Supplementary Material for "SHeRLoc: Synchronized Heterogeneous Radar Place Recognition for Cross-Modal Localization"

IEEE Robotics and Automation Letters

Hanjun Kim[1], Minwoo Jung[2], Wooseong Yang[2] and Ayoung Kim[2*]

[1]hanjun815@snu.ac.kr
[2][moonshot, yellowish, ayoungk]@snu.ac.kr

0. Motivation

As radar technology advances, multiple radar types have emerged, including spinning radars, X-band surveillance radars, and automotive 4D radars (see Fig A). Despite this growing diversity, most existing radar place recognition methods are designed for homogeneous settings, assuming the same radar type is used for both query and database. As illustrated in Fig B, from our submitted multimedia, this limits their generalizability and robustness when faced with intra-modality heterogeneity in range sensors.

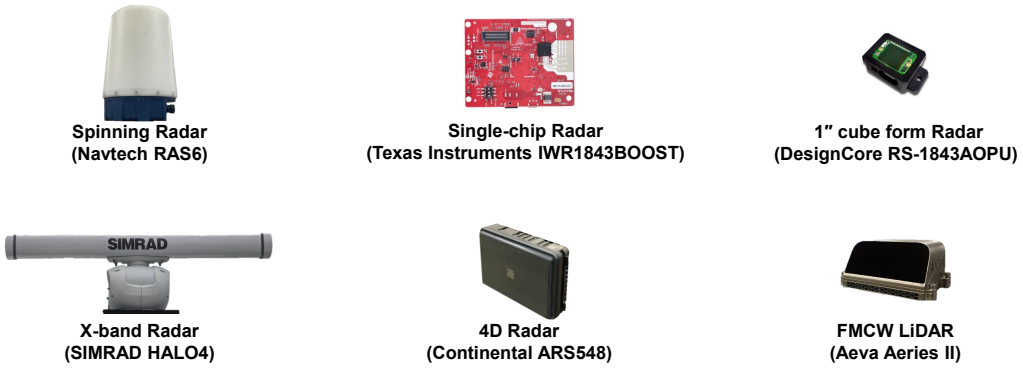

Figure A: Overview of various radar types and associated manufacturers.

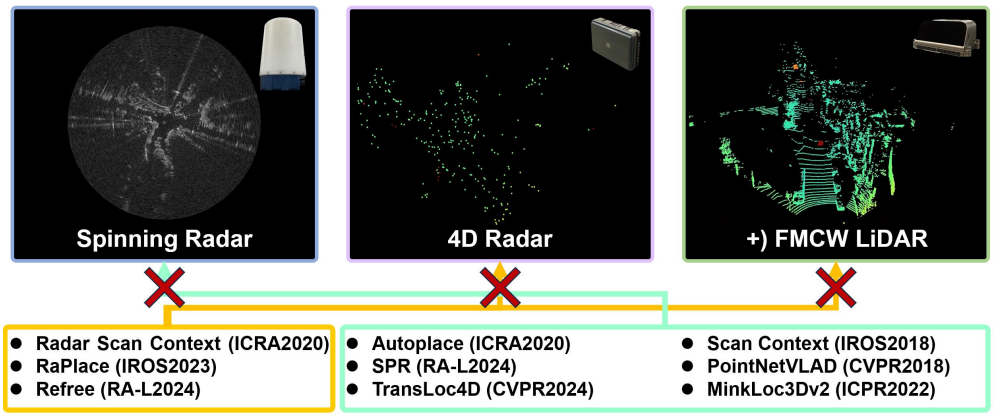

Figure B: Homogeneous radar place recognition models cannot be applied to different sensor types.

To highlight the benefits of our heterogeneous radar framework, we provided qualitative results in Fig. 6, where existing homogeneous models fail under the heterogeneous multi-session setup.

Moreover, as demonstrated in Fig. 5, the spinning radar's capability to build comprehensive mapping databases, combined with the 4D radar's real-time dynamic sensing, showcases the strengths of each sensor type in heterogeneous radar systems. This highlights the potential of our framework for handling challenging scenarios where the database and query are collected by different radar types.

