

Spatially Stratified Distillation for Heterogeneous Radar Place Recognition

Sagun Singh Shrestha¹, Samuel Harding², Abdelwahed Khamis¹, Saimunur Rahman¹, Peyman Moghadam^{1,2}

Abstract—Scalable, all-weather place recognition increasingly relies on heterogeneous radar place recognition to bridge diverse hardware platforms. A notable application is matching queries from cost-effective 4D automotive radars against high-fidelity reference maps built by dense spinning radars. This process is fundamentally limited by the extreme sparsity (and narrow field-of-view) of the 4D sensor, which captures only a fraction of the structural density present in the spinning radar database. Prior efforts address this issue by unifying different radar signals. That is, projecting both signals into a common representational space. Yet, they suffer performance degradation in multi-session environments. In this paper, we propose spatially-stratified distillation (SSD); a strategy that replaces standard uniform distillation with an asymmetric spatial alignment derived directly from physical radar returns. In regions where both radars exhibit overlapping returns, SSD enforces strong feature alignment. Crucially, in sparse regions where the 4D student lacks returns but the teacher contains valid structure within the shared field of view, SSD applies heavily discounted distillation weights. Extensive evaluations of the recent HeRCULES dataset demonstrate that SSD significantly outperforms prior place recognition methods, achieving state-of-the-art results on its challenging dynamic sequences.

Keywords: Spinning Radar, 4D Radar, Cross-modal Distillation, Radar Place Recognition.

I. INTRODUCTION

Robust autonomous navigation relies on place recognition [1]–[4]. Radar sensors are highly attractive for this task due to their immunity to adverse weather and lighting conditions that often compromise cameras and LiDAR. While high-fidelity environmental reference maps are typically constructed using expensive, dense 360° spinning radars, mass-market vehicles are increasingly equipped with cost-effective, solid-state 4D automotive radars. Consequently, scalable localization requires *heterogeneous radar place recognition*: matching sparse 4D radar queries against dense spinning radar databases. However, this process is fundamentally bottlenecked by severe modality asymmetry. As shown in Fig. 1, the 4D radar sensor captures a sparse 120° view of the scene, whereas the spinning radar captures a dense 360° panorama of the entire scene.

SHeRLoc [5], a recent work, bridges this cross-modal gap by projecting both data types into a synchronized polar representation and extracting features via a shared backbone. While effective for single-session retrieval, multi-session retrieval, where the map and query are recorded on different days along the same route, remains challenging (Sec. III-B).

4D radar



spinning radar



Fig. 1: Radar place recognition challenge addressed in this paper via SSD. The 4D radar view on the left is a sparse, narrow-field observation of the same scene shown by the spinning radar on the right. *Question: Can you identify where the 4D radar view fits inside the spinning-radar scan?* The answer is revealed on the next page.

The principal source of difficulty is the severe sparsity of the 4D radar queries compared to the dense spinning radar database: a shared-backbone metric learning based solution naturally aligns features where both sensors observe identical structures, but it does not attempt to leverage the dense structural context present in the spinning radar map that is physically absent from the 4D radar query.

To bridge this gap, we propose *Spatially-Stratified Distillation* (SSD), which reformulates cross-modal distillation specifically for asymmetric sensing. SSD operates through three core mechanisms designed to inject dense structural supervision into the sparse 4D student without violating physical sensing constraints. Specifically, we derive a distillation objective that strongly aligns jointly observed regions while utilizing unobserved, teacher-only regions as a heavily discounted structural prior. This elevates the 4D radar student learned representation while at the same time preventing it from hallucinating unobservable structures.

