RaVF: Learning Radar Velocity Fields via Spatial-Doppler Guidance

Shengpeng Wang [1*] wsp666@hust.edu.cn
Lingzhen Li [1*] lingzhenli@hust.edu.cn
Chunshen Li [1] leeechsh@hust.edu.cn
Fei Xiao [2] feixiao@hubu.edu.cn
Wei Wang [3†] wangw@whu.edu.cn

[1]School of Electronic Information and Communications, Huazhong University of Science and Technology
[2]Business School, Hubei University
[3]School of Computer Science, Wuhan University

Abstract

Scene-level Velocity field estimation from 4D millimeter-wave radar is critical for robust perception but remains challenged by the inherent sparsity and noise of point-based paradigms. Furthermore, Doppler ambiguity induced by the Nyquist sampling limit fundamentally disrupts the continuity of velocity field learning. To address these challenges, we propose RaVF, a physically grounded framework that directly learns dense velocity fields from high-fidelity spatial-Doppler spectra. RaVF is built upon three key designs tailored to radar-specific signal distortions: 1) To mitigate range-dependent propagation attenuation and nonlinear angular spatial quantization, we devise a Physically-aware Encoder that calibrates spectral feature extraction within the native radar beamforming. 2) Observing that ego-induced radial motion explains static Doppler responses, we factor out static motion and guide dynamic flow estimation for physically consistent velocity field learning. 3) To enhance kinematic plausibility, we introduce an ego-conditioned bidirectional pyramidal flow module that transfers backward flow into the forward frame and enforces anti-symmetric forward-backward consistency. Extensive experiments demonstrate that RaVF significantly outperforms state-of-the-art baselines on velocity fields. Our code is released at https://github.com/MetaIoT-WHU/RaVF.

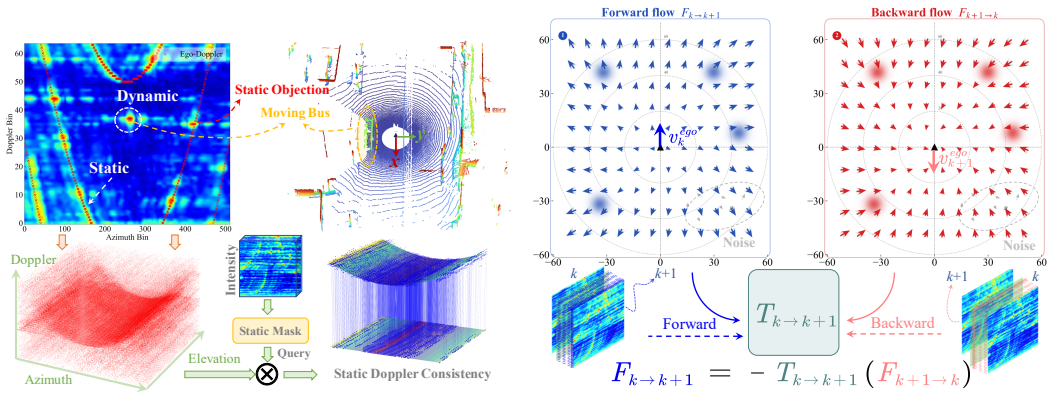
1 Introduction

Millimeter-wave radar is increasingly recognized as a robust sensing modality for autonomous perception [32, 29, 49, 17, 55, 50], owing to its resilience to poor illumination, fog, rain, and other challenging environmental conditions that often compromise cameras and LiDAR [8, 10, 54]. Beyond geometric perception [14, 27, 52], radar provides direct Doppler measurements that reveal radial motion cues of the surrounding scene, making it particularly attractive for motion-centric perception [16, 9]. Building dense velocity fields from radar signals can transform these local Doppler observations into coherent scene-level motion representations, supporting motion-aware perception beyond object detection alone. Accurate dense velocity estimation is thus a key component for Motion Estimation [7, 50, 51], Navigation [6, 35, 46], Object Detection [56, 36, 4], 3D Reconstruction/Generation [60, 61, 47], SLAM [51, 16, 57, 1], and Emobodied Spatial Intelligence [59, 26, 45] in safety-critical autonomous systems.

