Self-Supervised Doppler-Guided RF Odometry

Kunzhe Song **Jialuo Du** **Jingkai Lin** **Huacheng Zeng**
Michigan State University
{songkunz,dujialuo,linjingk,hzeng}@msu.edu

Abstract

Radio-frequency (RF) sensing enables robot ego-motion estimation in darkness, smoke, and fog, yet RF odometry remains challenging because multipath and moving reflectors make cross-frame matching highly ambiguous. Existing methods often rely on auxiliary inertial sensors or ground-truth pose supervision. We present RFWalk, a self-supervised, RF-only odometry framework that models dense correspondences across range-azimuth frames as random walks on a space-time graph and learns them from unlabeled sequences through temporal cycle consistency. To rule out correspondences that are cycle-consistent but physically implausible, we introduce a Doppler consistency loss that constrains correspondence-induced motion using radial velocity measured by the RF sensor itself. Ego-motion is then recovered through robust $\mathrm{SE}(2)$ registration that suppresses uncertain correspondences and those arising from independently moving objects. Without pose labels or auxiliary sensors, RFWalk remains competitive with supervised baselines in static scenes and reduces absolute trajectory error by 58% in dynamic scenes with moving pedestrians.

1 Introduction

Reliable ego-motion estimation is a fundamental capability for autonomous robots, underpinning localization, mapping, and navigation. Cameras and LiDAR have become the dominant sensing modalities for odometry [6, 12, 37], yet their performance can degrade substantially in darkness, smoke, fog, and dust. Radio-frequency (RF) sensing, particularly millimeter-wave (mmWave) radar, provides a complementary sensing modality that is robust to illumination changes and airborne obscurants while remaining compact and inexpensive enough for mobile robots [1, 14, 25, 38]. These properties have motivated growing interest in RF-based odometry and simultaneous localization and mapping (SLAM) [22, 26, 32].

Despite these advantages, estimating ego-motion from RF observations alone remains challenging. The core difficulty lies in establishing reliable correspondences between consecutive RF frames. Compared with optical images or LiDAR point clouds, range-azimuth (RA) heatmaps have coarse angular resolution and are heavily affected by multipath propagation and specular reflections, producing unstable or spurious structures that may vanish or shift across frames. In feature-poor or repetitive environments, multiple candidate matches can be geometrically indistinguishable. Moving reflectors such as pedestrians further complicate the problem, since correspondences on independently moving objects do not follow the rigid ego-motion shared by the static scene.

Existing RF odometry methods follow several strategies to address these challenges, but each leaves part of the problem unresolved. Scan-matching approaches [2, 28] align consecutive RF observations geometrically, yet remain vulnerable to multipath, coarse angular resolution, and geometric degeneracy. Radar-inertial systems [8, 20, 26, 39] improve motion observability by incorporating an IMU, at the cost of additional sensing, calibration, and synchronization. Doppler-based approaches [19, 32] exploit radar-measured radial velocity as a direct motion cue, but typically assume that most reflectors are static. Moreover, learning-based components in these pipelines [28,

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

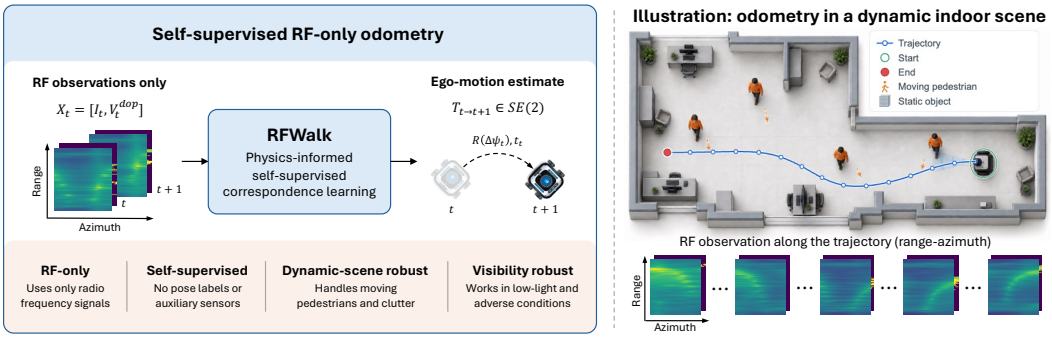


Figure 1: Self-supervised RF-only odometry. Left: from consecutive RF frames alone, RFWalk learns cross-frame correspondences without pose labels or auxiliary sensors and recovers the ego-motion $T_{t\to t+1}$. Right: illustration of RF-only odometry in a dynamic indoor scene with moving pedestrians.

32] require ground-truth trajectories or co-located LiDAR supervision, both of which are costly to collect at scale. As a result, accurate RF-only odometry without pose supervision remains an open problem, particularly in dynamic environments.

In this work, we introduce RFWalk, a self-supervised odometry framework that relies solely on RF observations and formulates ego-motion estimation as physics-informed cross-frame correspondence learning (Fig. 1). Rather than regressing poses directly, RFWalk models dense correspondences between RA frames as random walks on a space-time graph [16, 30], learns these correspondences from unlabeled RF sequences, and recovers ego-motion from the resulting correspondence field. Neither ground-truth poses nor auxiliary sensors are required during training or inference.

The key idea is to combine two complementary sources of self-supervision intrinsic to the RF signal. Temporal cycle consistency encourages correspondences that persist across time, while Doppler measurements provide an independent physical constraint on the radial displacement implied by each correspondence. Unlike prior Doppler-based methods, we do not use Doppler to estimate ego-velocity directly; instead, we use it to determine whether each local cross-frame association is physically plausible. This constraint applies to both static and moving reflectors and penalizes incorrect matches that remain cycle-consistent in repetitive or geometrically ambiguous scenes. We further enforce consistency between composed multi-step correspondences and their direct counterparts to reduce drift, and recover motion through robust $\mathrm{SE}(2)$ registration that suppresses uncertain or independently moving correspondences. Together, these components turn ambiguous RF matching into physically grounded, self-supervised motion estimation.

We evaluate RFWalk on the public Radarize benchmark [32] and on the RFWalk dataset, which we collect using a 60 GHz mmWave ISAC platform in indoor scenes with moving pedestrians. Without pose supervision or auxiliary sensing, RFWalk remains competitive with supervised baselines in static scenes and reduces absolute trajectory error by 58% relative to the strongest baseline in dynamic scenes. Its advantage becomes more pronounced as scene dynamics increase. From the least to the most dynamic scenes, its error grows by less than 1 m, compared with 2.3–8.2 m for the supervised baselines. These results suggest that physics-informed correspondence learning offers a promising path toward scalable RF odometry from unlabeled sequences.

We summarize our main contributions as follows.

- We formulate RF odometry as physics-informed, self-supervised correspondence learning, enabling dense cross-frame associations to be learned from unlabeled RF sequences without ground-truth poses or auxiliary sensors.
- We introduce Doppler consistency supervision, which converts per-cell radial velocity into a local physical constraint on correspondence-induced displacement. This constraint applies to both static and moving reflectors and suppresses physically incompatible matches that cycle consistency alone cannot rule out.
- We enforce compositional consistency across multi-step correspondences and recover motion through confidence-aware robust $\mathrm{SE}(2)$ registration, improving long-horizon stability and robustness to dynamic clutter.

2 Method

2.1 RF Observation Representation

We consider RF odometry in an RF-only setting, where only RF measurements are available during both training and inference. At each time step t, we represent the RF observation as a two-channel RA tensor

$$\mathbf{X}_t = \left[\mathbf{I}_t, \mathbf{V}_t^{\mathrm{dop}}\right] \in \mathbb{R}^{N_r \times N_\theta \times 2}, \tag{1}$$

where $\mathbf{I}_t$ is the RA log-power map and $\mathbf{V}_t^{\mathrm{dop}}$ is a signed radial-velocity map containing one scalar Doppler velocity for each RA cell. Here, N_r and N_θ denotes the numbers of range and azimuth bins, respectively. Let $N = N_r N_\theta$ be the number of RA cells. We index each RA cell by $i \in \{1, \ldots, N\}$ and associate it with the Cartesian coordinate of its bin center, $\mathbf{p}_i = [r_i \cos\theta_i \, , r_i \sin\theta_i]^\top \in \mathbb{R}^2$. Thus, each cell jointly encodes the reflected RF strength at a physical spatial location and the instantaneous radial motion observed along its line of sight.