By targeting the regions where sparsity is worst, SSD provides an external teaching signal exactly where the student needs it most, significantly improving single-session and multi-session retrieval performance without adding complexity at inference time. Our core contributions are:

- 1) **Spatially-Stratified Distillation (SSD):** We propose a cross-modal framework bridging sparse 4D and dense spinning radars. SSD employs an *asymmetric physical mask* to leverage dense teacher context without forcing the student to hallucinate unobservable structures.
- 2) **State-of-the-art Performance:** On the HeRCULES benchmark, SSD achieves state-of-the-art and outper-

¹ CSIRO Robotics, Brisbane, Australia

² University of Queensland, Brisbane, Australia

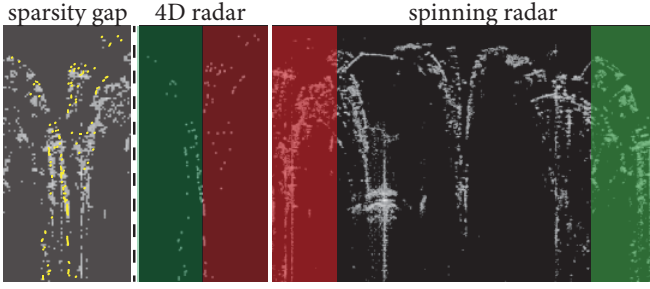


Fig. 2: **Right:** Highlighted region locating the sparse 4D radar view within the dense spinning-radar scan. **Left:** 4D radar returns (yellow colored) overlaid on the spinning-radar region. Although acquired simultaneously from the same scene, the modalities exhibit a disparity in measurement density in the polar RCS representation, where horizontal and vertical axes denote azimuth and range, respectively.

forms best performing baselines [5] on all single-session and multi-session sequences in heterogeneous (4D query vs. spinning database, AR@1 0.766 vs. 0.722) and homogeneous (4D query vs. 4D database, AR@1 0.954 vs 0.905) settings.

II. PROPOSED METHOD

Notation. We use s for the student branch operating on 4D radar and t for the teacher branch operating on spinning radar. Let $\mathbf{I}_s \in \mathbb{R}^{1 \times H \times W}$ and $\mathbf{I}_t \in \mathbb{R}^{1 \times H \times W}$ be the polar BEV inputs, and let $\phi_s, \phi_t : \mathbb{R}^{1 \times H \times W} \rightarrow \mathbb{R}^{C \times H_f \times W_f}$ be the corresponding backbones, producing feature maps $\mathbf{F}_s = \phi_s(\mathbf{I}_s)$ and $\mathbf{F}_t = \phi_t(\mathbf{I}_t)$ at stride $r = H/H_f$. We index spatial positions by $\mathbf{p} = (u, v) \in \Omega := \{1, \dots, H_f\} \times \{1, \dots, W_f\}$ and write $\mathbf{F}[\mathbf{p}] \in \mathbb{R}^C$ for the channel vector at \mathbf{p} . The per-location unit vector is $\hat{\mathbf{F}}[\mathbf{p}] := \mathbf{F}[\mathbf{p}] / \|\mathbf{F}[\mathbf{p}]\|_2$. We use $\mathbb{1}[\cdot]$ for the indicator function, G_σ for a 2D Gaussian kernel of bandwidth σ , and $*$ for 2D convolution. Tensors are written for a single sample throughout.

Base pipeline. Given a 4D-radar query frame and a spinning-radar database frame, polar BEV inputs $\mathbf{I}_s, \mathbf{I}_t \in \mathbb{R}^{1 \times H \times W}$ are produced with a matching 120° FOV following [5]. The feature extraction backbones produce local feature maps $\mathbf{F}_s = \phi_s(\mathbf{I}_s)$, $\mathbf{F}_t = \phi_t(\mathbf{I}_t) \in \mathbb{R}^{C \times H_f \times W_f}$ with $C = 256$ and $(H_f, W_f) \approx (16, 80)$ after $8 \times$ downsampling. Optimal transport based HOLMES pooling [5] was used to aggregate each local map to a global descriptor $\mathbf{d} \in \mathbb{R}^D$ with $D = 320$. Training uses FOV-aware adaptive-margin triplet loss $\mathcal{L}_{\text{trip}}(\mathbf{d}_q, \mathbf{d}_p, \mathbf{d}_n)$ [5], where the query \mathbf{d}_q comes from the student branch (4D radar) and the positives & negatives $\mathbf{d}_p, \mathbf{d}_n$ come from the teacher branch (spinning radar).