To achieve accurate velocity-field estimation, existing studies mainly learn motion through temporal variations of spectral energy [43], radar transformer [38], or solve motion completion from point-level

*Equal contribution.
†Corresponding author.

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



(a) (b)

Figure 1: **Our insights**: (a) *Folded ego-Doppler consistency.* Static radar returns preserve a folded ego-Doppler consistency under Nyquist-induced velocity folding: their Doppler responses agree with the wrapped radial projection of ego motion, while dynamic objects manifest as localized residual deviations. (b) *Ego-conditioned anti-symmetric bi-flow.* When the dense velocity field is viewed forward from frame k to frame $k+1$ and backward from frame $k+1$ to frame k, the two scene flows exhibit an ego-conditioned anti-symmetric structure. After transferring the backward flow into the forward frame by the ego-motion-conditioned operator $T_{k\to k+1}$, the forward flow should be consistent with its negative counterpart.

observations [44, 20, 21, 22]. However, these paradigms still faces three fundamental challenges. First, radar measurements suffer from range-dependent propagation attenuation and nonlinear angular resolution, which introduce spatially non-uniform uncertainty across the field of view [53]. As a result, far-range and oblique regions often provide weak, sparse, and noisy motion cues, easily causing velocity predictors to overfit reliable near-field observations or hallucinate physically inconsistent motions in uncertain areas. Second, Doppler measurements are inherently subject to velocity folding: the observed radial velocity can be written as $\tilde{v}_r = ((v_r + v_{\max}) \bmod 2v_{\max}) - v_{\max}$, where velocities differing by integer multiples of $2v_{\max}$ become observationally indistinguishable. This cyclic ambiguity corrupts motion magnitude and direction estimation, making temporal association and dense supervision unstable. Third, Doppler observations entangle ego-induced static motion with object-level dynamic motion. For static structures, the measured Doppler should be explainable by the ego-induced radial velocity, whereas dynamic objects manifest as localized residual deviations from this static consistency. This coupling makes dense velocity-field learning challenging: overly relying on global ego-motion suppresses dynamic object motion, while directly fitting local Doppler responses amplifies noise and folding artifacts.

To address these challenges, we mine two physics-induced consistency priors from real-world spatial-Doppler radar measurements and derive a physically grounded motivation for our framework. As shown in Fig. 1a, static radar returns preserve a folded ego-Doppler consistency even under Nyquist-induced velocity folding: their observed Doppler remains aligned with the wrapped line-of-sight projection of ego motion, while dynamic objects appear as localized residual violations of this consistency. This observation suggests a physics-guided factorization strategy that the readily observable static velocity field can be explained by ego-Doppler geometric consistency, which suppresses the dominant static response and exposes dynamic residual cues for more reliable dense flow learning. Additionally, as illustrated in Fig. 1b, dense scene flow across consecutive radar frames obeys an ego-conditioned anti-symmetric structure. The forward flow $F_{k\to k+1}$ should be consistent with the negative backward flow $F_{k+1\to k}$ after being transferred into the forward frame by the ego-motion-conditioned operator $T_{k\to k+1}$:$F_{k\to k+1} = -T_{k\to k+1}(F_{k+1\to k})$. This relation turns bidirectional flow estimation into a physical regularizer, enabling reliable motion evidence to be cross-validated across temporal directions and reducing implausible predictions in regions with weak, noisy, or ambiguous spectral measurements.

To translate these physics-induced observations into a practical velocity-field learner, we propose *RaVF*, a radar-native framework for scene-level velocity-field estimation from spatial-Doppler spectra. RaVF progressively converts radar measurement physics into learnable constraints through three

key designs. First, a physically aware spectral encoder compensates propagation attenuation and beamforming-induced angular bias via range-adaptive reweighting and coordinate-conditioned encoding. Second, occupancy-guided Doppler reasoning suppresses unreliable empty-space Doppler, refines ego motion, and unfolds folded radial velocities, thereby disentangling the ego-induced static field from dynamic residual motion. Third, an ego-conditioned bidirectional pyramidal flow module enforces anti-symmetric forward-backward consistency, stabilizing velocity prediction under noisy, weak, and Doppler-ambiguous measurements. Experiments on real-world 4D radar datasets validate the effectiveness of RaVF for radar-centric velocity-field estimation.