1. Preprocessing

We detail the refinement function used to remove clutter and ground returns. Given a raw 4D radar scan

$$\mathcal{S}_{\text{4D},k} = \{(x_i, y_i, z_i, v_i^d, \sigma_i)\}_{i=1}^{N},$$

the refinement function f_{removal} eliminates spurious points caused by multipath reflections, ground-plane returns, or noise. The refined set is defined as:

$$\mathcal{S}'_{\text{4D},k} = f_{\text{removal}}(\mathcal{S}_{\text{4D},k}) = \{(x_i, y_i, z_i, v_i^d, \sigma_i) \in \mathcal{S}_{\text{4D},k} \mid v_i^d < \tau_v,\ z_i \geq \tau_z,\ \sigma_i \geq \tau_\sigma\}.$$

In practice:

- Since the radar is mounted at the front lower part of the vehicle, points with vertical coordinate below $\tau_z = -1$ m are regarded as ground reflections and removed.
- Points with RCS smaller than $\tau_\sigma = -10$ dBsm are discarded as low-intensity clutter, as they are unlikely to correspond to meaningful structures.
- Points with Doppler velocity exceeding $\tau_v = 5$ m/s are filtered out to suppress fast-moving artifacts not relevant to place recognition.

This thresholding strategy effectively suppresses ground-plane reflections, random clutter, and spurious high-velocity points, while preserving stable returns from meaningful objects.

2. Correction Term C_{corr}

We clarify that C_{corr} is determined primarily by the physical characteristics of the radar sensor (e.g., antenna gain and transmitted power), rather than by environmental factors. Therefore, it is environment-independent and does not require re-estimation across different scenes or viewpoints. In our implementation, we simply used a subset of the `mountain 01` sequence for convenience, but any dataset could be used since the correction depends only on sensor properties. More specifically, we derive C_{corr} by matching the measured return power P_r with the Radar Cross Section (RCS) σ via the classical radar equation:

$$P_r = \frac{P_t \cdot G^2 \cdot \lambda^2 \cdot \sigma}{(4\pi)^3 \cdot R^4}, \tag{1}$$

where P_t is the transmitted power, G is the antenna gain, λ is the wavelength, and R is the range. Modern spinning radars adopt a cosecant-squared beam profile, effectively compensating for the R^4 attenuation, which allows the RCS to be expressed in decibel scale as:

$$\sigma_{\text{dBsm}} = P_r\,[\text{dB}] + C, \tag{2}$$

where C is the correction constant. As such, C_{corr} only needs to be re-estimated when a new radar sensor (with different P_t, G, or λ) is introduced, not when operating in new environments.

3. Temporal Max Pooling

Before temporal max pooling, we perform a refinement step that removes dynamic objects based on Doppler velocity. This preprocessing effectively suppresses moving agents, such as cars and pedestrians, ensuring their influence does not propagate into the temporally aggregated radar image. Consequently, the temporal max pooling step does not introduce motion blur from dynamic agents, as these returns have already been filtered out. This is also validated by our experimental results: despite the heavy traffic and large number of moving objects in the `River`

`Island` urban driving sequences, our method achieves consistently high recall in Tables I, III, and IV. Furthermore, Fig. 5 qualitatively demonstrates that place recognition remains robust even in scenarios with significant dynamic clutter. There are three different strategies for 4D radar scan aggregation, each with different trade-offs.

1) Max-pooling without motion compensation: As described in the paper, the simplest method directly aggregates adjacent scans via max pooling without applying motion compensation. This has the advantage of requiring no external sensors (e.g., IMU, odometry) and enables fast concatenation. Although a small degree of motion blur may occur, its effect is limited: the 4D radar operates at 20 Hz, implying a $\Delta t = 0.05\,\mathrm{s}$ interval between frames, and with a resolution of approximately 0.4 m per pixel, even at 36 km/h the inter-frame displacement corresponds to only about one pixel. For spinning radar (4 Hz), motion-induced distortion is a known issue [1, 2], but, to the best of our knowledge, all existing radar place recognition methods (e.g., [3, 4, 5]) do not employ explicit motion compensation, as the long radar wavelength inherently limits resolution and the task focuses on place recognition rather than precise geometric alignment.

