This ICCV paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

S^{3}E: Self-Supervised State Estimation for Radar-Inertial System

Shengpeng Wang[1], Yulong Xie[1], Qing Liao[2], Wei Wang[3] *
[1]Huazhong University of Science and Technology, [2]Harbin Institute of Technology, [3]Wuhan University
{wsp666, yulong_xie}@hust.edu.cn, liaoqing@hit.edu.cn, wangw@whu.edu.cn

Abstract

*Millimeter-wave radar for state estimation is gaining significant attention for its affordability and reliability in harsh conditions. Existing localization solutions typically rely on post-processed radar point clouds as landmark points. Nonetheless, the inherent sparsity of radar point clouds, ghost points from multi-path effects, and limited angle resolution in single-chirp radar severely degrade state estimation performance. To address these issues, we propose S^{3}E, a **S**elf-**S**upervised **S**tate **E**stimator that employs more richly informative radar signal spectra to bypass sparse points and fuses complementary inertial information to achieve accurate localization. S^{3}E fully explores the association between exteroceptive radar and proprioceptive inertial sensor to achieve complementary benefits. To deal with limited angle resolution, we introduce a novel cross-fusion technique that enhances spatial structure information by exploiting subtle rotational shift correlations across heterogeneous data. The experimental results demonstrate our method achieves robust and accurate performance without relying on localization ground truth supervision. To the best of our knowledge, this is the first attempt to achieve state estimation by fusing radar spectra and inertial data in a complementary self-supervised manner.*

1. Introduction

State estimation techniques have garnered widespread attention as crucial facilitators in cutting-edge applications, including autonomous driving [37], indoor localization [28], AR/VR [17, 34], Robotics Navigation [16, 33], and more. Currently, optical sensors such as LiDARs [24] and cameras [27] are considered mainstream sensors for state estimation and external perception. Numerous studies based on these sensors yield highly satisfactory outcomes under ideal conditions. However, optical sensors are not very practical in harsh conditions *e.g.* dust, fog, rain, snow, haze. In contrast, the millimeter-wave (mmWave) radar, operating longer wavelengths, affords robust measurements unaffected by minuscule particles [21]. Furthermore, mmWave radar can furnish relative radial velocity information for reflected objects, thereby indicating ego-motion and object mobility details. Leveraging these advantages, mmWave radar emerges as a promising catalyst for state estimation in diverse weather conditions.

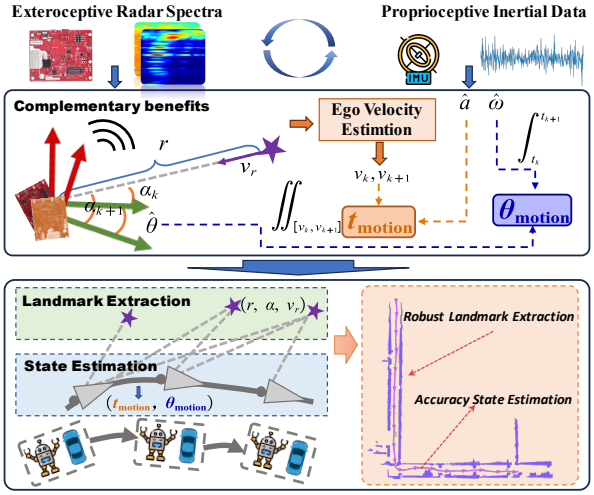


Figure 1. S^{3}E fully explores complementary benefits from *exteroceptive* radar and *proprioceptive* inertial sensor to achieve accurate state estimation.

Existing radar-based studies [2, 41] predominantly employ the Constant False Alarm Rate (CFAR) detector to extract point clouds with Doppler velocities from the processed radar data cube, which encompasses Range-Azimuth-Doppler (RAD) information. On this basis, some studies decouple the relative transformation between two consecutive keyframes through aligning co-observed landmarks [7]. However, multiple objects of different sizes disrupt the independence among CFAR training cells, diminish the probability of detection, and result in missed detections [9]. Additionally, high-intensity cells caused by signal multi-path effects can elevate the chances of false positives, giving rise to "ghost points". Consequently, these sparse and flawed point clouds struggle with reliable data asso-

*Corresponding author

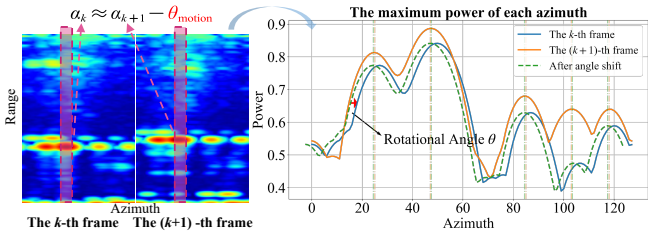


Figure 2. The subtle shifts in the rotational components map into the power's linear translations in the RAS.

ciation, posing a significant challenge for scan-matching techniques. Other studies solve ego velocity from stationary radar points with Doppler velocities [10, 20]. However, the ghost points and dynamic targets without spatial consistency will introduce inaccurate ego velocity estimation, degrading localization performance.

Our motivation is to employ more richly informative Range-Azimuth Spectra (RAS) to bypass sparse points, and leverage the complementary perception capabilities of the Radar-Inertial System (RIS) to attain robust state estimation. Specifically, as shown in Fig. 1, mmWave radar supplies exteroceptive information for IMU to compensate for motion cumulative drift. The proprioceptive IMU provides inertial data in kinetics to distill landmarks with motion consistency. Moreover, we observe that the quantity of principal energy moving between adjacent RAS depends on the rotational component of the motion transformation matrix. In Fig. 2, we take the maximum power along the azimuth for a Pow-Azimuth curve. It is worth noting that linearly translating the k-th curve by the motion angle yields its peak position coinciding exactly with that of the $(k+1)$-th curve. This makes rotation estimation more explicit.

Albeit inspiring, translating this intuition into a practical and reliable state estimator is non-trivial and encounters significant issues: *1) Fusion Incompatibility*. Though RAS make rotation more explicit, unlike point clouds with clear kinematic indicators, radar with inertial data cannot be directly embedded into existing fusion frameworks (*e.g.* Recursive Gaussian Filter, probabilistic factor graph optimization) due to the absence of explicit observation constraint between RAS and state factors. *2) Limited angle resolution*. Commercial single-chip radars are typically equipped with a limited number of antennas, making it difficult to extract reliable landmarks from RAS.

To address the above issues, we present S^{3}E, the first state estimator that fuses radar signal spectrum and inertial data. S^{3}E fully explores the kinematic association between RAS and inertial data to achieve complementary benefits without localization ground truth. As for the limited angle resolution of radar. we propose a novel cross-fusion technique to enhance the spatial structure information by injecting inertial information into RAS. To sum up, our primary contributions are evident in the following aspects.

- We propose a novel self-supervised state estimator, which leverages the complementary perception capabilities of RIS to attain accurate state estimation. To the best of our knowledge, this is the first attempt to achieve state estimation and landmark extraction from radar signal spectra and inertial data in a self-supervised manner.
- We propose a novel Rotation-based cross-fusion to effectively preserve motion-consistent features and enhance the spatial structure across adjacent spectra.
- Experimental results demonstrate that our method outperforms the previous state of the arts and achieves complementary benefits across exteroceptive radar and proprioceptive inertial data.