A. Spatially Stratified Distillation (SSD)

As illustrated in Fig. 3, we use the dense spinning radar teacher to enrich the learned 4D features. However, the teacher captures contextual cues physically unavailable to the sparse 4D student. This informational disparity makes exact feature alignment infeasible. Instead, we treat the detached teacher as a structural prior [6], aligning intermediate

activations via:

$$\mathcal{L}_{\text{FD}} = \frac{\sum_{\mathbf{p} \in \Omega} W(\mathbf{p}) d(\hat{\mathbf{F}}_s[\mathbf{p}], \hat{\mathbf{F}}_t[\mathbf{p}])}{\sum_{\mathbf{p} \in \Omega} W(\mathbf{p}) + \varepsilon}, \quad (1)$$

where the spatial mask $W(\mathbf{p})$ dictates the distillation weight and $d(\cdot, \cdot)$ measures feature discrepancy. Crucially, applying a uniform spatial mask is physically invalid due to the severe field-of-view mismatch. Uniform alignment forces the student to hallucinate unobservable structures, whereas completely masking unobserved regions discards the dense contextual prior. On the other hand, masking out unobserved regions entirely would discard the contextual prior that makes distillation valuable in the first place.

To resolve this asymmetry, we propose **Spatially-Stratified Distillation (SSD)**, which reformulates the distillation objective to account for both the sensing geometry and the modality differences of the problem. Specifically, SSD introduces a spatial weighting function $W(\mathbf{p})$ that distinguishes between regions of joint observability and regions unique to the teacher:

$$W(\mathbf{p}) = w_h \mathcal{M}_{\text{joint}}(\mathbf{p}) + w_\ell \mathcal{M}_{\text{gap}}(\mathbf{p}), \quad (2)$$

where $w_\ell \ll w_h$. $\mathcal{M}_{\text{joint}}$ represents the *Joint Observations Mask* where both the student and teacher have physical measurement support. \mathcal{M}_{gap} is the *Unobserved Context Mask*. This mask isolates locations where the teacher contains meaningful structural information (\mathcal{M}_t) but the student lacks sufficient physical support (student evidence is below a threshold τ_s). This weighting is central to our formulation. It preserves strong supervision where spatial correspondence is physically credible, while using teacher-only regions as a weak regularizing prior rather than as a hard matching target. The masks are leveraged as follows:

$$\mathcal{M}_{\text{joint}}(\mathbf{p}) = \mathcal{M}_s(\mathbf{p}) \mathcal{M}_t(\mathbf{p}). \quad (3)$$

$$\mathcal{M}_{\text{gap}}(\mathbf{p}) = (1 - \mathbb{1}[\mathcal{M}_s(\mathbf{p}) > \tau_s]) \mathcal{M}_t(\mathbf{p}), \quad (4)$$

Importantly, \mathcal{M}_{gap} does not defy physical plausibility; it simply captures teacher-supported locations that **fall into the student’s FoV gap** without forcing an impossible exact match. Rather than discarding these locations, we incorporate them with a discounted weight. The mechanism also ensures that the student is never penalized as if it had direct access to unobserved content, while remaining weakly regularized toward representations compatible with the teacher’s richer global scene structure.

Mask Generation. To effectively isolate these regions, the underlying masks (\mathcal{M}_t and \mathcal{M}_s) must accurately reflect the distinct informational roles of each branch. For the teacher branch, which processes a dense scene representation, we use a data-adaptive magnitude mask applied directly to the teacher’s feature maps: $\mathcal{M}_t(\mathbf{p}) = \mathbb{1}[\|\mathbf{F}_t[\mathbf{p}]\|_2 > \alpha \bar{F}_t]$ where $\bar{F}_t = \frac{1}{|\Omega|} \sum_{\mathbf{p}' \in \Omega} \|\mathbf{F}_t[\mathbf{p}']\|_2$ is the spatial feature magnitude.

