compute the corresponding 2D BBox $\boldsymbol{b}_{\texttt{init}} \in \mathbb{R}^4$, defined by its center coordinates (x_c, y_c) and its width w and height h. To achieve this, we define a projection function with a pinhole camera model as a concrete expression of (5):

$$\texttt{proj}_{\texttt{pinhole}} : \mathbb{R}^3 \to \mathbb{R}^2 : (X, Y, Z) \mapsto (p_x, p_y)\,. \tag{10}$$

In this model, the projection of the point $\boldsymbol{x}^*_{\texttt{camera}} = (X, Y, Z)$ onto the image plane is given by

$$p_x = \frac{f_x X}{Z} + c_x, \quad p_y = \frac{f_y Y}{Z} + c_y, \tag{11}$$

where f_x and f_y are the focal lengths along the x and y axes (in pixels), and (c_x, c_y) represents the coordinates of the principal point in the image. In homogeneous coordinates, this mapping can be expressed as

$$\lambda \begin{pmatrix} p_x \\ p_y \\ 1 \end{pmatrix} = \begin{pmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} X \\ Y \\ Z \end{pmatrix}, \tag{12}$$

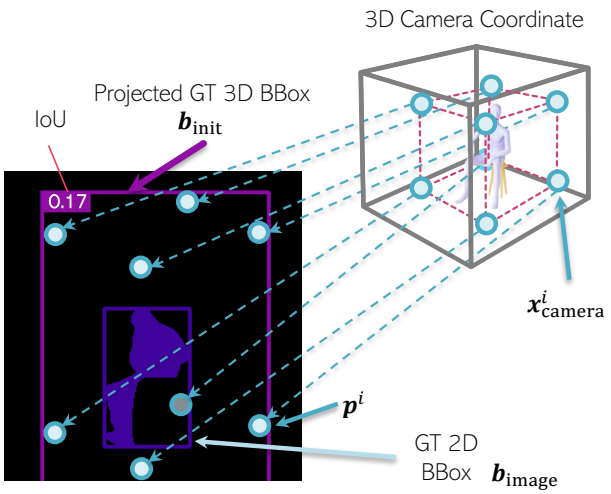


Figure 6. A direct projection of 3D BBoxes to the 2D image plane results in oversized 2D BBoxes. A learnable module is used to refine the projected BBoxes close to the 2D BBox GT.

with the scaling factor $\lambda = Z$. Thus, for each vertex, the projection onto the image plane is given by:

$$\boldsymbol{p}^i = \texttt{proj}_{\texttt{pinhole}}(\boldsymbol{x}^i_{\texttt{camera}}), \quad \text{for } i = 1, \ldots, 8, \tag{13}$$

where $\boldsymbol{p}^i = (p^i_x, p^i_y)$ represents the 2D coordinates of the projected point in the image plane. Once the eight vertices have been projected, the extreme coordinates on the image plane are determined as:

$$u_{\min} = \min_i\{p^i_x\}, \quad u_{\max} = \max_i\{p^i_x\}, \tag{14}$$

$$v_{\min} = \min_i\{p^i_y\}, \quad v_{\max} = \max_i\{p^i_y\}. \tag{15}$$

Using these extremes, the center coordinates, width, and height of the 2D BBox are computed by:

$$x_c = \frac{u_{\min} + u_{\max}}{2}, \quad y_c = \frac{v_{\min} + v_{\max}}{2}, \tag{16}$$

$$w = u_{\max} - u_{\min}, \quad h = v_{\max} - v_{\min}. \tag{17}$$

Thus, the final 2D BBox can be obtained as:

$$\boldsymbol{b}_{\texttt{init}} = (x_c, y_c, w, h)\,. \tag{18}$$

The 2D BBoxes obtained by projection, as shown by the purple box $\boldsymbol{b}_{\texttt{init}}$ in Fig. 6, are often too large. This occurs because projecting the eight vertices $\boldsymbol{x}^i_{\texttt{camera}}$ captures the depth information from the camera, which causes both the near and far parts of the object to be displayed in a 3D manner. As a result, to accurately predict the 2D BBox $\boldsymbol{b}_{\texttt{image}}$ on the image plane, we must use a refinement module. This module reduces the size of the initial BBox, as illustrated by the blue boxes in Fig. 6.