2. Related Work

Previous research on radar-based state estimation falls into two main categories: model-based and learning-based approaches. Model-based methods can be further categorized into ego-velocity-based and scan-matching-based methods.

2.1. Model-based State Estimation

Many efforts follow the scan matching program from LiDAR SLAM. Cen and Newman [7] design specific descriptors for landmarks to determine scan-matching correspondences. To mitigate outliers' effects on matching, Aldera *et al*. [1] utilize kinetic constraints to reject outliers and Burnett *et al*. [5] devote to compensate for motion distortion and the measurements bias from Doppler effects. Some effects [41, 44] respectively operate scan-to-scan and scan-to-map as radar odometry using recently eye-catching 4D imaging radars. However, these approaches are tailored for radars with high-angle resolution and are incompatible with commercial automotive radar or mobile-integrated single-chip radar. RALL [40] and DC-Loc [13] propose matching schemes for aligning point clouds to pre-built maps, however, pre-built maps are unavailable in most scenarios.

Other studies focus on solving instantaneous ego velocity from single-frame point clouds of Doppler radar. Kellner *et al*. [19] utilize Doppler velocities to resolve complete velocity through RANSAC [12]. However, using only one radar is constrained by a single-track model under Ackermann conditions, often leading to undefined yaw rate observations. Retan *et al*. [31] thoroughly analyze the explicit mapping between radial velocity and target position. Additionally, they construct a radar scan cost term to optimize the pose smoothly within a sliding window. Similarly, Li *et al*. [23] operate a 4D imaging radar with more dense measurements and derive ego-velocity pre-integration factors for pose graph optimization. While these arts achieve satisfactory state estimation performance, they often degrade or fail outright when more dynamic objects intersect the radar's scan viewsight [19]. Moreover, landmarks without

poor motion consistency retain some ghost clutter and dynamic objects, which affects velocity estimation.

2.2. Learning-based State estimation

Radar state estimation by deep learning is currently an emerging area of interest. Most techniques are designed for costly mechanical radars with higher spatial resolutions on par with LiDARs. However, these methods [3, 4, 6] are incompatible with low-resolution automotive radars. In distinction, Lu *et al*. [26] achieves end-to-end supervised pose estimation. Specifically, they convert radar point clouds into a depth map, extracting spatial features by [36] from the depth map and IMU time series, and feeding them into the pose regression network. The paradigm achieves pose estimation in low-capacity RIS. However, due to the sparsity of radar point clouds, it falls short in extracting reliable landmarks, which can further optimize the global pose [8]. Some studies [11, 38] have made notable progress in enhancing their quality and density; however, they fail to exploit the inherent connection between radar and IMU for state estimation. In addition, the IMU dataflow without initial velocity cannot afford an accurate translational component of the relative transformation.

3. Self-Supervised State Estimator for Radar-Inertial System

3.1. System Overview

In this section, the pipeline of our work S^{3}E is sequentially introduced. As shown in Fig. 3, S^{3}E includes five modules.

- **Rotation-based Cross Fusion** integrates motion information from IMU into delicately processed RAS, which preserves motion-consistent features and enhances the spatial structure information between adjacent RAS.
- **Consistent Landmark Extractor** predicts sub-pixel landmark locations by skip connection neural network and performs differentiable data association.
- **Differentiable Velocity Estimation** estimate sub-pixel Doppler velocities for extracted geometry-consistent landmarks and solve instantaneous ego velocity.
- **Self-Supervised Loss Function** configures Velocity Alignment, Geometry Constraint, and Kinematic Constraint to steer the network's learning process.
- **Localization and Mapping** achieves robust state estimation and landmark extraction.

3.2. Rotation-based Cross Fusion

Considering the different sampling rates of IMU and radar, IMU inter-frame pre-integration is used for data frame synchronization [30, 42]:

$$\begin{aligned} \boldsymbol{p}_{k+1}^{k} &= \iint_{t\in[k,k+1]} \mathbf{R}_t^k \left(\hat{\mathbf{a}}_t - \boldsymbol{b}_a\right) \mathrm{d}t^2 \\ \boldsymbol{v}_{k+1}^{k} &= \int_{t\in[k,k+1]} \mathbf{R}_t^k \left(\hat{\mathbf{a}}_t - \boldsymbol{b}_a\right) \mathrm{d}t \\ \boldsymbol{q}_{k+1}^{k} &= \int_{t\in[k,k+1]} \boldsymbol{q}_t^k \otimes \begin{bmatrix} 0 & \frac{1}{2}\left(\hat{\boldsymbol{\omega}}_t - \boldsymbol{b}_\omega\right) \end{bmatrix}^\top \mathrm{d}t \end{aligned} \tag{1}$$

where $\otimes$ is the quaternion multiplication operation, $(\cdot)_{k+1}^{k}$ denotes a measurement in IMU frame when receiving $(k+1)$-th RAS from mmwave radar with respect to the k-th IMU frame. We specify that the vehicle body coordinate system is aligned with the IMU coordinate system in this work. $\mathbf{R}_t^k \in SO(3)$ is an incremental rotation matrix from the current IMU frame to the k-th IMU frame *i.e.* $\boldsymbol{q}_t^k$ in form of quaternion. $\hat{\mathbf{a}}$ and $\hat{\boldsymbol{\omega}}$ refer to the measurements of the vehicle body's linear acceleration and angular velocity, respectively. $\boldsymbol{b}_a, \boldsymbol{b}_\omega$ are measurement biases of linear acceleration and angular velocity, which are regressed by the three-layer fully connected network with the IMU measurements and the instantaneous ego-velocity as inputs. $\boldsymbol{p}_{k+1}^{k}, \boldsymbol{v}_{k+1}^{k}, \boldsymbol{q}_{k+1}^{k}$ represents the increments of body's position, velocity, and orientation from the $(k+1)$-th to k-th frame with the k-th frame as the reference system. For discrete-time implementation, mid-point integration is applied.

By pre-integration in Eqn. (1) aligned with radar measurement, the slight rotation variance ϑ of vehicle motion between the k-th and $(k+1)$-th radar frames is translated by $\boldsymbol{q}_{k+1}^{k}$. Next, we utilize the property that such rotation represents a linear translation between the two Range-Azimuth Spectra along the azimuth axis. The rotation-based cross-fusion between Radar and IMU is shown in Fig. 4.