Unlike prior works [6] that rely on intermediate features, which decouple from physical evidence due to convolutional

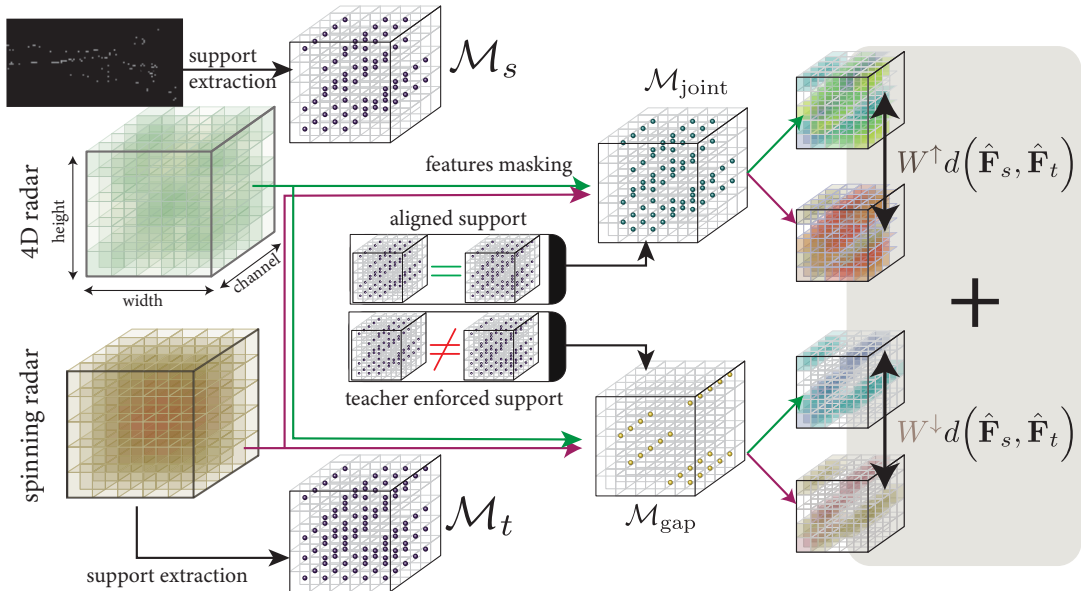


Fig. 3: **Overview of Spatially-Stratified Distillation (SSD)**. The distillation pipeline explicitly accounts for the field-of-view asymmetry between the sparse 4D radar student and the dense spinning radar teacher. First, support extraction isolates physical measurements (\mathcal{M}_s) and meaningful teacher activity (\mathcal{M}_t). These masks are combined to stratify the spatial domain into two regimes. In regions of aligned support ($\mathcal{M}_{\text{joint}}$), the feature alignment loss is strictly enforced with a high weight (W^\uparrow). Conversely, teacher-supported locations falling into the student’s unobserved field-of-view (\mathcal{M}_{gap}) are not discarded; they are distilled with a heavily discounted weight (W^\downarrow) to serve as a weak structural prior, ensuring the student is guided by global context without being penalized for physical sensing limitations.

diffusion, we derive the student mask directly from the down-sampled 4D input \mathbf{I}_s^\downarrow as $\mathcal{M}_s(\mathbf{p}) = (G_\sigma * \mathbb{1}[\mathbf{I}_s^\downarrow > 0]) (\mathbf{p})$, where G_σ is a Gaussian kernel.