2) Ground-truth compensated aggregation: When ground-truth ego-motion is available (e.g., AutoPlace [6]), scans can be concatenated using exact trajectory information. While this provides the most accurate alignment, it is impractical in real-world deployment scenarios where ground truth is not accessible.

3) Velocity-based motion compensation: Assuming constant velocity between consecutive 4D radar scans, the Cartesian displacement between scans is

$$s_{x,k-1,k} = \frac{v^e_{x,k-1}}{f_r}, \quad s_{y,k-1,k} = \frac{v^e_{y,k-1}}{f_r},$$

where f_r is the frame rate. The displacement is projected into the polar domain (r, θ) as

$$\Delta r_{k-1,k} = \frac{x \cdot s_{x,k-1,k} + y \cdot s_{y,k-1,k}}{\sqrt{x^2 + y^2}}, \quad \Delta\theta_{k-1,k} = \frac{x \cdot s_{y,k-1,k} - y \cdot s_{x,k-1,k}}{x^2 + y^2},$$

and normalized by the grid resolutions (h_r, h_θ) to yield the polar BEV coordinate offsets. Bilinear interpolation in this domain allows motion-compensated feature alignment. This approach, similar to the method used in TDFANet [7], is more practical than relying on ground-truth data but remains sensitive to velocity estimation errors, particularly during rotational motion when ego-motion is not purely translational.

4. Polar BEV vs Cartesian BEV Response

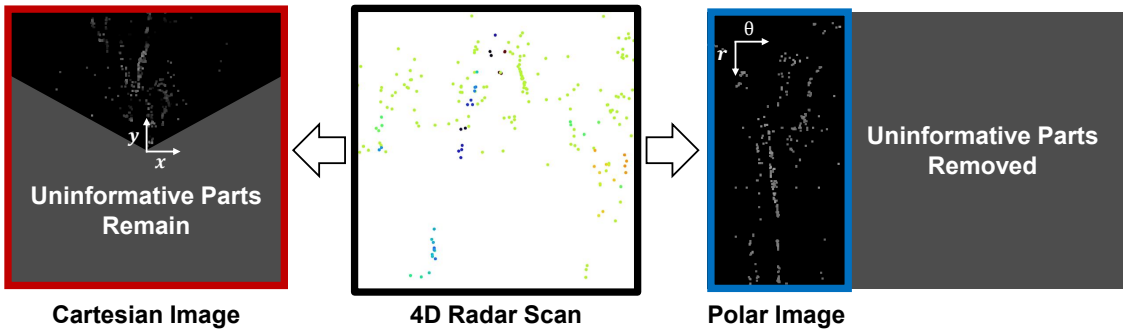

Figure C: Comparison of Cartesian representation versus Polar representation.

We clarify that the key advantage of the polar representation lies in its ability to align the field of view (FOV) between different radar modalities while discarding uninformative regions. In Cartesian projection, the FOV mismatch with spinning radar results in large empty regions (as illustrated in Fig. C), which cannot be easily cropped out because the Cartesian grid inherently

assumes rectangular coverage. These uninformative areas dilute the feature learning process and degrade recognition performance. In contrast, polar projection directly parameterizes the radar measurements in terms of range and azimuth, enabling us to (i) precisely crop the FOV to match between spinning and 4D radars, (ii) remove empty or uninformative regions outside the overlap, and (iii) ensure uniform angular sampling, which provides balanced spatial coverage for descriptor learning.

5. HOLMES: Multi-scale Descriptors with Optimal Transport

HOLMES introduces a sensor- and task-specific aggregation pipeline explicitly designed for heterogeneous range sensors. Specifically, it incorporates: (i) variance-adaptive entropy regularization to stabilize matching under noisy RCS distributions, (ii) ghostbin suppression to mitigate radar-specific artifacts, and (iii) compact yet discriminative descriptors that preserve cross-sensor robustness. These components directly address heterogeneity challenges such as noise, sparsity, and distinct intensity statistics, making HOLMES both more robust and more compact than SALAD or HeLiOS. To clarify, most traditional feature aggregation methods primarily focus on local feature pooling. For example, MAC [8] uses channel-wise maximum activations (local maxima only), SPoC [9] averages feature maps (global mean without structural context), and GeM [10] generalizes pooling with a learnable p-norm (local-to-global summarization). NetVLAD [11] indirectly captures scene-level distribution via clustering residuals, but does not employ an explicit global descriptor. In contrast, SALAD explicitly integrates a global token to enhance scene-level representation, and HeLiOS [12] similarly employs GeM pooling combined with MLP layers to incorporate global context. For clarity, we provide a comparison of representative aggregation methods in terms of their treatment of local and global contexts:

Table 1: Comparison on Feature Aggregation Method

Methods	Global Context	Size	Sports Complex				Library			
			01 - 02		01 - 03		01 - 02		01 - 03	
			R@1	R@1%	R@1	R@1%	R@1	R@1%	R@1	R@1%
NetVLAD [11]	△	32768	0.378	0.598	0.151	0.254	0.302	0.474	0.102	0.218
MAC [8]	×	512	0.700	0.846	0.404	0.534	0.525	0.657	0.258	0.514
SPoC [9]	×	512	0.282	0.517	0.112	0.210	0.155	0.258	0.036	0.150
GeM [10]	△	512	0.486	0.717	0.234	0.384	0.324	0.475	0.091	0.197
SALAD [13]	◯	8448	0.371	0.611	0.170	0.295	0.330	0.497	0.094	0.235
HOLMES	◯	**320**	**0.812**	**0.893**	**0.650**	**0.759**	**0.817**	**0.887**	**0.610**	**0.743**

Table 2: Comparison of Aggregation Methods in Terms of Local Feature Utilization and Global Context Integration

Method	**Local Feature Aggregation**	**Global Context Integration**
NetVLAD [11]	Residuals w.r.t. cluster centers	Indirect distributional info, no explicit global token
MAC [8]	Channel-wise max pooling	None
SPoC [9]	Sum/average pooling	Implicit global mean
GeM [10]	Generalized mean pooling (p-norm)	Learnable pooling, no explicit token
SALAD [13]	OT-based patch token alignment	**Explicit global token (scene descriptor)**
HeLiOS [12]	OT-based patch token alignment	**GeM pooling with MLP layers**
HOLMES	OT-based patch token alignment	**Hierarchical optimal transport, GeM pooling with MLP layers**

6. Zero-shot Generalization

Regarding generalization, we emphasize that our framework has already been evaluated in a heterogeneous range sensor PR setting, where FMCW LiDAR queries were tested against a spinning radar database. Although SHeRLoc is specifically designed for heterogeneous radar modalities, the results demonstrate that it also generalizes effectively to cross-modality scenarios involving LiDAR, achieving strong performance.

Table 3: Zero-shot Generalization Performance on Unseen Datasets

Methods	MulRan		Oxford Radar		AR@1
	DCC 01	KAIST 03	#1 to #3	#2 to #3	
Radar SC [14]	0.683	0.896	0.655	0.348	0.646
RaPlace [3]	0.641	0.932	0.666	0.437	0.669
RadVLAD [4]	0.507	0.833	0.813	0.693	0.712
FFT-RadVLAD [4]	0.706	0.948	0.666	0.427	0.687
ReFeree [5]	0.723	0.955	0.803	0.677	0.790
SHeRLoc-S	**0.732**	0.961	0.855	0.748	0.824
SHeRLoc	0.708	**0.963**	**0.900**	**0.890**	**0.865**

We also evaluated zero-shot experiments on the MulRan [14] dataset and the Oxford Radar RobotCar [15] dataset. The HeRCULES spinning radar is a Navtech RAS6, whereas MulRan employs the earlier Navtech CIR204-H model, and Oxford Radar RobotCar employs the Navtech CTS350-X model. These radars differ: RAS6 offers a more extended detection range, while CIR204-H exhibits coarser resolution. The CTS350-X used in the Oxford Radar RobotCar is an older scanning radar with lower angular resolution and a shorter maximum range than the RAS6. These differences provide diverse sources of domain shift, making both MulRan and Oxford Radar RobotCar suitable benchmarks for unseen evaluation.