Firstly, we filter noise with low power along each azimuth in RAS against speckle noise, and saturation noise picked up by the receiver. Let $\boldsymbol{M}_k = \left(\boldsymbol{\alpha}_1^k, \boldsymbol{\alpha}_2^k, \cdots, \boldsymbol{\alpha}_n^k\right) \in \mathbb{R}^{H\times W}$ denote the soft mask matrix, where filtered elements are set to 0, while others are retained in the k-th frame RAS. H, W indicates the range height and azimuth width of the RAS, respectively. Considering non-uniform angular resolution, the angle bins associated with the n column vectors of the soft mask are denoted as $\boldsymbol{\eta} = (\eta_1, \cdots, \eta_n)$. One of our key insights is that the static landmarks maintain a θ orientation at the k-th frame and a $\theta - \vartheta$ orientation at the $(k+1)$-th frame as the radar undergoes rotation by ϑ during a trivial period. Let $\boldsymbol{\beta}_i^k$ denote the expected vector that $\boldsymbol{\alpha}_i^k$ is transformed into from the k-th to $(k+1)$-th frame. Expected vectors are fished by designing the dot-product attention mechanism with the angular difference as metrics:

$$\boldsymbol{\beta}_m^k = \mathrm{Softmax}\left(-\frac{\left(\boldsymbol{\eta}^T - \vartheta - \eta_m\right)^2}{\kappa}\right) \cdot \left(\boldsymbol{\alpha}_1^k, \boldsymbol{\alpha}_2^k, \cdots, \boldsymbol{\alpha}_n^k\right) \tag{2}$$

where κ indicates the temperature parameter. The expected matrix $\boldsymbol{M}_{k+1}^k = \varrho(\boldsymbol{\beta}_1^k, , \boldsymbol{\beta}_2^k, \cdots, \boldsymbol{\beta}_n^k)$ are attached to the original matrix $\boldsymbol{M}_{k+1}$ at scale ϱ to enhance the landmark

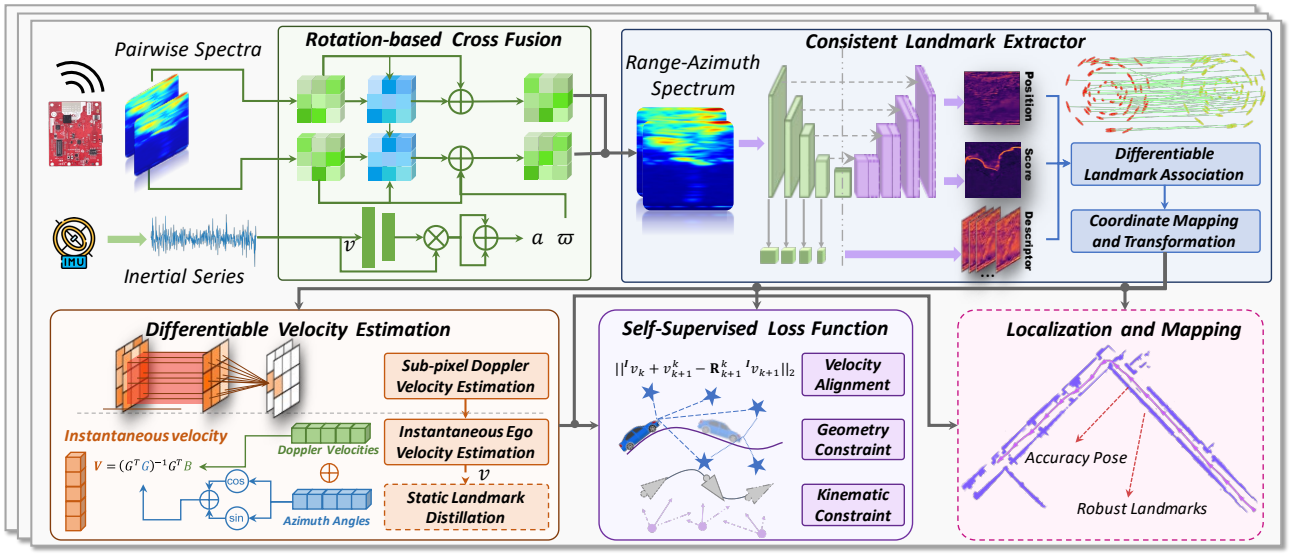


Figure 3. Overview of S^{3}E. Given a pair RAS and inertial data, it outputs accurate poses and eometry-consistent landmarks.

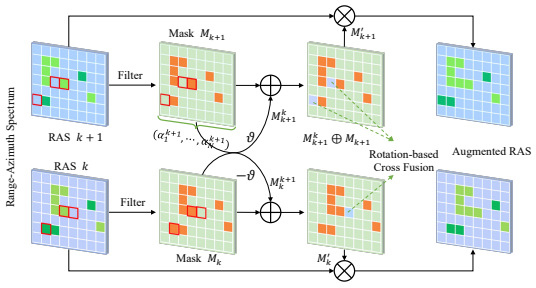


Figure 4. Schematic of Rotation-based Cross Fusion.

spatial features of the k-th RAS as:

$$\begin{aligned} \boldsymbol{M}_{k+1}^{'} &= \boldsymbol{M}_{k+1}^{k} \oplus \boldsymbol{M}_{k+1} \\ \boldsymbol{M}_{k+1}^{k} &= \text{Softmax}\left(-\frac{\left(\mathbf{1}\boldsymbol{\eta}^T - \vartheta\mathbf{1}\mathbf{1}^T - \boldsymbol{\eta}\mathbf{1}^T\right)^2}{\kappa}\right) \cdot \boldsymbol{M}_k \end{aligned} \tag{3}$$

where $\mathbf{1}$ denotes a column vector with elements all 1. Similarly, $\boldsymbol{M}_k$ can be augmented with $\boldsymbol{M}_k^{'} = \boldsymbol{M}_k^{k+1} \oplus \boldsymbol{M}_k$ where $\boldsymbol{M}_k^{k+1} = \text{Softmax}\left(-\frac{\left(\mathbf{1}\boldsymbol{\eta}^T + \vartheta\mathbf{1}\mathbf{1}^T - \boldsymbol{\eta}\mathbf{1}^T\right)^2}{\kappa}\right)$. Such cross-fusion mapping rotation information into RAS is applied to enhance motion-consistent spatial features.

3.3. Consistent Landmark Extractor

After Rotation-based Cross Fusion, RAS with enhanced geometry-consistent spatial features are employed to extract and match spatial feature points, serving as consistent landmarks across adjacent temporal frames. To this end, we utilize the impressive drawing power of U-Net [32] for the underlying location information with deep semantic information. Specifically, we feed the enhanced RAS into U-Net as inputs and design a multi-head architecture to decode locations of feature points and acquire N landmarks' locations. Assuming radar measurement uncertainty is isotropic, we decode an additional spatial weight matrix of dimensions $H \times W$ to represent scores as landmarks. Additionally, we reshape output of the encoder backbone blocks to a feature map with 248 channels before concatenating these blocks. Subsequently, we treat such feature maps with $N \times 248$ dimension as descriptors of N landmarks, which are used to match landmarks from different temporal frames by differentiable landmark association. Next, we specifically illustrate three headers: location, score, and descriptor.