Activation alignment. Finally, to address activation scale and correlation mismatches between branches, we align the student channel basis using a 1×1 convolution ψ and apply L_2 -normalization. We supervise the normalized features using cosine distance, $d(\mathbf{u}, \mathbf{v}) = 1 - \langle \mathbf{u}, \mathbf{v} \rangle$, and further align their second-order structure through $\mathcal{L}_{\text{cov}} = \|\text{cov}(\mathbf{u}) - \text{cov}(\mathbf{v})\|_F^2$. The resulting feature-distillation loss incorporates the spatial weighting $W(\mathbf{p})$ into the general objective (1) and is optimized jointly with the primary place-recognition triplet loss:

$$\mathcal{L} = \mathcal{L}_{\text{trip}} + \lambda_{\text{FD}} \mathcal{L}_{\text{FD}} + \lambda_{\text{cov}} \mathcal{L}_{\text{cov}}, \quad (5)$$

where λ_{FD} and λ_{cov} denote the weights of the feature-distillation and covariance-alignment losses, respectively.

III. EXPERIMENTAL EVALUATION

We evaluate SSD on the HeRCULES dataset [7], following the SHerLoc protocol [5]. Training is performed using *Mountain 01–03*, *Bridge 01*, *Stream 02*, and *Parking Lot 03–04* sequences, while evaluation is done on *Sports Complex 01–03*, *Library 01–03*, and *River Island 01–02* sequences in single-session and multi-session settings. Performance is measured using Recall@K (R@K) and Average Recall@K (AR@K), with predictions considered correct if they fall within 5 m of the ground truth. Our method is implemented in PyTorch, using a ResNet based backbone similar to SHerLoc, and also trained for 20 epochs on multiple GPUs.

A. Heterogeneous Single-session Place Recognition

Tab. I shows the results of SSD on single-session place recognition. In this evaluation, we use spinning radar data as the database and 4D radar data as queries across the Sports Complex, Library, and River Island sequences. The results show that SSD consistently surpasses all other methods including the very recent state-of-art baselines.

B. Heterogeneous Multi-session place recognition

We also evaluate SSD on multi-session evaluation. In this setting, spinning radar data from sequence 01 were used as the database, and 4D radar data from sequences 02 and 03 were used as queries. We show the results on Tab. II. The results show that SSD surpasses all methods including the both SHerLoc variants in a consistent manner by a significant margin (more than 4% over the best art).

C. Homogeneous Multi-session place recognition

Since SSD targets at improving 4D radar representation, we conduct homogeneous multi-session place recognition experiments to understand whether the learned improved representation applies beyond heterogeneous settings. Tab. II shows the results. It is interesting to see that both Sports Complex and Library results have been improved significantly. This demonstrate that SSD indeed improved 4D representation in a favourable manner.

D. Qualitative analysis.

Fig. 4 illustrates SSD’s robust retrieval performance on the challenging SC/01→SC/03 multi-session pair, where it recovers 407 queries while incurring only 297 regressions

TABLE I: Single-session place recognition on HeRCULES dataset. SSD results are mean over three runs.

Methods	Sports Complex						Library						River Island					
	01		02		03		01		02		03		01		02		03	
	R@1	R@1%	R@1	R@1%	R@1	R@1%	R@1	R@1%	R@1	R@1%	R@1	R@1%	R@1	R@1%	R@1	R@1%	R@1	R@1%
Radar SC [8]	0.047	0.135	0.042	0.061	0.044	0.064	0.040	0.060	0.041	0.063	0.042	0.066	0.010	0.020	0.004	0.012	0.005	0.030
RaPlace [9]	0.042	0.148	0.056	0.156	0.025	0.152	0.035	0.069	0.040	0.090	0.016	0.123	0.002	0.152	0.000	0.024	0.004	0.162
RadVLAD [10]	0.019	0.334	0.051	0.222	0.022	0.426	0.011	0.237	0.014	0.295	0.012	0.245	0.001	0.217	0.066	0.543	0.006	0.367
FFT-RadVLAD [10]	0.040	0.344	0.014	0.186	0.022	0.377	0.023	0.287	0.014	0.273	0.011	0.229	0.002	0.330	0.026	0.533	0.001	0.488
Autoplace [11]	0.015	0.198	0.046	0.098	0.017	0.186	0.010	0.129	0.031	0.131	0.018	0.122	0.003	0.051	0.001	0.009	0.001	0.021
SHerLoc-S [5]	0.857	0.924	0.975	0.985	0.945	0.981	0.866	0.917	0.925	0.950	0.850	0.891	0.899	0.965	0.868	0.963	0.850	0.945
SHerLoc [5]	0.900	0.936	0.962	0.987	0.958	0.980	0.881	0.936	0.938	0.964	0.868	0.912	0.880	0.957	0.860	0.952	0.858	0.959
SSD (Ours)	0.922	0.971	0.975	0.992	0.974	0.998	0.900	0.968	0.953	0.980	0.894	0.966	0.939	0.991	0.926	0.988	0.890	0.975