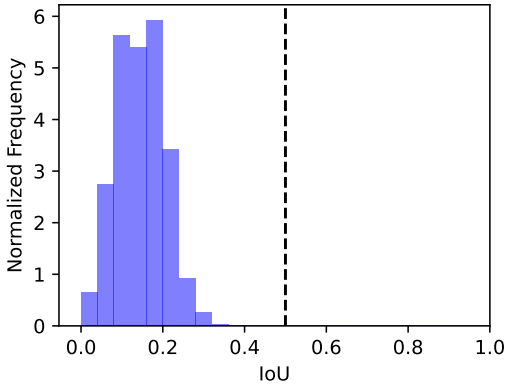


Figure 7. IoU histogram when no image plane supervision. Almost all IoU values are lower than 0.5, resulting in 0 AP.

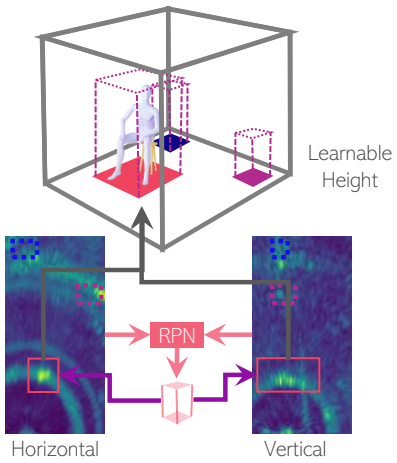


Figure 8. 3D Proposals with RFMask3D.

To better understand the need for refinement, we calculated the Intersection over Union (IoU) between the ground-truth (GT) 3D BBoxes (projected from the 3D space) and the GT 2D BBoxes (defined on the image plane). The histogram of IoU values in Fig. 7 shows a roughly Gaussian distribution with a peak around 0.15, and nearly all IoU values are below 0.5. In fact, in Fig. 6, the IoU is 0.17. This indicates that if we do not apply refinement, even when the 3D BBoxes are correctly predicted in the radar coordinate system, the average precision (AP) on the image plane would be zero. Therefore, our REXO method uses a refinement module.

D. Baselines

RFMask, DETR, and RETR Since RFMask [37] and DETR [5] 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 [40], 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.

RFMask3D As one of the baselines in our evaluation experiments, we constructed RFMask3D by extending RFMask [37] to 3D. RFMask uses a region proposal network (RPN) to extract regions of interest (RoIs) from a horizontal heatmap based on 2D anchor boxes and predicts 3D BBoxes in the 3D radar coordinate system by combining them with fixed heights. By designing an RPN that uses 3D anchor boxes, we explicitly extract 3D RoIs from both horizontal and vertical heatmaps, as shown in Fig. 8, enabling the estimation of 3D BBoxes. Unlike RFMask, this method allows for the learning of height as well.

E. Hyperparameters for Performance Evaluation

The hyper-parameters used in our experiments of Section 3 are shown in Table 2. The table is divided into three parts, Data, Model, and Training, each with parameter names, notations, and values for each dataset.

F. Definition of Metrics

Mean Intersection over Union: We adopt average precision on intersection over union (IoU) [10] as an evaluation metric. IoU is the ratio of the overlap to the union of a predicted BBox A and annotated BBox B as:

$$\texttt{IoU}\left(A, B\right) = \frac{|A \bigcap B|}{|A \bigcup B|}. \tag{19}$$

Average Precision: Average Precision (AP) can then be defined as the area under the interpolated precision-recall curve, which can be calculated using the following formula:

$$\texttt{AP} = \sum_{i=1}^{n-1} \left(r_{i+1} - r_i\right) p_{\texttt{interp}}\left(r_{i+1}\right), \tag{20}$$

$$p_{\texttt{interp}}\left(r\right) = \max_{r' \geq r} p\left(r'\right), \tag{21}$$

where the interpolated precision $p_{\texttt{interp}}$ at a certain recall level r is defined as the highest precision found for any recall level $r' \geq r$. We present three variants of average precision: $\texttt{AP}_{50}$, $\texttt{AP}_{75}$, and $\texttt{AP}$, where the former two represent the loose and strict constraints of IoU, while $\texttt{AP}$ is the averaged score over 10 different IoU thresholds in $[0.5, 0.95]$ with a step size of 0.05.