The location header outputs detection scores $L \in \mathbb{R}^{H \times W}$ with which each pixel is selected as a candidate landmark. To fully exploit the signal energy sidelobe information in RAS, we extract sub-pixel location (u_k, v_k) within a patch $\mathcal{U}_k$ centered on each pixel of M_k as:

$$\begin{aligned} u_k &= \sum_{(i,j)\in\mathcal{U}_k} u_{ij} \left[\text{Softmax}\left(L_{\mathcal{U}_k}\right)\right]_{ij} \\ v_k &= \sum_{(i,j)\in\mathcal{U}_k} v_{ij} \left[\text{Softmax}\left(L_{\mathcal{U}_k}\right)\right]_{ij} \end{aligned} \tag{4}$$

The score header exports the normalized credibility weight c_i^k that represents the score of each point as a landmark. This helps eliminate the effects of unpairable ghost points with highly reflective power.

The descriptor header records and concatenates encoder backbone blocks to identify the sub-pixel of the candidate landmark. We use the descriptors to match the candidate

landmarks extracted from the previous frame into the pixel features in the current frame. Let $Q_{k-1}^{N} \in \mathbb{R}^{N\times 248}, K_k^{H\times W}$ with $(H \times W, 248)$ dimension respectively denote the N candidate landmarks' descriptors in the $(k-1)$-th frame and the pixel features' descriptors in the k-th frame. Then, the candidate landmarks' pixel location V' in the k-th frame can be distilled through the attention mechanism [35]:

$$V_k' = \text{Softmax}\left(\frac{Q_{k-1}^{N} \cdot \left(K_k^{H\times W}\right)^T}{\kappa}\right) V \tag{5}$$

where V is a pixel location matrix of dimension $(H \times W, 2)$ by flattening all pixels (u_i, v_j) in RAS. We can transform the pixel locations V_k' to landmark locations $\boldsymbol{d}_i (i = 1, \cdots, N)$ in the k-th radar coordinate system. With $\boldsymbol{d}_i$, we design a differentiable velocity estimation module to solve vehicle instantaneous velocity and further eliminate clutter and dynamic feature points.

3.4. Differentiable Velocity Estimation

In this section, we probe the couple between the Doppler velocities of the landmarks and the vehicle's instantaneous velocity, and solve it in a differentiable way. Specifically, the proposed differentiable velocity estimation consists of the following three sub-modules.

Sub-pixel Doppler Velocity Estimation: we use the network with concatenation skip connections in Sec. 3.3 to extract sub-pixels landmark locations. To obtain the corresponding Doppler velocities, we extract the local landmark Doppler velocities from the global pixel Doppler map according to pixel coordinate distance:

$$v^r = \sum_{(i,j)\in M_k} v_{ij}^r \left[\text{Softmax}\left(-\frac{\Delta_k^2}{\kappa}\right)\right]_{ij} \tag{6}$$

where v_{ij}^r, v^r, Δ_k respectively refers to Doppler velocity corresponding to pixel (u_i, v_j), sub-pixel Doppler velocities of landmarks, and differences between targets' sub-pixels and the global pixels coordinate in the k-th frame.

Instantaneous Ego Velocity Estimation: we set the k-th IMU frame $\mathcal{F}_{\boldsymbol{I}_k}$ as the body frame, which is aligned with the ego vehicle frame. The instantaneous velocity ${}^{\boldsymbol{R}}\boldsymbol{v}_k$ of the vehicle in the k-th radar frame $\mathcal{F}_{\boldsymbol{R}_k}$ can be expressed in terms of its velocity ${}^{\boldsymbol{I}}\boldsymbol{v}_k$ in the body frame:

$${}^{\boldsymbol{R}}\boldsymbol{v}_k = \left(\mathbf{R}_{\boldsymbol{R}}^{\boldsymbol{I}}\right)^T \left({}^{\boldsymbol{I}}\boldsymbol{v}_k + {}^{\boldsymbol{I}}\boldsymbol{\omega}_{\boldsymbol{k}} \times {}^{\boldsymbol{I}}\boldsymbol{l}_{\boldsymbol{R}}\right) \tag{7}$$

where $\mathbf{R}_{\boldsymbol{R}}^{\boldsymbol{I}}$ means the relative rotation from the radar frame to the IMU frame, ${}^{\boldsymbol{I}}\boldsymbol{\omega}_k$ refers to vehicle angular velocity, and ${}^{\boldsymbol{I}}\boldsymbol{l}_{\boldsymbol{R}}$ donates the radar mounting position relative to the IMU. Next, as shown in Fig. 5, we observe a kinematic constraint between the vehicle velocity in the body frame and relative radial velocities *i.e.* Doppler velocities of the

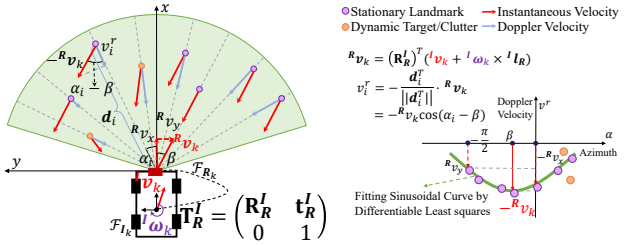


Figure 5. Schematic diagram of estimation for the vehicle's instantaneous velocity in a differentiable manner. Left reveals a kinematic constraint between the vehicle velocity ${}^{\boldsymbol{I}}\boldsymbol{v}_k$, azimuth angles and relative radial velocities of stationary landmarks. The right solves the instantaneous velocity and filters dynamic targets and clutters by differentiable least squares.

stationary landmarks. We combine all cosine constraints from radial velocity observations to yield a super-definite equation about the vehicle velocity:

$$-\begin{bmatrix} v_1^r \\ \vdots \\ v_N^r \end{bmatrix} = \begin{bmatrix} \cos\alpha_1 & \sin\alpha_1 \\ \vdots & \vdots \\ \cos\alpha_N & \sin\alpha_N \end{bmatrix} \begin{bmatrix} {}^{\boldsymbol{R}}\boldsymbol{v}\cos\beta \\ {}^{\boldsymbol{R}}\boldsymbol{v}\sin\beta \end{bmatrix} \tag{8}$$

where α_i, β respectively indicate the direction of landmark i and ego vehicle' velocity direction in the body frame. Eqn. (8) can be written as $\boldsymbol{B} = \boldsymbol{G}\boldsymbol{X}$, and in consequence, the velocity can be solved in a differential way as $(\boldsymbol{G}^T\boldsymbol{G})^{-1}\boldsymbol{G}^T\boldsymbol{B}$ by least squares.

Static Landmark Criteria and Distillation: We adopt Eqn. (8) as a criterion for stationary landmarks and further filter dynamic landmarks and clutter as well as mismatched pairs by RANSAC. Notably, this process is cooled down during the training phase and activated until the test phase.