TABLE II: Multi-session place recognition on HeRCULES. Heterogeneous: 4D query vs spinning database; Homogeneous-4D: 4D query vs 4D database (student-only evaluation of the SSD-trained backbone). R@1 at 5 m radius; R@1% reports recall of top 1%. AR@1 = mean over the four pairs. SSD results are mean over three runs.

Methods	Sports Complex				Library				AR@1	
	01 → 02		01 → 03		01 → 02		01 → 03			
	R@1	R@1%	R@1	R@1%	R@1	R@1%	R@1	R@1%		
Heterogeneous	RaPlace [9]	0.024	0.056	0.016	0.057	0.021	0.044	0.011	0.023	0.018
	RadVLAD [10]	0.011	0.118	0.005	0.214	0.012	0.138	0.011	0.073	0.010
	FFT-RadVLAD [10]	0.006	0.121	0.007	0.182	0.015	0.217	0.016	0.136	0.011
	Autoplace [11]	0.007	0.071	0.012	0.070	0.022	0.129	0.007	0.119	0.012
SHerLoc-S [5]	0.796	0.890	0.580	0.696	0.822	0.892	0.618	0.757	0.704	
SHerLoc [5]	0.812	0.893	0.650	0.759	0.817	0.887	0.610	0.743	0.722	
SSD (Ours)	0.838	0.907	0.700	0.818	0.863	0.948	0.663	0.858	0.766	
4D	Autoplace [11]	0.799	0.967	0.725	0.945	0.812	0.986	0.619	0.901	0.738
	MinkLoc3Dv2 [12]	0.837	0.982	0.743	0.977	0.725	0.981	0.619	0.963	0.735
	TransLoc4D [13]	0.833	0.970	0.804	0.976	0.801	0.991	0.676	0.940	0.779
	SHerLoc-S [5]	0.866	0.945	0.803	0.925	0.914	0.986	0.908	0.988	0.873
SHerLoc [5]	0.904	0.961	0.843	0.949	0.950	0.995	0.923	0.988	0.905	
SSD (Ours)	0.949	0.963	0.942	0.973	0.964	0.996	0.959	0.987	0.954	

SC/01 → SC/03 → where our method gains (3226 queries total)

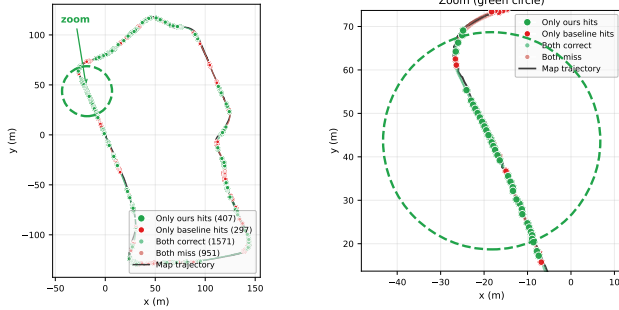


Fig. 4: Per-query difference between SSD and SHerLoc on SC/01→SC/03. **Green:** SSD-only hits (407); **Red:** SHerLoc-only hits (297); **Dots:** agreement. **Right:** zoom on the densest win cluster along a curved sub-route.

relative to SHerLoc. As shown, these performance gains are concentrated on a curved sub-route where sparse returns under-constrain place identity. In the most severe failure cases, the baseline predicts locations over 240 m away, while SSD keeps it within a 1.2 m error margin.