We use a scalar radial-velocity map instead of retaining all Doppler FFT bins for two reasons. First, radial velocity is an explicit physical quantity that can be directly related to correspondence-induced radial displacement, enabling the physics-informed supervision introduced in Sec. 2.3 without requiring the network to infer the relevant Doppler component from a spectrum. Second, the number, spacing, and velocity range of Doppler bins vary substantially across radar waveforms and configurations. Collapsing the spectrum into metric radial velocity therefore provides a consistent two-channel input interface after platform-specific signal processing, while avoiding sensor-specific changes to the network input dimensionality. Given two consecutive frames $\mathbf{X}_t$ and $\mathbf{X}_{t+1}$, our goal is to estimate the relative planar motion

$$\mathbf{T}_{t\to t+1} = \big(\mathbf{R}(\Delta\psi_t), \mathbf{t}_t\big) \in \mathrm{SE}(2), \tag{2}$$

where $\mathbf{R}(\Delta\psi_t) \in \mathbb{R}^{2\times 2}$ is the relative rotation and $\mathbf{t}_t \in \mathbb{R}^2$ is the translation. The inter-frame interval Δt is obtained from RF timestamps.

2.2 Cross-Frame Correspondence

We formulate RF correspondence on the RA grid as a probabilistic walk on a space-time graph, where each RA cell is treated as a node and inter-frame edges encode soft correspondence probabilities. An overview is shown in Fig. 2. Following GMRW [30], a CNN encoder and global matching transformer map two RF observations $\mathbf{X}_t$ and $\mathbf{X}_{t+1}$ to cell embeddings $\mathbf{F}_t, \mathbf{F}_{t+1} \in \mathbb{R}^{N\times d}$. We define the transition probability from source cell i at time t to target cell j at time $t+1$ as

$$A_{t\to t+1}(i,j) = \frac{\exp\big(\mathbf{f}_{t,i}^\top \mathbf{f}_{t+1,j}/\tau\big)}{\sum_{\ell=1}^{N} \exp\big(\mathbf{f}_{t,i}^\top \mathbf{f}_{t+1,\ell}/\tau\big)}, \tag{3}$$

where τ is a temperature parameter. Each row of $A_{t\to t+1}$ defines a probability distribution over target RA cells in frame $t+1$. To capture longer-range temporal structure, we compose adjacent-frame transitions into a k-step walk

$$\bar{A}_{t\to t+k} = \prod_{s=0}^{k-1} A_{t+s\to t+s+1}. \tag{4}$$

Under this composition, transient clutter and unstable reflections tend to yield diffuse paths whose probability decays across steps, whereas persistent scene structures induce sharper paths across time. We adopt the palindrome training scheme of CRW [16], pairing a forward walk from t to $t+k$ with its reverse walk. Reversing an RF sequence also requires inverting the sign of the Doppler channel, since radial velocity flips under time reversal. We enforce cycle consistency by requiring a random walker to return to its starting node after the forward–backward traversal:

$$\mathcal{L}_{\mathrm{rw}}^{(k)} = \mathcal{L}_{\mathrm{CE}}\big(\bar{A}_{t\to t+k}\bar{A}_{t+k\to t}, I\big), \tag{5}$$

where $\mathcal{L}_{\mathrm{CE}}$ denotes row-wise cross-entropy and I is the identity matrix. Cycle consistency admits a trivial solution in which the model simply copies absolute RA indices. To prevent this, we randomly shift the 2D sinusoidal positional encoding during training, as in [30].

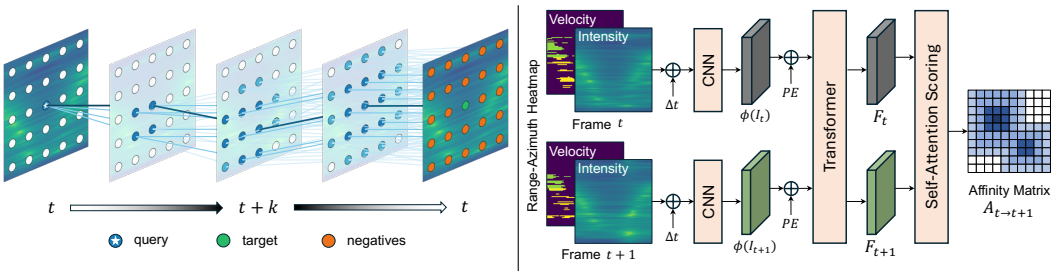


Figure 2: Overview of the proposed cross-frame correspondence framework. **Left**: training with a palindrome random walk objective, which encourages a random walker to return to its starting node after a forward-backward traversal. **Right**: given two RA tensors, we construct a cross-frame graph over RA cells and apply a global matching architecture to estimate soft correspondence probabilities.

2.3 Doppler Consistency Supervision

Pure cycle consistency does not guarantee physically correct correspondences. An incorrect transition may still participate in a self-consistent forward–backward walk if the reverse transition compensates for it. To address this limitation, we impose Doppler consistency directly on each one-step transition matrix, providing physics-grounded local supervision that requires the correspondence-induced radial displacement to agree with the Doppler-measured relative motion. This constraint reduces the feasible correspondence space by penalizing cycle-consistent but physically incompatible transitions. It is particularly beneficial in feature-poor or repetitive RF scenes, where intensity-based matching may otherwise admit multiple geometrically plausible correspondences.

As illustrated in Fig. 3(a), the learned transition matrix $A_{t\to t+1}$ induces a soft displacement field over RA cells. For a source cell i, we define its expected target location and correspondence-induced displacement as

$$\tilde{\mathbf{p}}_i = \sum_{j=1}^{N} A_{t\to t+1}(i,j)\,\mathbf{p}_j, \qquad \boldsymbol{\delta}_i = \tilde{\mathbf{p}}_i - \mathbf{p}_i, \tag{6}$$

where $\mathbf{p}_i = [r_i\cos\theta_i\,, r_i\sin\theta_i]^\top$ is the Cartesian coordinate of the source cell i. Here, $\tilde{\mathbf{p}}_i$ denotes the expected target location corresponding to source cell i, not the location of target cell i. Let

$$\mathbf{u}_i = \frac{\mathbf{p}_i}{\|\mathbf{p}_i\|_2} = [\cos\theta_i\,,\ \sin\theta_i]^\top. \tag{7}$$

denote the unit line-of-sight direction of that cell. Doppler measures the radial component of the relative velocity between the RF sensor and the observed reflector. Therefore, over a short inter-frame interval Δt, the radial projection of the correspondence-induced displacement should satisfy the first-order relation

$$\mathbf{u}_i^\top \boldsymbol{\delta}_i \;\approx\; \int_{t}^{t+\Delta t} -v_i^{\mathrm{dop}}(\tau)\,d\tau \;\approx\; -\Delta t\, v_i^{\mathrm{dop}}, \tag{8}$$

where v_i^{dop} is the measured Doppler velocity at cell i and $v_i^{\mathrm{dop}} > 0$ indicates that the reflector approaches the sensor. The relation in Eq. 8 holds for both static and dynamic reflectors, as illustrated in Fig. 3(b). For a static reflector, the relative radial motion arises solely from sensor ego-motion, whereas for a dynamic reflector it additionally includes the reflector's self-motion. Since both Doppler and correspondence-induced displacement measure the same sensor–reflector relative motion, the constraint remains valid in either case when the same physical scattering point is tracked.