3.5. Self-Supervised Loss Function

Constructing geometric and kinematic constraints between the proprioceptive IMU and exteroceptive radar is crucial for enabling self-supervision in RAS. A straightforward approach is to respectively decouple the inter-frame transformation matrix $\begin{bmatrix}\mathbf{R}_k^{k+1} & \mathbf{t}_k^{k+1}; 0 & 1\end{bmatrix}$ from the radar and the IMU as constraints. Unfortunately, the IMU inter-frame integrals are missing initial velocities, making the translational component of the transformation matrix unimpressive. In this case, we novelly propose to exploit the body velocity implicit in the radar Doppler information to compensate for the lack of initial IMU velocities. This allows us to construct constraints using the landmark points. Specifically, we design joint loss function considering the geometric constraint between IMU transformation and landmark points, the kinematic constraint applied to the Doppler measurements, and the velocity disparity between multiple sensors as follows:

- *Geometry Constraint*: Each landmark pair (p_k^i, q_{k+1}^i), $i = 1, \cdots, N$ from the k-th and $k+1$-th frames follows the IMU transformation matrix as $q_{k+1}^i = \mathbf{R}_k^{k+1} p_k^i + \mathbf{t}_k^{k+1}$. Hence, the geometry loss can be listed by accumulating residuals in the Mahalanobis Distance [25]:

$$\mathcal{L}_1 = \frac{1}{2} \sum_{k \in \mathcal{B}} \sum_{i=1}^{N} c_i^k e_i^T e_i \tag{9}$$

where $\mathcal{B}$ denotes the batch number and $e_i = \mathbf{R}_k^{k+1} p_k^i + \mathbf{t}_k^{k+1} - q_{k+1}^i$ refers to the residual. The unimpressive translational component $\mathbf{t}_k^{k+1} = \int_{t \in [k,k+1]} \mathbf{R}_t^k \cdot {}^{\boldsymbol{I}}\boldsymbol{v}_k \mathrm{d}t + \boldsymbol{p}_{k+1}^k$ is patched by Eqn. (8)(1).
- *Kinematic Constraint*: All distilled static landmark points satisfy the cosine constraint from radial velocity observations in Eqn. (8), which can be described as:

$$\mathcal{L}_2 = ||\boldsymbol{G}(\boldsymbol{G}^T\boldsymbol{G})^{-1}\boldsymbol{G}^T\boldsymbol{B} - \boldsymbol{B}||_2 \tag{10}$$

- *Velocity Alignment*: The body velocity in $(k+1)$-th frame at the k-th coordinate system can be expressed by IMU and radar observations or direct velocity transformation. Align these expressions to get velocity loss:

$$\mathcal{L}_3 = ||{}^{\boldsymbol{I}}\boldsymbol{v}_k + \boldsymbol{v}_{k+1}^k - \mathbf{R}_{k+1}^k \cdot {}^{\boldsymbol{I}}\boldsymbol{v}_{k+1}||_2 \tag{11}$$

Our total loss function joins the above losses as:

$$\mathcal{L}_{total} = \mathcal{L}_1 + \lambda_1 \mathcal{L}_2 + \lambda_2 \mathcal{L}_3 \tag{12}$$

Finally, in the Localization and Mapping module, Static Landmark Criteria and Distillation in Sec. 3.4 is activated, and the landmark points are selected according to stationary metric in Eqn. (8). Subsequently, all static landmarks are aligned to the initial coordinate frame and integrated into a spatially coherent map by relative pose transformations.

4. EXPERIMENTS AND EVALUATION

4.1. Experimental Setup

ColoRadar Dataset. This work focuses on extracting geometry-consistent landmark points from Range-Azimuth Spectrum, in conjunction with IMU data streams. We conduct our method on ColorRadar Dataset [22], which contains unprocessed analog-to-digital convert sample data and provides access to RAS through signal processing. The ColoRadar contains both indoor and outdoor scenes with a total of 52 sequences and a distance of 107.4km. It furnishes LiDAR, single-chip radar (TI AWR1843BOOST-EVM), and IMU data. To validate our method's superiority, we utilize lower-resolution single-chip radar data alongside IMU data for our experiments. For a fair comparison with other learning-based baselines, we select 36 sequences as the training set and 16 for testing.

Self-collected Dataset We built a multimodal data acquisition platform (Figure 6) and collected data in 12 indoor and outdoor scenes. The localization ground truth was obtained by using Fast-Livo [43]. To demonstrate the generalization of S3E as a self-supervised method over supervised learning [26], we evaluate various methods in such unseen scenes. The dataset will be available with code at https://github.com/wsp666/S3E.

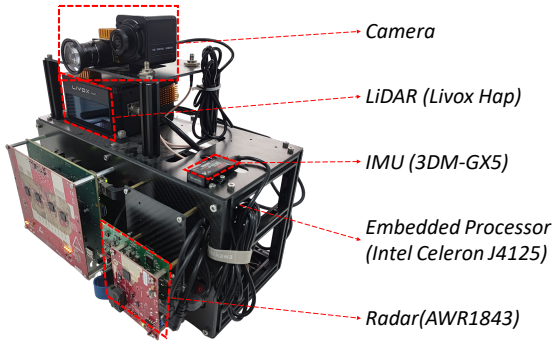


Figure 6. Our multi-modal sensing platform.

Implementation Details. We implement our system S^{3}E using Pytorch 1.11.0 with CUDA 12.4. The extractor network is fine-tuned using the pre-trained model from RadarHD [29] to make S^{3}E converge better. The parameters λ_1 and λ_2 of the weighted kinematic constraint and velocity alignment loss function are respectively set to 0.05 and 0.1. The temperature coefficient κ and scale parameter ϱ set to 0.01, and 0.3, respectively. We train the networks for 120 epochs on ColoRadar Dataset with Adam optimizer, which has a learning rate of 10^{-4} and controls it with the ReduceLROnPlateau scheduler. Specifically, The pixel values within the first six range bins of RAS are set to zero to mitigate effects of fixed installations within the radar's FOV. Moreover, to enhance spatial feature density, we interpolate each RAS from 128×128 to 256×256 before filtering speckle noise below 0.8 of the mean value per azimuth. We train the model with a machine with one NVIDIA GeForce RTX 4090 and Intel Xeon Gold 6226R CPU.

4.2. Overall Performance

We compare our method against model-based and learning-based localization approaches for the Radar-Inertial system in terms of translation and rotation error. For model-based methods, we utilize EKF-RIO [10], PG-RIO [41] and 4D-RIO [39] as our baseline. As for learning-based methods, we adopt Milliego [26] as another baseline. In addition to absolute pose error, we portray the relative error distribution for a more intuitive evaluation of odometry.

Relative pose error distributions are analyzed to evaluate our odometry results against the baselines as illustrated in Fig. 7. The scatterplots show our method reduces rela-

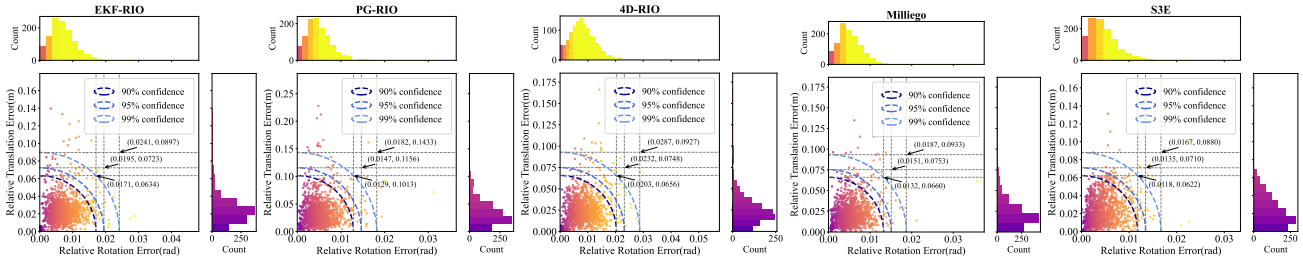


Figure 7. The relative rotation and translation error distributions, and quarter ellipse contours at 90%, 95%, and 99% confidence levels for different methods.