IV. CONCLUSION

We presented Spatially-Stratified Distillation (SSD), a cross-modal distillation framework that bridges the sparsity

gap between 4D and spinning radar for place recognition. Extensive evaluations on the HeRCULES dataset demonstrate that SSD consistently outperforms state-of-the-art baselines across all evaluated single-session and multi-session environments. By effectively bridging the structural gap between dense database maps and sparse queries, we hope this work encourages further exploration into asymmetric cross-modal supervision for cost-effective 4D radar localization.

V. ACKNOWLEDGMENT

The authors would like to acknowledge the collaboration with Boeing.

REFERENCES

- [1] K. Vidanapathirana, M. Ramezani, P. Moghadam, S. Sridharan, and C. Fookes, “Logg3d-net: Locally guided global descriptor learning for 3d place recognition,” in *IEEE ICRA*, pp. 2215–2221, IEEE, 2022.
- [2] K. Vidanapathirana, P. Moghadam, S. Sridharan, and C. Fookes, “Spectral geometric verification: Re-ranking point cloud retrieval for metric localization,” *IEEE RA-L*, vol. 8, no. 5, pp. 2494–2501, 2023.
- [3] S. Hausler and P. Moghadam, “Pair-vpr: Place-aware pre-training and contrastive pair classification for visual place recognition with vision transformers,” *IEEE RA-L*, 2025.
- [4] J. Knights, J. Reid, K. Roy, D. Hall, M. Cox, and P. Moghadam, “Wildcross: A cross-modal large scale benchmark for place recognition and metric depth estimation in natural environments,” *arXiv preprint arXiv:2603.01475*, 2026.
- [5] H. Kim, M. Jung, W. Yang, and A. Kim, “Sherloc: Synchronized heterogeneous radar place recognition for cross-modal localization,” *IEEE RA-L*, 2025.
- [6] G. Bang, K. Choi, J. Kim, D. Kum, and J. W. Choi, “Radardistill: Boosting radar-based object detection performance via knowledge distillation from lidar features,” in *IEEE/CVF CVPR*, 2024.
- [7] H. Kim *et al.*, “Hercules: A heterogeneous radar and lidar dataset for autonomous driving,” *arXiv preprint arXiv:2502.01946*, 2025.
- [8] G. Kim, Y. S. Park, Y. Cho, J. Jeong, and A. Kim, “Mulran: Multimodal range dataset for urban place recognition,” in *IEEE ICRA*, pp. 6246–6253, IEEE, 2020.
- [9] H. Jang, M. Jung, and A. Kim, “Raplace: Place recognition for imaging radar using radon transform and mutable threshold,” in *IEEE/RSJ IROS*, pp. 11194–11201, IEEE, 2023.
- [10] M. Gadd and P. Newman, “Open-radvlad: Fast and robust radar place recognition,” in *IEEE Radar Conference (RadarConf24)*, pp. 1–6, IEEE, 2024.
- [11] K. Cai, B. Wang, and C. X. Lu, “Autoplace: Robust place recognition with single-chip automotive radar,” in *IEEE ICRA*, pp. 2222–2228, IEEE, 2022.
- [12] J. Komorowski, “Improving point cloud based place recognition with ranking-based loss and large batch training,” in *ICPR*, pp. 3699–3705, IEEE, 2022.
- [13] L. Zhang *et al.*, “Transloc4d: Transformer-based 4d radar place recognition,” in *IEEE/CVF CVPR*, 2024.