For the MulRan dataset, we conducted single-session place recognition experiments on the `DCC01` and `KAIST03` sequences. For the Oxford Radar RobotCar dataset, we performed multi-session place recognition experiments by using the `2019-01-18 #3` sequence as the database and `2019-01-10 #1` and `2019-01-16 #2` as the query. The results, now included in Table 3, demonstrate that our method maintains strong recall in these challenging zero-shot settings, confirming its generalization under radar-domain shifts.

7. Triplet Loss with Adaptive Margin

Substituting Eq. 14 into Eq. 13, the triplet loss becomes:

$$\mathcal{L}_{\text{triplet}} = \max\Big(d(x_i^q, x_j^p) - \min_n d(x_i^q, x_j^n) + \gamma(\text{Sim}(\mathbf{q}, \mathbf{p}) - \text{Sim}(\mathbf{q}, \mathbf{n})), 0\Big). \tag{3}$$

Focusing on the hardest negative, this can be rewritten as:

$$\mathcal{L}_{\text{triplet}} = \max\Big([d(x_i^q, x_j^p) + \gamma\text{Sim}(\mathbf{q}, \mathbf{p})] - [d(x_i^q, x_j^n) + \gamma\text{Sim}(\mathbf{q}, \mathbf{n})], 0\Big). \tag{4}$$

Our adaptive margin modulates each triplet individually based on FOV similarity, resulting in sample-specific pull/push dynamics:

- Hard negatives with high similarity to the query receive smaller margins, forcing the network to discriminate finer details.
- Easy negatives with low similarity receive larger margins, reducing their gradient contribution and preventing overfitting to trivial cases.

A global rescaling or fixed margin cannot replicate this selective, adaptive behavior.

When the adaptive margin is not activated, we set α_{sim} to a constant value, i.e., a fixed margin is used. This setup allows a direct comparison between the fixed-margin and adaptive-margin formulations, as reported in the ablation study on SHeRLoc components. Empirically, we observe that enabling the adaptive margin improves the average performance by 8.95%, demonstrating its effectiveness in enhancing discriminability and robustness.

8. Shared Backbone vs Modality-specific Backbone

From a representation-learning perspective, different radar modalities exhibit substantial structural similarities in the spatial and temporal distributions of radar reflections. By training a single backbone, the model learns a shared feature space that promotes generalization across modalities and avoids redundant parameterization. This design is particularly advantageous in heterogeneous localization, where direct correspondence between radar modalities is required. In contrast, modality-specific backbones may overfit to modality-dependent biases and hinder cross-modal matching. Theoretical analyses also support that learning in a shared latent space reduces population risk and provides more stable representations compared to separate training.

Table 4: Performance Comparison of Shared and Modality-Specific Backbone Designs

Methods	# Backbone Params	Sports Complex				Library			
		01 - 02		01 - 03		01 - 02		01 - 03	
		R@1	R@1%	R@1	R@1%	R@1	R@1%	R@1	R@1%
SHeRLoc*	$22.3M$	0.558	0.748	0.295	0.426	0.511	0.631	0.199	0.350
SHeRLoc	**11.2M**	**0.812**	**0.893**	**0.650**	**0.759**	**0.817**	**0.887**	**0.610**	**0.743**

SHeRLoc*: SHeRLoc with modality-specific backbones.

From a practical deployment perspective, a shared backbone yields a lighter model with significantly fewer parameters and reduced training overhead. This efficiency is essential for real-time robotics applications and resource-constrained platforms, while also simplifying adaptation to new modalities without the need to train and maintain multiple distinct networks. Thus, the shared design provides both performance and scalability benefits.

9. Translation and Rotation Equivariance/Invariance

We formalize these concepts in the context of global localization using a representation function $\mathcal{F}$ that maps a scan S to a feature space.

Let $S_{\delta t}$ and $S_{\delta\theta}$ denote the scan translated by δt and rotated by $\delta\theta$, respectively. Similarly, let $\mathcal{F}_{\delta t}$ and $\mathcal{F}_{\delta\theta}$ represent the transformed feature outputs under translation and rotation. We define:

- **Translation Equivariance:** $\mathcal{F}(S_{\delta t}) = \mathcal{F}(S)_{\delta t}$. Translating the input scan results in a corresponding translation in the feature representation.
- **Rotation Equivariance:** $\mathcal{F}(S_{\delta\theta}) = \mathcal{F}(S)_{\delta\theta}$. Rotating the input scan results in a corresponding rotation in the feature representation.
- **Translation Invariance:** $\mathcal{F}(S_{\delta t}) = \mathcal{F}(S)$. Translating the scan does not change the feature representation.
- **Rotation Invariance:** $\mathcal{F}(S_{\delta\theta}) = \mathcal{F}(S)$. Rotating the scan does not change the feature representation.