Average Recall: Average recall (AR) [19] between 0.5 and 1 of $\texttt{IoU overlap threshold}$ can be computed by averaging over the overlaps of each annotation gt_i with the closest matched proposal, that is integrating over the y : $\texttt{recall}$ axis of the plot instead of the x : $\texttt{IoU overlap threshold}$ axis. Let o be the $\texttt{IoU}$ overlap and $\texttt{recall}\left(o\right)$ the function. Let $\texttt{IoU}\left(\text{gt}_i\right)$ denote the $\texttt{IoU}$ between the annotation gt_i and the closest detection proposal:

$$\texttt{AR} = 2 \int_{0.5}^{1} \texttt{recall}(o) \text{d}o \tag{22}$$

$$= \frac{2}{n} \sum_{i=1}^{n} \max\left(\texttt{IoU}\left(\text{gt}_i\right) - 0.5, 0\right). \tag{23}$$

The following are some variations of $\texttt{AR}$:
- $\texttt{AR}_1$: $\texttt{AR}$ given one detection per frame;
- $\texttt{AR}_{10}$: $\texttt{AR}$ given 10 detection per frame;
- $\texttt{AR}_{100}$: $\texttt{AR}$ given 100 detection per frame.

G. Analysis of Failure Cases:

We provide failure cases in Fig. 9. These are all results of "Unseen," which means the environment that is not included in the training data (d8). As with d8s1 and d8s3, REXO may sometimes predict inaccurate positions, although less frequently than RETR and RFMask. In addition, there are cases where false negatives occur, such as with d8s2, d8s4, d8s5, and d8s6. In particular, it is thought to be difficult to capture the characteristics of individuals that are far away from the radar, such as with d8s2, because the resolution becomes coarse. In addition, REXO frequently gets false positives such as d8s2 - d8s6, so adjusting the threshold is important.

Table 2. Details of hyper-parameters. Fixed height for the HIBER dataset depends on the environment.

	Name	Notation	Value			
			P1S1	P1S2	P2S1	P2S2
Data	# of training	-	86579	70266	190441	118280
	# of validation	-	10538	24398	23899	33841
	# of test	-	10785	13238	23458	85677
	Input radar heatmap size	$H \times W$	256×128	256×128	256×128	256×128
	Segmentation mask size	$H \times W$	240×320	240×320	240×320	240×320
	Resolution of range	cm	11.5	11.5	11.5	11.5
	Resolution of azimuth	deg.	1.3	1.3	1.3	1.3
	Resolution of elevation	deg.	1.3	1.3	1.3	1.3
	Scale	-	log	log	log	log
Model	Backbone	-	ResNet18	ResNet18	ResNet18	ResNet18
	# of input consecutive radar frames	M	4	4	4	4
	Extracted feature map size	$H/s \times W/s$	64×32	64×32	64×32	64×32
	The number of BBoxes	$N_{\texttt{train}}$	10	10	10	10
	Threshold for detection	-	0.5	0.5	0.5	0.5
	Loss weight for `GIoU` on radar coordinate system	$\lambda_{\texttt{GIoU}}$	2.0	2.0	2.0	2.0
	Loss weight for `GIoU` on image plane	$\lambda_{\texttt{GIoU}}$	2.0	2.0	2.0	2.0
	Loss weight for $\texttt{L}_1$ on radar coordinate system	$\lambda_{\texttt{L}_1}$	5.0	5.0	5.0	5.0
	Loss weight for $\texttt{L}_1$ on image plane	$\lambda_{\texttt{L}_1}$	5.0	5.0	5.0	5.0
	Loss weight for `radar`	$\lambda_{\texttt{3D}}$	1.0	1.0	1.0	1.0
	Loss weight for `image`	$\lambda_{\texttt{2D}}$	1.0	1.0	1.0	1.0
Training	Batch size	-	32	32	32	32
	Epoch for detection	-	100	100	100	100
	Patience for early stopping	-	5	5	5	5
	Check val every n epoch for early stopping	-	2	2	2	2
	Optimizer	-	AdamW	AdamW	AdamW	AdamW
	Learning rate	-	1e-4	1e-4	1e-4	1e-4
	Sheduler	-	Cosine	Cosine	Cosine	Cosine
	Maximum number of epochs for sheduler	-	100	100	100	100
	Weight decay	-	1e-3	1e-3	1e-3	1e-3
	# of workers	-	8	8	8	8
	GPU (NVIDIA)	-	A40	A40	A40	A40
	# of GPUs	-	1	1	1	1
	Approximate training time	day	1	1	2	2