Eq. 8 further provides a geometric constraint on the target correspondence. For a given source cell i and measured Doppler velocity v_i^{dop}, all target locations satisfying Eq. 8 lie on a line perpendicular to the line-of-sight direction $\mathbf{u}_i$, as illustrated in Fig. 3(b). Requiring exact equality is overly restrictive due to RA discretization and the first-order approximation over a finite Δt. We therefore introduce a tolerance margin $\delta > 0$, which expands the ideal line into an admissible band of width 2δ, and penalize only deviations beyond this band. Accordingly, we define the Doppler consistency loss as

$$\mathcal{L}_{\mathrm{dop}} = \frac{1}{\sum_i q_i}\sum_{i=1}^{N} q_i\,\max\Big(0,\ \Big|\mathbf{u}_i^\top \boldsymbol{\delta}_i + \Delta t\, v_i^{\mathrm{dop}}\Big| - \delta\Big)\,, \tag{9}$$

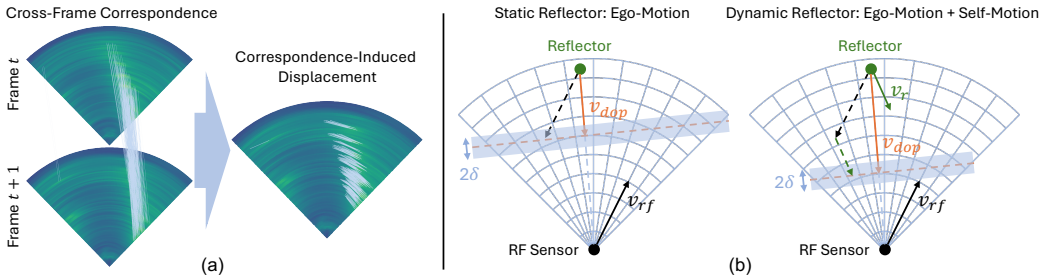


Figure 3: Doppler provides a first-order physical constraint on correspondence-induced motion. (**a**) Cross-frame correspondence induces a per-cell displacement field on the RA grid. (**b**) Doppler measures the sensor-reflector relative radial velocity, which arises from sensor ego-motion alone for a static reflector and from ego-motion together with the reflector's own motion for a dynamic one.

where $q_i \in [0, 1]$ is a data-dependent weight that downweights unreliable Doppler measurements.

2.4 Training Objective

Group composition consistency. Long-horizon odometry is obtained by recursively composing pairwise motions, so local correspondence errors can accumulate into trajectory drift. We therefore encourage the learned correspondences to be temporally compositional.

Given a temporal group $(\mathbf{X}_t, \mathbf{X}_{t+1}, \ldots, \mathbf{X}_{t+g})$, we construct the endpoint correspondence from $\mathbf{X}_t$ and $\mathbf{X}_{t+g}$ in two ways. The composed transition $\bar{A}_{t\to t+g}$ aggregates the intermediate one-step transitions via Eq. 4, while the direct transition $A_{t\to t+g}$ is obtained by applying Eq. 3 to the endpoint pair. Since both describe correspondence between the same frames, we enforce their agreement using a row-wise symmetric KL divergence:

$$\mathcal{L}_{\text{gc}}^{(g)} = \frac{1}{2} D_{\text{KL}}^{\text{row}}\left(\bar{A}_{t\to t+g} \,\|\, A_{t\to t+g}\right) + \frac{1}{2} D_{\text{KL}}^{\text{row}}\left(A_{t\to t+g} \,\|\, \bar{A}_{t\to t+g}\right). \tag{10}$$

This regularizes locally plausible but mutually inconsistent transitions and reduces drift under recursive pose composition.

Overall objective. Combining the above terms, the final training objective is

$$\mathcal{L} = \lambda_{\text{rw}} \frac{1}{|\mathcal{K}|} \sum_{k\in\mathcal{K}} \mathcal{L}_{\text{rw}}^{(k)} + \lambda_{\text{dop}} \mathcal{L}_{\text{dop}} + \lambda_{\text{gc}} \frac{1}{|\mathcal{G}|} \sum_{g\in\mathcal{G}} \mathcal{L}_{\text{gc}}^{(g)}, \tag{11}$$

where $\mathcal{K}$ denotes the set of palindrome lengths used for random-walk supervision, and $\mathcal{G}$ denotes the set of temporal group sizes used for composition consistency. λ_{rw}, λ_{dop}, and λ_{gc} are scalar weights. All three losses are self-supervised and require no ground-truth poses or correspondence annotations.

2.5 Odometry Estimation

The transition matrix $A_{t\to t+1}$ defines, for each source RA cell i, a discrete distribution over candidate target cells in frame $t+1$. Rather than collapsing this distribution with an $\arg\max$, we recover a soft Cartesian correspondence by taking its expectation over the target-cell coordinates:

$$\tilde{\mathbf{p}}_i = \sum_{j=1}^{N} A_{t\to t+1}(i,j)\, \mathbf{p}_j, \qquad \mathbf{\Sigma}_i = \sum_{j=1}^{N} A_{t\to t+1}(i,j)(\mathbf{p}_j - \tilde{\mathbf{p}}_i)(\mathbf{p}_j - \tilde{\mathbf{p}}_i)^\top. \tag{12}$$

Here, $\tilde{\mathbf{p}}_i$ is the expected target location of source cell i, while $\mathbf{\Sigma}_i$ measures the spatial dispersion of its correspondence distribution. Not all RA cells should contribute equally to ego-motion estimation. Low-SNR bins and multipath reflections often induce diffuse transition distributions and should contribute less to the final motion estimate. We therefore assign each source cell a match-quality prior

$$\bar{w}_i = m_i \, \exp(-\beta \, \text{tr}(\mathbf{\Sigma}_i)) \left(1 - \frac{\mathcal{H}(A_i)}{\log N}\right), \qquad \mathcal{H}(A_i) = -\sum_{j=1}^{N} A_{t\to t+1}(i,j) \log A_{t\to t+1}(i,j), \tag{13}$$

where A_i denotes the i-th row of $A_{t\to t+1}$, $\mathcal{H}(A_i)$ is its entropy, and $m_i \in [0,1]$ is a measurement-confidence prior computed from the source RA response. This weight favors source cells with strong measurements and sharp, spatially concentrated correspondence distributions.

A more fundamental challenge is that only the static part of the scene should obey a single global $\mathrm{SE}(2)$ motion. Reflectors on walking people, opening doors, or other independently moving objects violate this assumption and should be treated as outliers rather than forced into the rigid fit. To this end, we estimate odometry with a robust M-estimator:

$$(\mathbf{R}_t, \mathbf{t}_t) = \arg \min_{\mathbf{R}\in \mathrm{SO}(2),\, \mathbf{t}\in\mathbb{R}^2} \sum_{i=1}^{N} \bar{w}_i \, \rho(\|\mathbf{R}\mathbf{p}_i + \mathbf{t} - \tilde{\mathbf{p}}_i\|_2)\,, \tag{14}$$

where $\rho(\cdot)$ is the Huber loss,

$$\rho(r) = \begin{cases} \frac{1}{2}r^2, & r \le c, \\ c\left(r - \frac{1}{2}c\right), & r > c, \end{cases} \tag{15}$$

with threshold $c > 0$. Eq. 14 preserves high-confidence static correspondences while suppressing large-residual outliers caused by dynamic objects or gross matching errors. We solve it via iteratively reweighted least squares and provide the complete closed-form derivation in Appendix A.

3 Experiments

3.1 Experimental Setup

Datasets. We evaluate on two complementary datasets that differ substantially in carrier frequency, bandwidth, antenna configuration, sensing range, and scene dynamics.

Radarize dataset. Radarize [32] is a public dataset collected with a TI IWR1843 single-chip mmWave radar operating in the 76–81 GHz band. The radar has 3 transmit and 4 receive antennas and provides a maximum range of 4.284 m, a range resolution of 4 cm, and a frame rate of 30 Hz. The dataset contains 146 trajectories recorded across 4 buildings, covering approximately 4.7 km in total.

RFWalk dataset. We collect our own data with a custom-designed 60 GHz monostatic integrated sensing and communication (ISAC) platform (Fig. 4). The platform simultaneously transmits and receives OFDM communication waveforms, from which RA frames are formed. It has a bandwidth of 1.2288 GHz, corresponding to a range resolution of about 12 cm, uses 16 transmit and 16 receive phased-array antennas, and records frames at 10 Hz. The RFWalk dataset contains 80 trajectories with a total length of 3.2 km. In contrast to Radarize, it is recorded in the presence of moving pedestrians, which introduce substantial dynamic clutter into the RF observations.