As shown in Fig. D, even when the scan is rotated, the image set generated by polar-domain multi-view generation remains unchanged. Consequently, after passing through the feature extraction network $\mathcal{G}$ and the HOLMES module $\mathcal{H}$, the resulting descriptors differ only in ordering, but the descriptor set itself remains identical. Theoretically, as mentioned in the paper, an infinite number of multi-views would be required, and aliasing may occur due to downsampling. However, we demonstrate in Section IV that robust performance can be achieved with a limited number of multi-views.

Furthermore, we summarize in Table 5 the equivariance and invariance properties of each component (f_{rcs}, $f_{\text{multiview}}$, CNN-based $\mathcal{G}$, and HOLMES $\mathcal{H}$), both before and after applying the polar transform. Specifically, $f_{\text{multiview}}$ is translation equivariant, and $\mathcal{G}$ is translation equivariant up

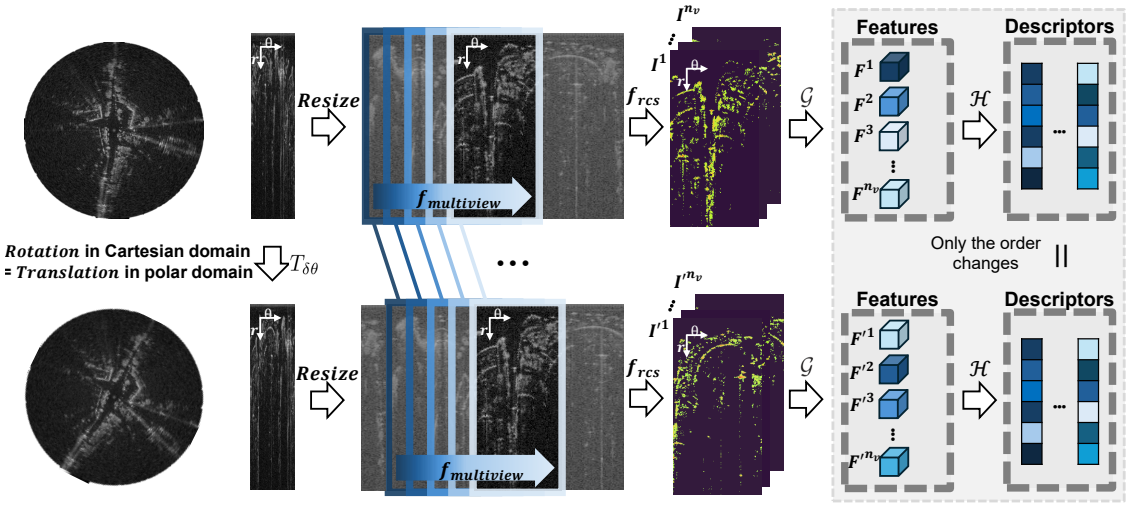

Figure D: Even if the scan rotates, the descriptors just shift in order but come out identically.

Table 5: Properties of Equivariance and Invariance for Each Component. Here, R denotes Rotation, T denotes Translation, Eq denotes Equivariance, and Inv denotes Invariance.

	f_{rcs}	$f_{multiview}$	CNN-based network $\mathcal{G}$	HOLMES $\mathcal{H}$	Overall
Intrinsic property	RT Eq	T Eq	T Eq (up to edge effects)	T Inv	T Inv
With f_{polar} transform*	RT Eq	R Eq	R Eq (up to edge effects)	R Inv	R Inv

*Rotation in the Cartesian domain is identical to translation in the polar domain.

to edge effects. Since rotation in the Cartesian domain is equivalent to translation in the polar domain, both modules can be regarded as rotation equivariant when combined with the polar transform. Finally, due to the invariant properties of HOLMES $\mathcal{H}$ (e.g., summation and concatenation operations), the overall pipeline achieves rotation invariance.