In summary, our contributions are fourfold: 1) We introduce *RaVF*, a physically grounded framework for learning radar velocity fields directly from spatial-Doppler spectra by disentangling ego-motion-induced static flow from independent object motion. 2) We develop an ego-conditioned bidirectional pyramidal flow with anti-symmetric consistency, enabling robust scene flow estimation under noisy, sparse, and Doppler-ambiguous observations. 3) We design a physics-aware spectral encoder that accounts for range-dependent attenuation and non-uniform angular resolution, improving feature reliability across challenging radar regions. 4) Extensive experiments on real-world 4D radar datasets demonstrate consistent gains in velocity estimation, occupancy prediction, and dynamic-motion reasoning.

2 Related Work

Radar motion estimation from point-level Doppler. Millimeter-wave radar naturally provides radial velocity measurements, which have been widely exploited for ego-motion estimation [7, 8, 18], moving object segmentation [69, 37, 24], and motion-aware perception [50, 39, 9]. Classical methods estimate ego velocity by fitting the relationship between radial Doppler velocity and azimuth angle, often relying on static-scene assumptions or robust geometric optimization [21, 22]. Registration-based methods such as Doppler-aware ICP variants [20, 48] further incorporate point-level geometry and radial velocity cues to improve motion estimation. More recently, learning-based methods [67, 68, 40] directly operate on sparse radar point clouds. Radar Velocity Transformer learns moving object segmentation from noisy single-scan radar detections by embedding Doppler velocity into transformer features [62, 63]. RadarEVE [38] introduces a spatial-temporal transformer for radar-only moving object segmentation and ego-velocity estimation, exploiting sparse radar Doppler points across consecutive frames [38]. These methods demonstrate the value of Doppler for motion reasoning, but they remain fundamentally tied to sparse point-level detections. Consequently, their motion estimates are sensitive to detection quality, spatial sparsity, and unreliable Doppler measurements in weak or ambiguous radar returns.

Radar velocity-field learning. Velocity-field estimation aims to recover a 3D motion field across consecutive frames [34, 66]. Foundational point-level scene flow methods have established essential paradigms, including learning-based approaches [28, 64] that infer point-wise displacement from local geometric features or neural motion representations [33, 65, 31], and registration-based approaches [30, 2] that estimate motion through geometric alignment, correspondence search, and iterative optimization. However, these methods are primarily designed for geometrically informative point clouds and do not explicitly account for radar-specific sensing properties. Recent radar motion learning methods extend scene-flow estimation to millimeter-wave radar point clouds. MilliFlow estimates point-wise motion from mmWave radar for human motion sensing [13]. learn radar scene flow from sparse 4D radar detections with self-supervision or cross-modal supervision [11, 15, 12]. More recently, RadarMP moves beyond point detections by learning motion from low-level radar tensors and spectral energy variations [43]. Despite these advances, existing methods mainly formulate radar motion as point-level displacement estimation, cross-modal pseudo-label learning, or spectral energy transport. They do not explicitly model scene-level radar velocity fields in the spatial-Doppler measurement space, nor do they address fundamental radar limitations such as Doppler folding as a physical constraint.

3 Method

Given two consecutive radar frames k and $k+1$, our goal is to estimate a scene-level radar velocity field $\mathbf{F}_{k\to k+1}$ and the corresponding occupancy field $\mathcal{O}$. We interpret the velocity field as the