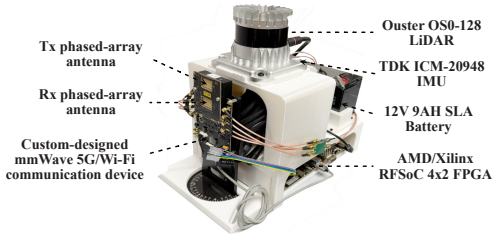


Figure 4: Custom-designed 60 GHz mmWave monostatic ISAC platform.

Evaluation metrics. Following the protocol of Radarize [32], we adopt two standard trajectory metrics. Absolute Trajectory Error (ATE) measures the global consistency of the estimated trajectory, whereas Relative Error (RE) measures local odometry accuracy as the drift accumulated over short trajectory segments. We report both metrics separately for translation (m) and rotation (deg) and compute them with the `evo` toolbox [13].

Baselines. We compare against five baselines that span the major paradigms of RF odometry. The first two are classical scan-matching methods that require no training. (i) ICP [3] iteratively registers consecutive RF scans by minimizing point-to-point distances. (ii) NDT [5] models each scan as a set of local normal distributions and registers consecutive scans by maximizing their likelihood. The remaining three are learning-based methods. (iii) RadarHD [28] upsamples range-azimuth heatmaps with a U-Net and feeds the enhanced heatmaps to a scan-matching SLAM backend. (iv) milliEgo [26] is a learning-based fusion approach that combines mmWave radar with an inertial measurement unit (IMU). (v) Radarize [32] decomposes ego-motion into Doppler-based translation estimation and

Table 1: Odometry accuracy on the Radarize (static) and RFWalk (dynamic) datasets. ATE measures global trajectory consistency and RE measures local odometry drift, each reported for translation (Pos, m) and rotation (Rot, deg). Lower is better for all metrics, and the best result in each column is shown in bold. Supervision indicates whether a method requires ground-truth poses for training.

Methods	Supervision	Radarize (Static Scenes)				RFWalk (Dynamic Scenes)			
		ATE		RE		ATE		RE	
		Pos (m)	Rot (deg)	Pos (m)	Rot (deg)	Pos (m)	Rot (deg)	Pos (m)	Rot (deg)
ICP [3]	✗	8.613	7.104	0.027	0.022	16.482	8.036	0.052	0.035
NDT [5]	✗	7.421	6.372	0.023	0.019	14.208	7.184	0.045	0.031
RadarHD [28]	✓	6.367	5.820	0.020	0.016	12.724	6.215	0.041	0.027
milliEgo [26]	✓	4.732	7.774	0.014	0.020	7.406	5.123	0.022	0.024
Radarize [32]	✓	**1.657**	**2.445**	**0.007**	**0.008**	5.642	4.938	0.018	0.022
RFWalk (Ours)	✗	1.934	3.191	0.008	0.010	**2.357**	**2.607**	**0.010**	**0.013**

correlation-based rotation estimation. All three learning-based baselines are trained with ground-truth pose supervision, whereas RFWalk uses none. All methods are evaluated on the same test trajectories.

Implementation details. The CNN encoder consists of two convolutional blocks followed by a linear projection that maps each RA cell to a d=128-dimensional embedding. The transformer has 4 layers with 8 attention heads, and the temperature of the transition matrix is set to $\tau=\sqrt{d}$. During training, we use palindrome walk lengths $\mathcal{K}=\{1,2,3\}$ and group composition sizes $\mathcal{G}=\{2,3\}$. The loss weights are $\lambda_{\mathrm{rw}}=1.0$, $\lambda_{\mathrm{dop}}=0.5$, and $\lambda_{\mathrm{gc}}=0.2$, and the tolerance margin in the Doppler consistency loss is $\delta=0.02$ m. For odometry estimation, we set the dispersion decay factor to $\beta=5.0$ and the Huber threshold to $c=0.1$ m. All models are trained on 4 NVIDIA RTX A6000 GPUs.

3.2 Main Results

For the Radarize dataset, we follow the official train/test split [32]. For the RFWalk dataset, we randomly assign 60% of the trajectories to training and the remaining 40% to testing. The learning-based baselines are trained with ground-truth poses obtained from LiDAR and IMU, whereas RFWalk is trained only on unlabeled RF sequences. Tab. 1 summarizes the quantitative results.

Performance on static scenes. On the Radarize dataset, the supervised Radarize baseline achieves the best accuracy on all four metrics, as its Doppler-based translation estimator provides a strong inductive bias when the scene is static. Without any pose supervision, RFWalk remains close to Radarize, and it outperforms all other baselines on every metric. The training-free ICP and NDT perform worst, since coarse angular resolution, multipath, and sparse returns make direct geometric registration of RF scans unreliable. These results show that self-supervised correspondence learning alone recovers ego-motion with accuracy approaching that of the strongest supervised method in static environments, as also reflected in the trajectories in Fig. 5.

Performance on dynamic scenes. On the RFWalk dataset with frequent pedestrian motion, RFWalk achieves the best accuracy on all four metrics. ICP and NDT yield the largest errors, because both assume a rigid scene and therefore register moving pedestrians as if they were static structure. We attribute the robustness of RFWalk to two complementary mechanisms. First, the cycle-consistent random walk objective suppresses correspondences on transient clutter, since moving objects induce diffuse paths whose probability decays under multi-step composition. Second, the robust M-estimator in Eq. 14 downweights the remaining large-residual correspondences on independently moving objects. Together, these mechanisms allow RFWalk to handle dynamic objects without explicit motion segmentation, as illustrated in Fig. 6.

3.3 Ablation Study

We further analyze the robustness of RFWalk to scene dynamics and the contribution of its key design choices. All experiments in this section are performed on the RFWalk dataset, which provides a more challenging testbed due to scene dynamics and geometric degeneracy.

Robustness to dynamic objects. To examine how scene dynamics affect each method, we partition the test trajectories into three groups by the number of moving pedestrians and report the ATE Pos of

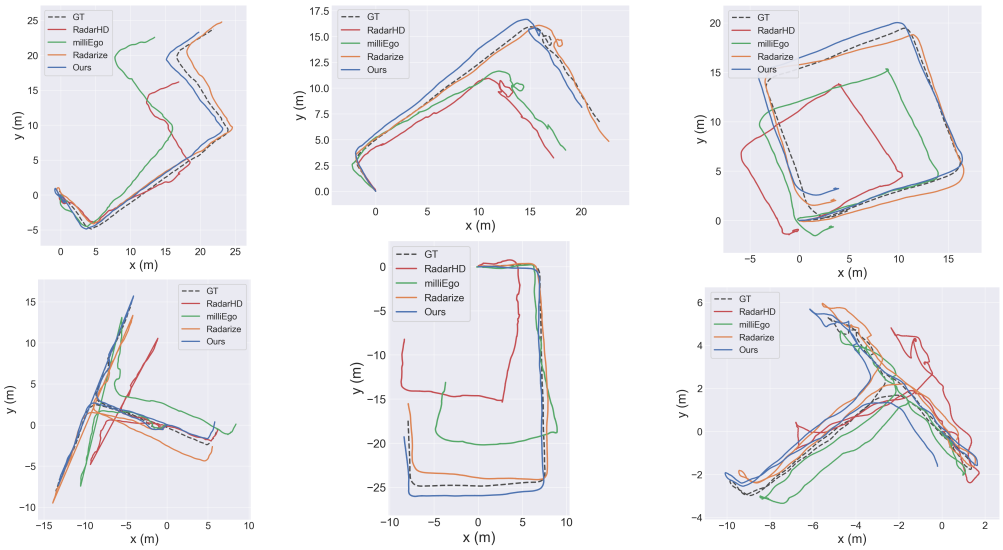


Figure 5: Qualitative trajectory comparison on the Radarize dataset. Each subplot shows the estimated trajectory of all methods overlaid with the ground truth (GT) marked by the dashed black line.