(i) *Rotation.* A rigid rotation in the Cartesian plane corresponds to a *cyclic shift along the azimuth* in polar (r, θ). Our pipeline is designed to be *invariant* to this cyclic shift (hence rotation-invariant). (ii) *Translation.* A Cartesian translation induces a *position-dependent warp* in (r, θ) (not a global shift). We therefore *do not* claim translation invariance/equivariance in the Cartesian sense; instead, we *intentionally keep translation variance* to preserve place discriminability (cf. Sec . III-E).

Let $S : \mathbb{R}^2 \to \mathbb{R}$ be a scan and $\tilde{S}(r, \theta) = S(r \cos\theta, r \sin\theta)$ its polar image. For a rotation by $\delta\theta$,

$$\tilde{S}_{\text{rot}}(r, \theta) = S\left(R_{\delta\theta}^{-1} \begin{bmatrix} r\cos\theta \\ r\sin\theta \end{bmatrix} \right) = \tilde{S}(r, \theta - \delta\theta),$$

so rotation becomes a *cyclic azimuthal shift*. On a discretized grid with W azimuth bins, this is a column permutation $\mathbf{T}_\delta \in \mathbb{R}^{W \times W}$.

Multi-view extraction and CNN feature extraction are equivariant to azimuthal shifts. Denote the polar image by $\mathbf{S} \in \mathbb{R}^{H \times W}$. Our multi-view operator $f_{\text{multiview}}$ extracts windows

$$M^j = \mathbf{S}[:, a_j : (a_j + W_v)] \in \mathbb{R}^{H \times W_v}, \quad j = 1, \ldots, n_v.$$

For a cyclic shift $\mathbf{T}_\delta$, indices are shifted modulo W, hence

$$f_{\text{multiview}}(\mathbf{S}\mathbf{T}_\delta) = \left\{ \mathbf{S}\mathbf{T}_\delta[:, a_j : (a_j + W_v)] \right\}_j = \mathbf{T}_\delta \, f_{\text{multiview}}(\mathbf{S}),$$

i.e., $f_{\text{multiview}}$ is *equivariant* to azimuthal cyclic shifts.

For the CNN $\mathcal{G}$, with weight sharing and *circular padding* along θ, standard convolution is (discrete) translation-equivariant along the azimuthal axis (up to boundary effects) [16]:

$$\mathcal{G}(\mathbf{S}\mathbf{T}_\delta) \;=\; \mathcal{G}(\mathbf{S})\,\mathbf{T}_\delta \quad \text{(azimuthal shift-equivariance)}.$$

In practice, (i) we use circular padding along the azimuthal axis θ, (ii) employ overlapping windows ($\Delta \ll W_v$) with a sufficiently large n_v to mitigate slicing artifacts, and (iii) the spatial resolution becomes progressively lower as the network depth increases. These design choices collectively reduce the impact of border effects, addressing the reviewer's concern.

HOLMES is invariant to cyclic azimuthal shifts. Let $\mathbf{F} \in \mathbb{R}^{N\times C}$ be the spatially flattened feature map. HOLMES computes scores $\mathbf{S} \in \mathbb{R}^{N\times m}$ and an OT coupling $\mathbf{R} \in \mathbb{R}^{N\times m}$ via Sinkhorn, then aggregates

$$V_{j,k} = \sum_{i=1}^{N} R_{i,k}\, F_{i,j}.$$

If the input is azimuthally shifted by a permutation Π, then $\mathbf{F}' = \Pi\mathbf{F}$. Since Sinkhorn is permutation-equivariant [17], $\mathbf{R}' = \Pi\mathbf{R}$. Therefore

$$V'_{j,k} = \sum_i R'_{i,k}\, F'_{i,j} = \sum_i (\Pi R)_{i,k}\,(\Pi F)_{i,j} = \sum_i R_{\pi(i),k}\, F_{\pi(i),j} = \sum_\ell R_{\ell,k}\, F_{\ell,j} = V_{j,k}.$$

Thus $\mathbf{V}' = \mathbf{V}$. GeM pooling and concatenation preserve invariance, so the descriptor is invariant to azimuthal cyclic shifts (i.e., rotation-invariant).