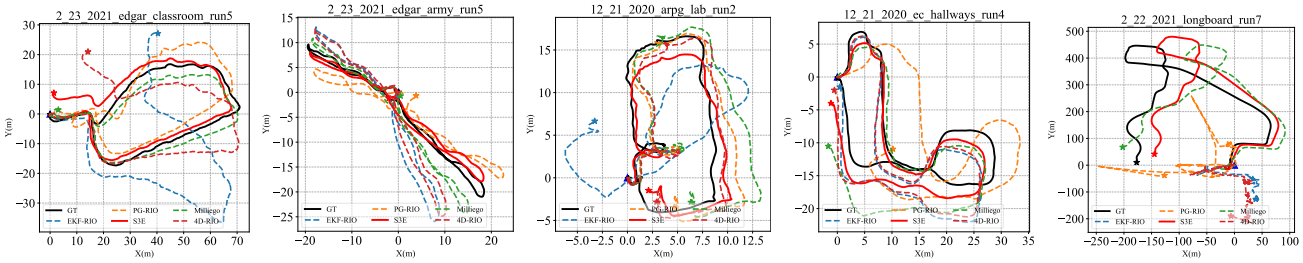


Figure 8. Pose estimation results are shown from different methods across various scenarios, where △ indicates a unified initial point and ⋆ signifies the termination point of each respective algorithm.

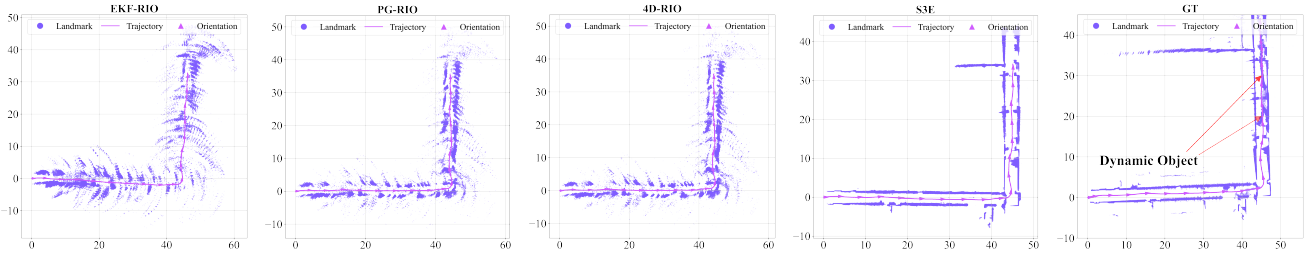


Figure 9. Pose and map visualizations from various methods and ground truth in self-collected dataset. All poses indicate transformations relative to the initial coordinate system. Note that Milliego is unable to extract landmark point clouds.

tive rotation and translation errors by 11% and 5.7%, compared to Milliego, with a 90% confidence level. This indicates more effective localization bias reduction and more reliable relative pose estimation. For an overall trajectory as Fig. 8 shows, our system significantly contributes to a more accurate, less drifting trajectory. In addition, we evaluate the absolute pose error of different methods in various scenarios using the KITTI toolkit [14] in Tab. 1. S^{3}E still maintains superior performance in longboard scenarios where the CFAR detector generates fairly sparse point clouds. In contrast, EKF-RIO, PG-RIO and 4D-RIO fail due to inaccurate velocity estimation and poor point-wise matching. Among the scenes with valid evaluations, S^{3}E achieves the lowest average pose error, with a translation error of 3.10% and a rotation error of 4.05 deg/100m, which demonstrates significant enhancement of 16.8% in rotation and 22.2% over supervised Milliego. This is attributed to more richly informative RAS rather than highly clustered points. Besides, S^{3}E in a self-supervised manner can be seamlessly adapted to new environments as shown in Tab. 2. Our approach consistently maintains robust and accurate localization performance.

Furthermore, in contrast to Milliego, S^{3}E can additionally extract motion-consistent landmark features by leveraging the spatial structure information within RAS. Qualitatively, as shown in Fig. 9, our approach yields a more structured map with denser, more consistent landmark point clouds, and substantially fewer ghost points. Quantitatively, the performance of landmark extractors is evaluated using four metrics: Chamfer distance [15], Hausdorff distance [18], RPCDL [9] and Clutter Radio, as shown in Fig. 10. The Chamfer and Hausdorff distances quantify the average distance between each point in two sets and its nearest neighbor in the opposite set. A higher RPCDL means the landmark point clouds are closer to the ground truth provided by LiDAR. Besides, a landmark point is defined as

Methods	longboard		edgar_classroom		outdoors		ec_hallways		edgar_army		arpg_lab		aspen_run	
	Trans.↓	Rot.↓	Trans.↓	Rot.↓	Trans.↓	Rot.↓	Trans.↓	Rot.↓	Trans.↓	Rot.↓	Trans.↓	Rot.↓	Trans.↓	Rot.↓
EKF-RIO [10]	-	-	5.32	6.08	4.65	5.61	5.68	**2.16**	6.19	5.38	6.90	11.3	9.73	19.1
PG-RIO [41]	-	-	2.61	2.99	7.67	14.7	5.58	5.26	3.94	3.99	**2.31**	9.49	2.74	15.8
4D-RIO [39]	-	-	5.62	2.56	8.37	4.69	5.91	4.33	4.97	3.01	2.55	12.1	2.48	18.2
Milliego [26]	9.14	2.23	**2.32**	**1.91**	**2.02**	5.92	4.46	5.01	**2.13**	2.45	2.84	7.89	3.17	**11.0**
S^3E	**5.69**	**2.07**	2.45	2.21	2.16	**3.83**	**3.90**	2.53	2.33	**2.11**	2.82	**4.28**	**2.34**	11.3

Table 1. Comparison of rotation and translation error on the Coloradar Dataset. The rotation error (abbreviated as Rot.) is measured in deg/100m and translation *i.e.* Trans. in %.

Methods	indoor_seq0		indoor_seq1		indoor_seq2		indoor_seq3		outdoor_seq0		outdoor_seq1		outdoor_seq2	
	Trans.↓	Rot.↓	Trans.↓	Rot.↓	Trans.↓	Rot.↓	Trans.↓	Rot.↓	Trans.↓	Rot.↓	Trans.↓	Rot.↓	Trans.↓	Rot.↓
EKF-RIO [10]	4.35	5.16	3.66	4.91	4.12	3.02	5.01	3.14	-	-	-	-	-	-
PG-RIO [41]	3.68	4.10	5.27	3.65	4.32	3.71	4.25	2.96	-	-	-	-	-	-
4D-RIO [39]	5.65	3.23	5.34	3.17	4.16	2.79	4.62	2.77	-	-	-	-	-	-
Milliego [26]	9.84	3.16	12.4	3.97	8.95	2.92	9.51	3.01	14.3	3.62	16.9	3.76	18.6	3.55
S^3E	**2.69**	**2.13**	**2.97**	**2.41**	**2.35**	**1.98**	**2.84**	**2.15**	**5.37**	**3.11**	**6.53**	**3.55**	**6.92**	**3.14**

Table 2. Evaluation on the fresh scenes, where the dataset is collected by our multi-modal sensing platform.

a clutter point if its nearest neighbor's distance from the ground truth exceeds 0.65m, and the Clutter Ratio indicates the fraction of landmark points identified as clutter.