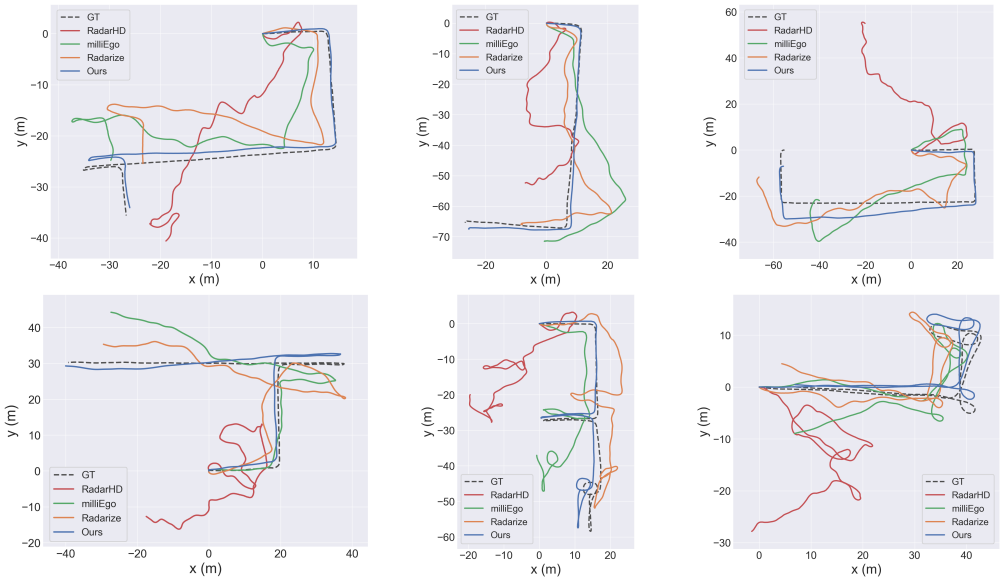


Figure 6: Qualitative trajectory comparison on the RFWalk dataset with dynamic pedestrians. All supervised baselines exhibit notable degradation in the presence of moving objects.

each group in Tab. 2. RFWalk achieves the lowest error at every level, and its error increases by only 0.93 m from the Low to the High group, compared with 2.29 m to 8.24 m for the supervised baselines. RadarHD degrades the most, since its scan-matching backend assumes a rigid scene and is therefore directly corrupted by moving pedestrians. milliEgo degrades more gracefully owing to its IMU branch, yet its fusion network still ingests corrupted RF features and cannot fully compensate. Radarize attains lower errors than milliEgo at every level, but its error grows faster, from 4.83 m to 8.81 m, because its Doppler-based translation estimator conflates pedestrian velocities with ego-motion.

Effect of group composition size. Tab. 3 varies the temporal group sizes $\mathcal{G}$ in the group composition consistency loss. Without this loss ($\mathcal{G}=\emptyset$), the palindrome objective constrains each forward walk only through its own backward return, so locally plausible but mutually incompatible transitions can still accumulate into drift. Adding $\mathcal{G}=\{2\}$ and then $\{2,3\}$ steadily improves translational accuracy, and

Table 2: ATE Pos (m) under increasing levels of scene dynamics.

Methods	Low	Medium	High
RadarHD [28]	10.243	13.175	18.478
milliEgo [26]	6.812	7.934	9.105
Radarize [32]	4.831	6.042	8.813
RFWalk	**2.285**	**2.672**	**3.214**

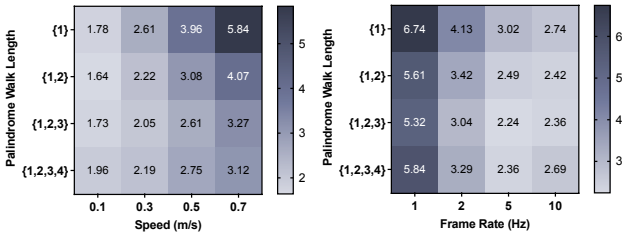
Low, Medium, and High correspond to scenes with 0–2, 3–4, and 5+ moving pedestrians, respectively.

Table 3: Effect of group composition sizes $\mathcal{G}$. $\emptyset$ denotes training without $\mathcal{L}_{\text{gc}}$.

$\mathcal{G}$	ATE		RE	
	Pos (m)	Rot (deg)	Pos (m)	Rot (deg)
$\emptyset$	2.675	**2.487**	0.013	**0.011**
$\{2\}$	2.402	2.575	0.011	0.013
$\{2,3\}$	**2.357**	2.607	**0.010**	0.013
$\{2,3,4\}$	2.728	3.253	0.012	0.016

$\{2,3\}$ reduces ATE Pos by 12% and RE Pos by 23% relative to $\emptyset$. Extending the groups to $\{2,3,4\}$ degrades all metrics, as direct transitions over longer horizons become less reliable and enforcing agreement with them propagates their errors back to the one-step transitions. We also observe a mild trade-off in rotation, where $\mathcal{L}_{\text{gc}}$ slightly increases ATE Rot from 2.487 deg to 2.607 deg. Since translational drift dominates the overall trajectory error, we adopt $\mathcal{G}{=}\{2,3\}$ by default.

Effect of palindrome walk length. The palindrome walk length $\mathcal{K}$ controls the temporal horizon over which cycle consistency is enforced, and its optimal choice depends on the effective inter-frame motion. We disentangle this dependence along two complementary axes in Fig. 7.



(a) Varying speed. (b) Varying frame rate.

Figure 7: Effect of palindrome walk length $\mathcal{K}$ across speeds (m/s) and frame rates (Hz), measured by ATE Pos (m).

Fig. 7(a) varies the test-time platform speed while training each model with a different $\mathcal{K}$. The penalty of longer chains is most pronounced at slow speeds, where per-frame displacements are small and multi-step composition mainly amplifies correspondence noise. As speed grows, the gap between walk lengths widens substantially, since longer palindromes provide stronger multi-step regularization when inter-frame displacement is large. Fig. 7(b) jointly varies the frame rate at training and inference. At 1 Hz, sparse temporal sampling produces large per-step motions, and walks beyond two steps drift outside the reliable correspondence horizon. As the frame rate increases, the effective per-step displacement shrinks and longer chains become beneficial, with $\mathcal{K}{=}\{1,2,3\}$ attaining the lowest error.

Effect of individual components. To understand where the gains come from, we progressively add each component to the base random walk objective $\mathcal{L}_{\text{rw}}$ and report the results in Fig. 8. Together, these components reduce ATE Pos by 90% and ATE Rot by 52%.

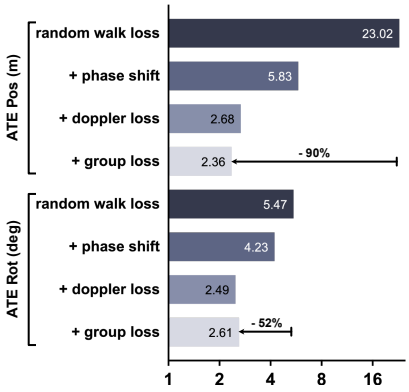


Figure 8: Cumulative ablation of training components. Each row adds one component to the row above.

The random phase shift yields the largest translational gain, reducing ATE Pos from 23.02 m to 5.83 m. Without it, the positional encodings of consecutive frames are aligned, so cycle consistency can be trivially satisfied by copying RA indices and the transition matrix collapses toward the identity. The phase shift removes this shortcut and forces the model to match cells by their RF appearance. The Doppler consistency loss $\mathcal{L}_{\text{dop}}$ yields the largest rotational gain, reducing ATE Rot from 4.23 deg to 2.49 deg, and also lowers ATE Pos by 54%. In feature-poor scenes, translational ambiguity in the correspondences couples into rotation during $\mathrm{SE}(2)$ fitting, and anchoring radial displacements to measured Doppler velocities resolves this ambiguity. Finally, the group composition consistency loss $\mathcal{L}_{\text{gc}}$ further reduces ATE Pos to 2.36 m by mitigating drift under recursive pose composition.