Let a small Cartesian translation be $\Delta\mathbf{s} = (\Delta x, \Delta y)$. For a point $(x, y) = (r\cos\theta, r\sin\theta)$,

$$\Delta r \approx \frac{x\,\Delta x + y\,\Delta y}{r}, \qquad \Delta\theta \approx \frac{x\,\Delta y - y\,\Delta x}{r^2}.$$

This depends on (r, θ), hence is not a global shift. Thus we do not enforce translation invariance/equivariance; instead, we intentionally keep translation variance to preserve location specificity. $\mathcal{G}$ and $f_{\text{multiview}}$ are equivariant to azimuthal cyclic shifts (Cartesian rotation), and HOLMES is invariant to such shifts via OT-based aggregation. Cartesian translations induce warps in polar, so we retain translation variance.

References

[1] K. Harlow, H. Jang, T. D. Barfoot, A. Kim, and C. Heckman, "A new wave in robotics: Survey on recent mmwave radar applications in robotics," *IEEE Transactions on Robotics*, 2024.

[2] K. Burnett, A. P. Schoellig, and T. D. Barfoot, "Do we need to compensate for motion distortion and doppler effects in spinning radar navigation?" *IEEE Robotics and Automation Letters*, vol. 6, no. 2, pp. 771–778, 2021.

[3] H. Jang, M. Jung, and A. Kim, "Raplace: Place recognition for imaging radar using radon transform and mutable threshold," 2023.

[4] M. Gadd and P. Newman, "Open-radvlad: Fast and robust radar place recognition," in *2024 IEEE Radar Conference*, 2024, pp. 1–6.

[5] H. Kim, B. Choi, E. Choi, and Y. Cho, "Referee: Radar-based lightweight and robust localization using feature and free space," 2024.

[6] K. Cai, B. Wang, and C. X. Lu, "Autoplace: Robust place recognition with single-chip automotive radar," 2022, pp. 2222–2228.

[7] S. Lu, G. Zhuo, H. Wang, Q. Zhou, H. Zhou, R. Huang, M. Huang, L. Zheng, and Q. Shu, "Tdfanet: Encoding sequential 4d radar point clouds using trajectory-guided deformable feature aggregation for place recognition," *arXiv preprint arXiv:2504.05103*, 2025.

[8] G. Tolias, R. Sicre, and H. Jégou, "Particular object retrieval with integral max-pooling of cnn activations," *arXiv preprint arXiv:1511.05879*, 2015.

[9] A. Babenko and V. Lempitsky, "Aggregating deep convolutional features for image retrieval," *arXiv preprint arXiv:1510.07493*, 2015.

[10] F. Radenović, G. Tolias, and O. Chum, "Fine-tuning cnn image retrieval with no human annotation," vol. 41, no. 7, pp. 1655–1668, 2018.

[11] R. Arandjelovic, P. Gronat, A. Torii, T. Pajdla, and J. Sivic, "Netvlad: Cnn architecture for weakly supervised place recognition," 2016, pp. 5297–5307.

[12] M. Jung, S. Jung, H. Gil, and A. Kim, "Helios: Heterogeneous lidar place recognition via overlap-based learning and local spherical transformer," *arXiv preprint arXiv:2501.18943*, 2025.

[13] S. Izquierdo and J. Civera, "Optimal transport aggregation for visual place recognition," 2024, pp. 17 658–17 668.

[14] G. Kim, Y. S. Park, Y. Cho, J. Jeong, and A. Kim, "Mulran: Multimodal range dataset for urban place recognition," 2020, pp. 6246–6253.

[15] D. Barnes, M. Gadd, P. Murcutt, P. Newman, and I. Posner, "The oxford radar robotcar dataset: A radar extension to the oxford robotcar dataset," 2020, pp. 6433–6438.

[16] T. Cohen and M. Welling, "Group equivariant convolutional networks," in *International conference on machine learning*. PMLR, 2016, pp. 2990–2999.

[17] M. Cuturi, "Sinkhorn distances: Lightspeed computation of optimal transport," vol. 26, 2013.