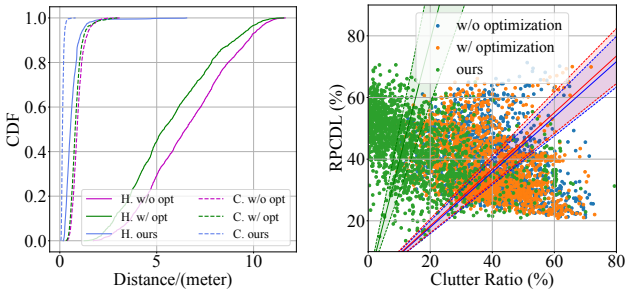


Figure 10. Quality assessment of landmark extractors.

As shown in Fig. 10, we evaluate our landmark points against raw radar points used in EKF-RIO and optimized points by PG-RIO and 4D-RIO. The left shows that S^3E achieves the Chamfer and Hausdorff distances with the median errors of 0.14m and 0.55m, respectively, which reduce to 10-17% compared to baseline. The right illustrates the distribution of RPCDL and Clutter Ratio at different frames. Specifically, although PG-RIO filters and optimizes landmark points better than EKF-RIO, it still exhibits a higher Clutter Ratio and lower RPCDL compared to our method. Therefore, S^3E provides more geometry-consistent spatial landmark points with more accurate and robust perception than CFAR-based methods.

4.3. Ablation Study

To assess the individual contribution of the Rotation-based Cross Fusion module based on our insight in Fig. 1, we conduct an ablation experiment, presenting median errors for Chamfer Distance (C.D.), Hausdorff Distance (H.D.), and pose errors in translation and rotation, as shown in Tab. 3. The results indicate an 11-19% improvement in state estimation and landmark extraction, demonstrating that Rotation-based Cross Fusion effectively preserves motion-consistent features and enhances the spatial structure between adjacent spectra.

S^3E	RCF†	C.D(m)↓	H.D(m)↓	Trans.(%)↓	Rot.(deg/100m)↓
✓	✗	0.16	0.62	3.23	2.57
✓	✓	**0.14**	**0.55**	**2.84**	**2.15**

† Rotation-based Cross Fusion in Fig. 4.

Table 3. ABLATION STUDY RESULTS

In a nutshell, our method S^3E thoroughly leverages geometric spatial consistency constraints while enhancing state estimation and landmark feature extraction in a complementary self-supervised manner, and two heterogeneous sensors synergize and complement each other, creating a harmonious interplay that enhances both.

5. Conclusion

This paper introduces S^3E, a novel self-supervised state estimator and landmark extractor for Radar-Inertial Systems that leverages two key observations: subtle shifts in rotational components translate into linear shifts in the Range-Azimuth spectrum, and Doppler velocities can compensate for initial velocity, preventing cumulative IMU offsets. The experiments demonstrate that S^3E harnesses the complementary advantages of two heterogeneous sensors, achieving robust and accurate performance.

Acknowledgment

This work was supported in part by National Natural Science Foundation of China with Grant 62471194.

References

[1] Roberto Aldera, Matthew Gadd, Daniele De Martini, and Paul Newman. What goes around: Leveraging a constant-curvature motion constraint in radar odometry. *IEEE Robotics and Automation Letters*, 7(3):7865–7872, 2022. 2

[2] Yasin Almalioglu, Mehmet Turan, Chris Xiaoxuan Lu, Niki Trigoni, and Andrew Markham. Milli-rio: Ego-motion estimation with low-cost millimetre-wave radar. *IEEE Sensors Journal*, 21(3):3314–3323, 2020. 1

[3] Yasin Almalioglu, Mehmet Turan, Niki Trigoni, and Andrew Markham. Deep learning-based robust positioning for all-weather autonomous driving. *Nature Machine Intelligence*, 4(9):749–760, 2022. 3

[4] Dan Barnes and Ingmar Posner. Under the radar: Learning to predict robust keypoints for odometry estimation and metric localisation in radar. In *2020 IEEE international conference on robotics and automation (ICRA)*, pages 9484–9490. IEEE, 2020. 3

[5] Keenan Burnett, Angela P Schoellig, and Timothy D Barfoot. Do we need to compensate for motion distortion and doppler effects in spinning radar navigation? *IEEE Robotics and Automation Letters*, 6(2):771–778, 2021. 2

[6] Keenan Burnett, David J Yoon, Angela P Schoellig, and Timothy D Barfoot. Radar odometry combining probabilistic estimation and unsupervised feature learning. In *Robotics: Science and Systems*, 2021. 3

[7] Sarah H Cen and Paul Newman. Radar-only ego-motion estimation in difficult settings via graph matching. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 298–304. IEEE, 2019. 1, 2

[8] Changhao Chen, Bing Wang, Chris Xiaoxuan Lu, Niki Trigoni, and Andrew Markham. A survey on deep learning for localization and mapping: Towards the age of spatial machine intelligence. *arXiv preprint arXiv:2006.12567*, 2020. 3

[9] Yuwei Cheng, Jingran Su, Mengxin Jiang, and Yimin Liu. A novel radar point cloud generation method for robot environment perception. *IEEE Transactions on Robotics*, 38(6): 3754–3773, 2022. 1, 7

[10] Christopher Doer and Gert F Trommer. An ekf based approach to radar inertial odometry. In *2020 IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI)*, pages 152–159. IEEE, 2020. 2, 6, 8

[11] Cong Fan, Shengkai Zhang, Kezhong Liu, Shuai Wang, Zheng Yang, and Wei Wang. Enhancing mmwave radar point cloud via visual-inertial supervision. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 9010–9017. IEEE, 2024. 3

[12] Martin A Fischler and Robert C Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981. 2

[13] Pengen Gao, Shengkai Zhang, Wei Wang, and Chris Xiaoxuan Lu. Dc-loc: Accurate automotive radar based metric localization with explicit doppler compensation. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 4128–4134. IEEE, 2022. 2

[14] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012. 7

[15] Yulan Guo, Hanyun Wang, Qingyong Hu, Hao Liu, Li Liu, and Mohammed Bennamoun. Deep learning for 3d point clouds: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 43(12):4338–4364, 2020. 7

[16] Yicong Hong, Yang Zhou, Ruiyi Zhang, Franck Dernoncourt, Trung Bui, Stephen Gould, and Hao Tan. Learning navigational visual representations with semantic map supervision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3055–3067, 2023. 1