4 Related Work

4.1 Self-Supervised Space-Time Correspondence

Learning correspondences across frames underlies video segmentation, tracking, and 3D reconstruction [27, 29]. Supervised optical flow [11, 33, 34] and tracking-any-point (TAP) methods [9, 10, 15, 17, 18] achieve high accuracy on natural videos but rely largely on synthetic annotations or pseudo-labels. Self-supervised methods instead learn correspondences from unlabeled videos through proxy tasks [23, 35], or through temporal cycle consistency [24, 36]. CRW [16] formulates correspondence as a contrastive random walk on a space-time graph. This formulation has been extended to multiscale pixel trajectories [4], combined with a global matching transformer for all-pairs any-point tracking in GMRW [30], and applied to cross-modal matching among RGB, depth, and thermal images [31].

More fundamentally, cycle consistency is derived from the observations alone and is satisfied by any set of transitions whose composition returns to the start. In natural images, rich appearance makes it a reliable proxy for correct matching, but RA heatmaps have coarse resolution and repetitive structures that admit many cycle-consistent yet physically incorrect matches. RFWalk brings the random-walk formulation to RF and imposes Doppler consistency directly on its soft transition probabilities rather than on regressed flow, so that every correspondence between dense RA cells must agree with the radial motion measured by the radar.

4.2 RF-Based Odometry

Geometric radar odometry estimates motion by aligning consecutive RF observations. Early methods extract and match landmarks from scanning radar [7]. CFEAR [2] registers sparse oriented surface points, and RadarHD [28] learns to generate LiDAR-like point clouds from single-chip mmWave radar for subsequent scan matching. However, coarse angular resolution, multipath ghosts, and geometric degeneracy in feature-poor environments make such registration unreliable, especially for low-cost single-chip radars. Radar-inertial systems improve motion observability by fusing an IMU through filtering or optimization [8, 20, 39] or through deep sensor fusion [26]. These systems, however, require additional sensing, calibration, and synchronization. Learning-based radar odometry has largely relied on pose supervision. Recent mmWave pipelines [28, 32] require co-located LiDAR or ground-truth trajectories during training.

Doppler provides a complementary motion cue by measuring the radial velocity of each reflector. Radarize [32] combines Doppler-based translation with correlation-based rotation estimation, while PanoRadar [21] and CartoRadar [22] incorporate Doppler into motion compensation and RF SLAM. These methods treat Doppler as a measurement of ego-motion and therefore assume that reflectors are static. RFWalk instead uses Doppler as local supervision for correspondence learning, and recovers ego-motion through robust $\mathrm{SE}(2)$ registration that suppresses independently moving correspondences. The resulting pipeline requires neither ground-truth trajectories nor auxiliary sensors.

5 Conclusion

We presented RFWalk, a self-supervised, RF-only odometry framework that learns cross-frame correspondences without ground-truth poses or auxiliary sensors. By combining cycle-consistent learning with Doppler-based physical supervision, RFWalk achieves competitive accuracy in static environments and substantial improvements in dynamic scenes.

Limitations and future work. Several limitations remain. First, global correspondence between RA frames scales quadratically with the number of RA cells N. Second, the current formulation estimates only planar $\mathrm{SE}(2)$ motion. Third, each model is trained and evaluated on a single radar configuration, and zero-shot transfer across RF sensors with different carrier frequencies, bandwidths, and antenna layouts remains unexplored. However, the label-free training formulation provides a natural path toward addressing this limitation. Since RFWalk requires no pose labels and represents RF observations using a common intensity and radial-velocity interface across radar configurations, it can in principle be trained on unlabeled sequences pooled from heterogeneous RF platforms. Future work will therefore explore efficient correspondence architectures, full $\mathrm{SE}(3)$ motion estimation, and large-scale self-supervised pretraining of general-purpose RF motion representations across sensors.

Acknowledgments

We sincerely thank the reviewers for their insightful comments and suggestions. This work is supported in part by the National Science Foundation (NSF) under Award No. ECCS-2434001.

References

[1] Fadel Adib, Chen-Yu Hsu, Hongzi Mao, Dina Katabi, and Frédo Durand. 2015. Capturing the human figure through a wall. *ACM Transactions on Graphics (TOG)* 34, 6 (2015), 1–13.

[2] Daniel Adolfsson, Martin Magnusson, Anas Alhashimi, Achim J Lilienthal, and Henrik Andreasson. 2021. Cfear radarodometry-conservative filtering for efficient and accurate radar odometry. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 5462–5469.

[3] Paul J Besl and Neil D McKay. 1992. Method for registration of 3-D shapes. In *Sensor fusion IV: control paradigms and data structures*, Vol. 1611. Spie, 586–606.

[4] Zhangxing Bian, Allan Jabri, Alexei A Efros, and Andrew Owens. 2022. Learning pixel trajectories with multiscale contrastive random walks. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 6498–6509.

[5] Peter Biber and Wolfgang Straßer. 2003. The normal distributions transform: A new approach to laser scan matching. In *Proceedings 2003 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2003)(Cat. No. 03CH37453)*, Vol. 3. IEEE, 2743–2748.

[6] Carlos Campos, Richard Elvira, Juan J Gómez Rodríguez, José MM Montiel, and Juan D Tardós. 2021. Orb-slam3: An accurate open-source library for visual, visual–inertial, and multimap slam. *IEEE transactions on robotics* 37, 6 (2021), 1874–1890.

[7] Sarah H Cen and Paul Newman. 2018. Precise ego-motion estimation with millimeter-wave radar under diverse and challenging conditions. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 6045–6052.

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

[9] Carl Doersch, Ankush Gupta, Larisa Markeeva, Adria Recasens, Lucas Smaira, Yusuf Aytar, Joao Carreira, Andrew Zisserman, and Yi Yang. 2022. Tap-vid: A benchmark for tracking any point in a video. *Advances in Neural Information Processing Systems* 35 (2022), 13610–13626.

[10] Carl Doersch, Yi Yang, Mel Vecerik, Dilara Gokay, Ankush Gupta, Yusuf Aytar, Joao Carreira, and Andrew Zisserman. 2023. Tapir: Tracking any point with per-frame initialization and temporal refinement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 10061–10072.

[11] Alexey Dosovitskiy, Philipp Fischer, Eddy Ilg, Philip Hausser, Caner Hazirbas, Vladimir Golkov, Patrick Van Der Smagt, Daniel Cremers, and Thomas Brox. 2015. Flownet: Learning optical flow with convolutional networks. In *Proceedings of the IEEE international conference on computer vision*. 2758–2766.

[12] Jakob Engel, Vladlen Koltun, and Daniel Cremers. 2017. Direct sparse odometry. *IEEE transactions on pattern analysis and machine intelligence* 40, 3 (2017), 611–625.

[13] Michael Grupp. 2017. evo: Python package for the evaluation of odometry and SLAM. `https://github.com/MichaelGrupp/evo`.

[14] Junfeng Guan, Sohrab Madani, Suraj Jog, Saurabh Gupta, and Haitham Hassanieh. 2020. Through fog high-resolution imaging using millimeter wave radar. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 11464–11473.

[15] Adam W Harley, Zhaoyuan Fang, and Katerina Fragkiadaki. 2022. Particle video revisited: Tracking through occlusions using point trajectories. In *European Conference on Computer Vision*. Springer, 59–75.

[16] Allan Jabri, Andrew Owens, and Alexei Efros. 2020. Space-time correspondence as a contrastive random walk. *Advances in neural information processing systems* 33 (2020), 19545–19560.

[17] Nikita Karaev, Yuri Makarov, Jianyuan Wang, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. 2025. Cotracker3: Simpler and better point tracking by pseudo-labelling real videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 6013–6022.

[18] Nikita Karaev, Ignacio Rocco, Benjamin Graham, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. 2024. Cotracker: It is better to track together. In *European conference on computer vision*. Springer, 18–35.

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

[20] Andrew Kramer, Carl Stahoviak, Angel Santamaria-Navarro, Ali-Akbar Agha-Mohammadi, and Christoffer Heckman. 2020. Radar-inertial ego-velocity estimation for visually degraded environments. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 5739–5746.