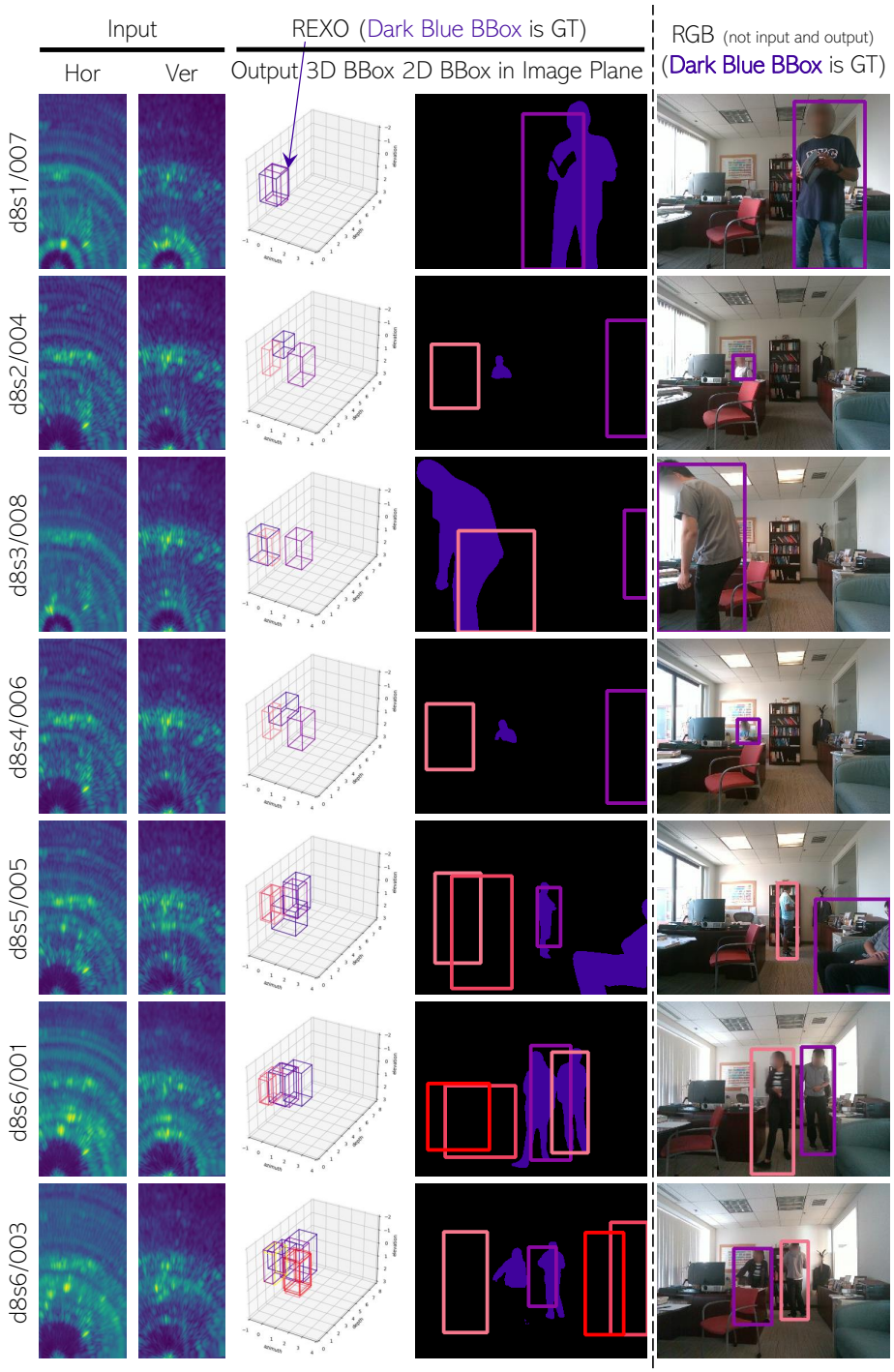


Figure 9. Visualization of failure cases. Each row indicates the segment name used from the P2S2 test dataset.