[17] Junwen Huang, Hao Yu, Kuan-Ting Yu, Nassir Navab, Slobodan Ilic, and Benjamin Busam. Matchu: Matching unseen objects for 6d pose estimation from rgb-d images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10095–10105, 2024. 1

[18] D.P. Huttenlocher, G.A. Klanderman, and W.J. Rucklidge. Comparing images using the hausdorff distance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(9):850–863, 1993. 7

[19] Dominik Kellner, Michael Barjenbruch, Jens Klappstein, Jürgen Dickmann, and Klaus Dietmayer. Instantaneous ego-motion estimation using doppler radar. In *16th International IEEE Conference on Intelligent Transportation Systems (ITSC 2013)*, pages 869–874. IEEE, 2013. 2

[20] Dominik Kellner, Michael Barjenbruch, Jens Klappstein, Jürgen Dickmann, and Klaus Dietmayer. Instantaneous ego-motion estimation using multiple doppler radars. In *2014 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1592–1597. IEEE, 2014. 2

[21] Youngseok Kim, Sanmin Kim, Jun Won Choi, and Dongsuk Kum. Craft: Camera-radar 3d object detection with spatio-contextual fusion transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1160–1168, 2023. 1

[22] Andrew Kramer, Kyle Harlow, Christopher Williams, and Christoffer Heckman. Coloradar: The direct 3d millimeter wave radar dataset. *The International Journal of Robotics Research*, 41(4):351–360, 2022. 6

[23] Xingyi Li, Han Zhang, and Weidong Chen. 4d radar-based pose graph slam with ego-velocity pre-integration factor. *IEEE Robotics and Automation Letters*, 2023. 2

[24] Jiuming Liu, Guangming Wang, Chaokang Jiang, Zhe Liu, and Hesheng Wang. Translo: A window-based masked point transformer framework for large-scale lidar odometry. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1683–1691, 2023. 1

[25] Ze Liu, Yingfeng Cai, Hai Wang, Long Chen, Hongbo Gao, Yunyi Jia, and Yicheng Li. Robust target recognition and

tracking of self-driving cars with radar and camera information fusion under severe weather conditions. *IEEE Transactions on Intelligent Transportation Systems*, 23(7):6640–6653, 2021. 6

[26] Chris Xiaoxuan Lu, Muhamad Risqi U Saputra, Peijun Zhao, Yasin Almalioglu, Pedro PB De Gusmao, Changhao Chen, Ke Sun, Niki Trigoni, and Andrew Markham. milliego: single-chip mmwave radar aided egomotion estimation via deep sensor fusion. In *Proceedings of the 18th Conference on Embedded Networked Sensor Systems*, pages 109–122, 2020. 3, 6, 8

[27] Yao Lu, Xiaoli Xu, Mingyu Ding, Zhiwu Lu, and Tao Xiang. A global occlusion-aware approach to self-supervised monocular visual odometry. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2260–2268, 2021. 1

[28] Sinno Jialin Pan, Vincent Wenchen Zheng, Qiang Yang, and Derek Hao Hu. Transfer learning for wifi-based indoor localization. In *Association for the advancement of artificial intelligence (AAAI) workshop*. The Association for the Advancement of Artificial Intelligence Palo Alto, 2008. 1

[29] Akarsh Prabhakara, Tao Jin, Arnav Das, Gantavya Bhatt, Lilly Kumari, Elahe Soltanaghai, Jeff Bilmes, Swarun Kumar, and Anthony Rowe. Radarhd: Demonstrating lidar-like point clouds from mmwave radar. In *Proceedings of the 29th Annual International Conference on Mobile Computing and Networking*, pages 1–3, 2023. 6

[30] Tong Qin, Peiliang Li, and Shaojie Shen. Vins-mono: A robust and versatile monocular visual-inertial state estimator. *IEEE Transactions on Robotics*, 34(4):1004–1020, 2018. 3

[31] Kyle Retan, Frasher Loshaj, and Michael Heizmann. Radar odometry on $se(3)$ with constant velocity motion prior. *IEEE Robotics and Automation Letters*, 6(4):6386–6393, 2021. 2

[32] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer, 2015. 4

[33] Erik Sandström, Yue Li, Luc Van Gool, and Martin R Oswald. Point-slam: Dense neural point cloud-based slam. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 18433–18444, 2023. 1

[34] Xiaoyu Shi, Zhaoyang Huang, Weikang Bian, Dasong Li, Manyuan Zhang, Ka Chun Cheung, Simon See, Hongwei Qin, Jifeng Dai, and Hongsheng Li. Videoflow: Exploiting temporal cues for multi-frame optical flow estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12469–12480, 2023. 1

[35] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 5

[36] Sen Wang, Ronald Clark, Hongkai Wen, and Niki Trigoni. Deepvo: Towards end-to-end visual odometry with deep recurrent convolutional neural networks. In *2017 IEEE international conference on robotics and automation (ICRA)*, pages 2043–2050. IEEE, 2017. 3

[37] Sijie Wang, Qiyu Kang, Rui She, Wee Peng Tay, Andreas Hartmannsgruber, and Diego Navarro Navarro. Robustloc: Robust camera pose regression in challenging driving environments. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 6209–6216, 2023. 1

[38] Shengpeng Wang, Xin Luo, Yulong Xie, and Wei Wang. Sddiff: Boost radar perception via spatial-doppler diffusion. *arXiv preprint arXiv:2506.16936*, 2025. 3

[39] Yang Xu, Qiucan Huang, Shaojie Shen, and Huan Yin. Incorporating point uncertainty in radar slam. *IEEE Robotics and Automation Letters*, 2025. 6, 8

[40] Huan Yin, Runjian Chen, Yue Wang, and Rong Xiong. Rall: end-to-end radar localization on lidar map using differentiable measurement model. *IEEE Transactions on Intelligent Transportation Systems*, 23(7):6737–6750, 2021. 2

[41] Jun Zhang, Huayang Zhuge, Zhenyu Wu, Guohao Peng, Mingxing Wen, Yiyao Liu, and Danwei Wang. 4dradarslam: A 4d imaging radar slam system for large-scale environments based on pose graph optimization. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 8333–8340. IEEE, 2023. 1, 2, 6, 8

[42] Shengkai Zhang, Sheyang Tang, Wei Wang, Tao Jiang, and Qian Zhang. Conquering textureless with rf-referenced monocular vision for mav state estimation. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 146–152. IEEE, 2021. 3

[43] Chunran Zheng, Wei Xu, Zuhao Zou, Tong Hua, Chongjian Yuan, Dongjiao He, Bingyang Zhou, Zheng Liu, Jiarong Lin, Fangcheng Zhu, Yunfan Ren, Rong Wang, Fanle Meng, and Fu Zhang. Fast-livo2: Fast, direct lidar–inertial–visual odometry. *IEEE Transactions on Robotics*, 41:326–346, 2025. 6

[44] Yuan Zhuang, Binliang Wang, Jianzhu Huai, and Miao Li. 4d iriom: 4d imaging radar inertial odometry and mapping. *IEEE Robotics and Automation Letters*, 2023. 2