[21] Haowen Lai, Gaoxiang Luo, Yifei Liu, and Mingmin Zhao. 2024. Enabling visual recognition at radio frequency. In *Proceedings of the 30th Annual International Conference on Mobile Computing and Networking*. 388–403.

[22] Haowen Lai, Zhiwei Zheng, and Mingmin Zhao. 2025. Rf-based 3d slam rivaling vision approaches. In *Proceedings of the 31st Annual International Conference on Mobile Computing and Networking*. 170–185.

[23] Zihang Lai, Erika Lu, and Weidi Xie. 2020. Mast: A memory-augmented self-supervised tracker. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 6478–6487.

[24] Xueting Li, Sifei Liu, Shalini De Mello, Xiaolong Wang, Jan Kautz, and Ming-Hsuan Yang. 2019. Joint-task self-supervised learning for temporal correspondence. *Advances in Neural Information Processing Systems* 32 (2019).

[25] Chris Xiaoxuan Lu, Stefano Rosa, Peijun Zhao, Bing Wang, Changhao Chen, John A Stankovic, Niki Trigoni, and Andrew Markham. 2020. See through smoke: robust indoor mapping with low-cost mmwave radar. In *Proceedings of the 18th International Conference on Mobile Systems, Applications, and Services*. 14–27.

[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. 2020. milliEgo: single-chip mmWave radar aided egomotion estimation via deep sensor fusion. In *Proceedings of the 18th Conference on Embedded Networked Sensor Systems*. 109–122.

[27] Federico Perazzi, Jordi Pont-Tuset, Brian McWilliams, Luc Van Gool, Markus Gross, and Alexander Sorkine-Hornung. 2016. A benchmark dataset and evaluation methodology for video object segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 724–732.

[28] Akarsh Prabhakara, Tao Jin, Arnav Das, Gantavya Bhatt, Lilly Kumari, Elahe Soltanaghai, Jeff Bilmes, Swarun Kumar, and Anthony Rowe. 2023. High resolution point clouds from mmwave radar. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 4135–4142.

[29] Johannes L Schonberger and Jan-Michael Frahm. 2016. Structure-from-motion revisited. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 4104–4113.

[30] Ayush Shrivastava and Andrew Owens. 2024. Self-supervised any-point tracking by contrastive random walks. In *European Conference on Computer Vision*. Springer, 267–284.

[31] Ayush Shrivastava and Andrew Owens. 2025. Self-supervised spatial correspondence across modalities. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 6383–6393.

[32] Emerson Sie, Xinyu Wu, Heyu Guo, and Deepak Vasisht. 2024. Radarize: Enhancing radar slam with generalizable doppler-based odometry. In *Proceedings of the 22nd Annual International Conference on Mobile Systems, Applications and Services*. 331–344.

[33] Deqing Sun, Xiaodong Yang, Ming-Yu Liu, and Jan Kautz. 2018. Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 8934–8943.

[34] Zachary Teed and Jia Deng. 2020. Raft: Recurrent all-pairs field transforms for optical flow. In *European conference on computer vision*. Springer, 402–419.

[35] Carl Vondrick, Abhinav Shrivastava, Alireza Fathi, Sergio Guadarrama, and Kevin Murphy. 2018. Tracking emerges by colorizing videos. In *Proceedings of the European conference on computer vision (ECCV)*. 391–408.

[36] Xiaolong Wang, Allan Jabri, and Alexei A Efros. 2019. Learning correspondence from the cycle-consistency of time. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2566–2576.

[37] Ji Zhang, Sanjiv Singh, et al. 2014. LOAM: Lidar odometry and mapping in real-time.. In *Robotics: Science and systems*, Vol. 2. Berkeley, CA, 1–9.

[38] Mingmin Zhao, Tianhong Li, Mohammad Abu Alsheikh, Yonglong Tian, Hang Zhao, Antonio Torralba, and Dina Katabi. 2018. Through-wall human pose estimation using radio signals. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 7356–7365.

[39] Yuan Zhuang, Binliang Wang, Jianzhu Huai, and Miao Li. 2023. 4d iriom: 4d imaging radar inertial odometry and mapping. *IEEE Robotics and Automation Letters* 8, 6 (2023), 3246–3253.

A Solving the Robust SE(2) Registration

In the main text, Eq. 14 formulates odometry estimation as a robust weighted registration problem on $\mathrm{SE}(2)$. Here we provide the complete derivation of its solution.

Recall that the optimization problem is

$$(\mathbf{R}^*, \mathbf{t}^*) = \arg\min_{\mathbf{R}\in\mathrm{SO}(2),\, \mathbf{t}\in\mathbb{R}^2} \sum_{i=1}^{N} \bar{w}_i \, \rho(\|\mathbf{R}\mathbf{p}_i + \mathbf{t} - \tilde{\mathbf{p}}_i\|_2)\,, \tag{A.1}$$

where $\mathbf{p}_i \in \mathbb{R}^2$ is the Cartesian position of source RA cell i, $\tilde{\mathbf{p}}_i \in \mathbb{R}^2$ is the expected target position defined in Eq. 12, $\bar{w}_i \geq 0$ is the match-quality prior from Eq. 13, and $\rho(\cdot)$ is the Huber loss with threshold $c > 0$.

A.1 Iteratively Reweighted Least Squares

Because the Huber loss is non-quadratic, we solve Eq. A.1 via iteratively reweighted least squares (IRLS). The key observation is that the Huber loss admits a *half-quadratic* interpretation. For any residual $r_i = \|\mathbf{R}\mathbf{p}_i + \mathbf{t} - \tilde{\mathbf{p}}_i\|_2$, the first derivative of ρ is

$$\rho'(r) = \begin{cases} r, & r \leq c, \\ c, & r > c. \end{cases} \tag{A.2}$$

The IRLS weight function, defined as the ratio $\psi(r) = \rho'(r)/r$ for $r > 0$, is

$$\psi(r) = \begin{cases} 1, & r \leq c, \\ c/r, & r > c. \end{cases} \tag{A.3}$$

At each IRLS iteration, the effective weight for cell i becomes

$$\omega_i = \bar{w}_i \cdot \psi(r_i), \tag{A.4}$$

and the original problem reduces to a *weighted least squares* rigid registration

$$(\mathbf{R}^*, \mathbf{t}^*) = \arg\min_{\mathbf{R}\in\mathrm{SO}(2),\, \mathbf{t}\in\mathbb{R}^2} \sum_{i=1}^{N} \omega_i \, \|\mathbf{R}\mathbf{p}_i + \mathbf{t} - \tilde{\mathbf{p}}_i\|_2^2\,. \tag{A.5}$$

This subproblem has a closed-form solution, which we derive below.

A.2 Closed-Form Weighted SE(2) Registration

We first eliminate the translational variable by centering both point sets. Define the weighted centroids

$$\boldsymbol{\mu}_{\mathbf{p}} = \frac{\sum_{i=1}^{N} \omega_i \, \mathbf{p}_i}{\sum_{i=1}^{N} \omega_i}, \qquad \boldsymbol{\mu}_{\tilde{\mathbf{p}}} = \frac{\sum_{i=1}^{N} \omega_i \, \tilde{\mathbf{p}}_i}{\sum_{i=1}^{N} \omega_i}, \tag{A.6}$$

and the centered coordinates

$$\hat{\mathbf{p}}_i = \mathbf{p}_i - \boldsymbol{\mu}_{\mathbf{p}}, \qquad \hat{\tilde{\mathbf{p}}}_i = \tilde{\mathbf{p}}_i - \boldsymbol{\mu}_{\tilde{\mathbf{p}}}. \tag{A.7}$$

Substituting $\mathbf{t} = \boldsymbol{\mu}_{\tilde{\mathbf{p}}} - \mathbf{R}\boldsymbol{\mu}_{\mathbf{p}} + \boldsymbol{\epsilon}$ into Eq. A.5 and expanding, the objective decouples into a term depending only on $\mathbf{R}$ (through centered coordinates) and a term depending only on $\boldsymbol{\epsilon}$:

$$\sum_{i=1}^{N} \omega_i \left\| \mathbf{R}\hat{\mathbf{p}}_i - \hat{\tilde{\mathbf{p}}}_i + \boldsymbol{\epsilon} \right\|_2^2 = \sum_{i=1}^{N} \omega_i \left\| \mathbf{R}\hat{\mathbf{p}}_i - \hat{\tilde{\mathbf{p}}}_i \right\|_2^2 + \left(\sum_{i=1}^{N} \omega_i \right) \|\boldsymbol{\epsilon}\|_2^2 , \tag{A.8}$$

where the cross term vanishes because $\sum_i \omega_i \hat{\mathbf{p}}_i = \sum_i \omega_i \hat{\tilde{\mathbf{p}}}_i = \mathbf{0}$ by definition of the centroids. Therefore $\boldsymbol{\epsilon}^* = \mathbf{0}$, and the optimal translation is

$$\mathbf{t}^* = \boldsymbol{\mu}_{\tilde{\mathbf{p}}} - \mathbf{R}^* \boldsymbol{\mu}_{\mathbf{p}}. \tag{A.9}$$

Optimal rotation. It remains to solve

$$\mathbf{R}^* = \arg \min_{\mathbf{R} \in \mathrm{SO}(2)} \sum_{i=1}^{N} \omega_i \left\| \mathbf{R}\hat{\mathbf{p}}_i - \hat{\tilde{\mathbf{p}}}_i \right\|_2^2 . \tag{A.10}$$

Expanding the squared norm and dropping terms independent of $\mathbf{R}$, we obtain the equivalent maximization

$$\mathbf{R}^* = \arg \max_{\mathbf{R} \in \mathrm{SO}(2)} \sum_{i=1}^{N} \omega_i \, \hat{\tilde{\mathbf{p}}}_i^{\top} \mathbf{R} \hat{\mathbf{p}}_i = \arg \max_{\mathbf{R} \in \mathrm{SO}(2)} \mathrm{tr}\big(\mathbf{R}\mathbf{H}^{\top}\big) , \tag{A.11}$$

where $\mathbf{H} \in \mathbb{R}^{2\times 2}$ is the weighted cross-covariance matrix

$$\mathbf{H} = \sum_{i=1}^{N} \omega_i \, \hat{\tilde{\mathbf{p}}}_i \, \hat{\mathbf{p}}_i^{\top} . \tag{A.12}$$

In two dimensions, $\mathbf{R}$ is parameterized by a single angle $\Delta\psi$,

$$\mathbf{R}(\Delta\psi) = \begin{bmatrix} \cos\Delta\psi & -\sin\Delta\psi \\ \sin\Delta\psi & \cos\Delta\psi \end{bmatrix} . \tag{A.13}$$

Writing $\mathbf{H}$ in components as

$$\mathbf{H} = \begin{bmatrix} h_{11} & h_{12} \\ h_{21} & h_{22} \end{bmatrix} , \tag{A.14}$$

the trace objective in Eq. A.11 evaluates to

$$\mathrm{tr}\big(\mathbf{R}\mathbf{H}^{\top}\big) = (h_{11} + h_{22}) \cos\Delta\psi + (h_{21} - h_{12}) \sin\Delta\psi. \tag{A.15}$$

This is a sinusoidal function in $\Delta\psi$. Taking its derivative with respect to $\Delta\psi$ and setting it to zero,

$$-(h_{11} + h_{22}) \sin\Delta\psi^* + (h_{21} - h_{12}) \cos\Delta\psi^* = 0, \tag{A.16}$$

which yields the optimal rotation angle

$$\boxed{\Delta\psi^* = \mathrm{atan2}(h_{21} - h_{12},\ h_{11} + h_{22}) .} \tag{A.17}$$

We can verify that this is indeed a maximum by checking the second derivative,

$$\frac{d^2}{d(\Delta\psi)^2} \mathrm{tr}\big(\mathbf{R}\mathbf{H}^{\top}\big) \bigg|_{\Delta\psi = \Delta\psi^*} = -\sqrt{(h_{11} + h_{22})^2 + (h_{21} - h_{12})^2} < 0, \tag{A.18}$$

confirming that $\Delta\psi^*$ maximizes the trace objective (assuming $\mathbf{H}$ is nonzero).

A.3 Complete IRLS Procedure

Combining the results above, we summarize the complete solver in Algorithm 1.

Algorithm 1 Robust SE(2) Odometry Estimation via IRLS.

Require: Source positions $\{\mathbf{p}_i\}_{i=1}^N$, soft target positions $\{\tilde{\mathbf{p}}_i\}_{i=1}^N$, quality priors $\{\bar{w}_i\}_{i=1}^N$, Huber threshold c, maximum iterations L, convergence tolerance ϵ
Ensure: Optimal rotation $\mathbf{R}^*$ and translation $\mathbf{t}^*$

Initialize $\omega_i \leftarrow \bar{w}_i$ for all i ▷ First iteration uses unit IRLS weights
for $\ell = 1, 2, \ldots, L$ **do**
  Compute weighted centroids $\boldsymbol{\mu}_{\mathbf{p}}$, $\boldsymbol{\mu}_{\tilde{\mathbf{p}}}$ via Eq. A.6
  Center the point sets $\hat{\mathbf{p}}_i \leftarrow \mathbf{p}_i - \boldsymbol{\mu}_{\mathbf{p}}$, $\hat{\tilde{\mathbf{p}}}_i \leftarrow \tilde{\mathbf{p}}_i - \boldsymbol{\mu}_{\tilde{\mathbf{p}}}$
  Compute cross-covariance $\mathbf{H} \leftarrow \sum_i \omega_i \, \hat{\tilde{\mathbf{p}}}_i \, \hat{\mathbf{p}}_i^\top$ ▷ Eq. A.12
  Compute optimal angle $\Delta\psi \leftarrow \operatorname{atan2}(h_{21} - h_{12},\ h_{11} + h_{22})$ ▷ Eq. A.17
  Form rotation $\mathbf{R} \leftarrow \mathbf{R}(\Delta\psi)$
  Compute translation $\mathbf{t} \leftarrow \boldsymbol{\mu}_{\tilde{\mathbf{p}}} - \mathbf{R}\,\boldsymbol{\mu}_{\mathbf{p}}$ ▷ Eq. A.9
  Compute residuals $r_i \leftarrow \|\mathbf{R}\mathbf{p}_i + \mathbf{t} - \tilde{\mathbf{p}}_i\|_2$ for all i
  Update effective weights $\omega_i \leftarrow \bar{w}_i \cdot \psi(r_i)$ via Eq. A.3 and Eq. A.4
  if $|\Delta\psi^{(\ell)} - \Delta\psi^{(\ell-1)}| < \epsilon$ **and** $\|\mathbf{t}^{(\ell)} - \mathbf{t}^{(\ell-1)}\| < \epsilon$ **then**
    break ▷ Convergence reached
  end if
end for
return $\mathbf{R}^* \leftarrow \mathbf{R}$, $\mathbf{t}^* \leftarrow \mathbf{t}$

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We clearly state our contributions in the abstract and introduction and conduct extensive experiments.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss the limitations in the Discussion session.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.

- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

   Justification: This work does not include theoretical proof.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We clearly detailed our methods and parameters for the training of the network. The code, and the dataset will be open-sourced once the paper is accepted.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often

one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.

- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: Our code and dataset will be open-sourced after the review process.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: We have all the training and testing details, including data splits, hyperparameters, ablation studies on hyperparameters and the description of the used optimizer.

   Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: We exclude error bars because it would be too computationally expensive.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
   - If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We have all details in the experiment setup.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We adhere to the NeurIPS Code of Ethics and preserve anonymity

Guidelines:

- The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We mention both potential societal impacts in the discussion.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: Our model and data do not pose such risks.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: We cite the corresponding paper when needed.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: We discuss the model architecture details, training details, dataset collection process in the paper.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: Our paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: Our paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: This work does not use LLMs.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.