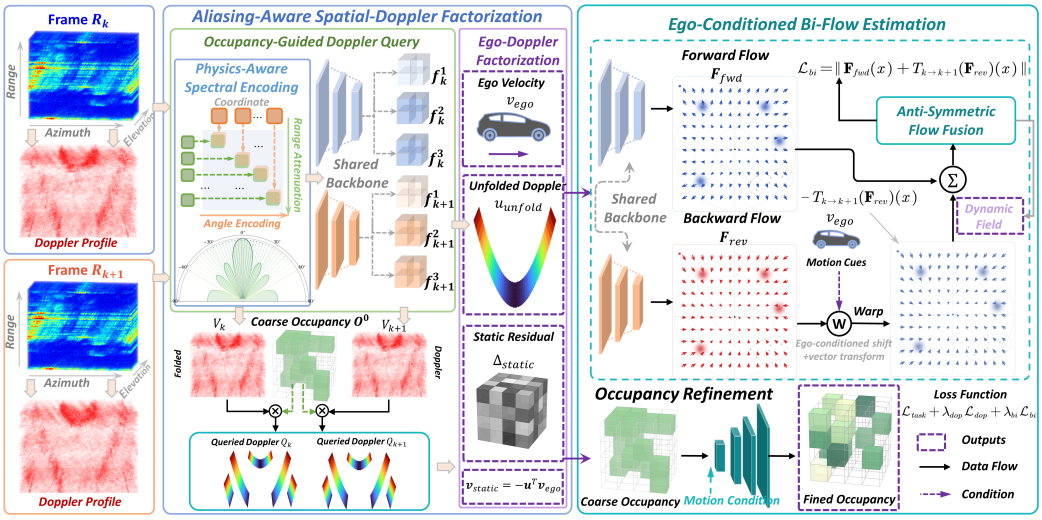


Figure 2: An overview of our RaVF pipeline for learning radar velocity fields.

composition of an ego-induced static component under the ego velocity $\mathbf{v}_{ego}$, and a residual dynamic component caused by independently moving objects. Let $\mathcal{R}_t \in \mathbb{R}^{D\times H\times W}$ denote the radar spectrum at frame t, where D, H, W correspond to range, azimuth, and elevation. Let $\tilde{\mathcal{V}}_t$ denote the folded Doppler observation, whose valid elements are bounded by $v_{\max}$. Since $\tilde{\mathcal{V}}_t$ is derived from Doppler-bin selection, it is physically reliable only at occupied radar responses indicated by $\mathcal{O}$.

RaVF converts this radar measurement physics into two learnable constraints as illustrated in Fig. 2. First, it performs *aliasing-aware spatial-Doppler factorization*: the dominant static field is explained by ego-Doppler geometric consistency, while dynamic motion is exposed as residual violations after Doppler unfolding. Second, it performs *ego-conditioned bi-flow regularization*: forward and backward velocity fields are estimated jointly and cross-validated through an anti-symmetric relation after ego-conditioned transfer. These two components allow RaVF to reason over folded, sparse, and noisy Doppler measurements while preserving physically plausible velocity fields.

3.1 Aliasing-Aware Spatial-Doppler Factorization

Physics-Aware Spectral Encoding. To reduce spatially non-uniform uncertainty before motion inference (*Details in Supplementary Materials*), we first adapt radar features to propagation and beamforming effects. According to the monostatic radar equation [25], the received echo power decays approximately as $P_r \propto R^{-4}$, making far-range responses substantially weaker than near-field returns. Meanwhile, for a uniform linear array, the inter-element phase progression is proportional to $\sin\theta$, so beamforming-based angular estimation maps spatial frequency to physical angle through an $\arcsin(\cdot)$ relation [19]. This nonlinear angular sampling induces non-uniform spatial resolution across the field of view. Rather than explicitly inverting these sensing effects, RaVF exposes them to the encoder through range-adaptive reweighting and coordinate-conditioned feature encoding. Specifically, For each frame $t \in \{k, k+1\}$, the encoded input $\mathbf{X}_t$ and multi-scale radar features are defined as

$$\mathbf{X}_t = \Pi_{\text{phys}}(\mathcal{R}_t) = \text{Concat}\big(\gamma(r) \odot \mathcal{R}_t^0, \mathcal{R}_t^{1:}, \mathbf{c}(r,a,e)\big), \quad \{\mathbf{f}_t^1, \mathbf{f}_t^2, \mathbf{f}_t^3\} = E_\theta(\mathbf{X}_t), \tag{1}$$

where $\mathcal{R}_t^0$ denotes the primary radar spectral response, $\mathcal{R}_t^{1:}$ denotes optional auxiliary input channels, $\gamma(r)$ is a learnable range-adaptive gain shared over azimuth and elevation, $\mathbf{c}(r,a,e)$ denotes normalized range-azimuth-elevation coordinate channels, and E_θ is a shared 3D residual encoder. This design does not assume a hand-crafted inverse radar model; instead, it provides the network with physically structured cues to compensate range-dependent attenuation and learn position-dependent biases induced by beamforming geometry, producing radar-native features that preserve local spectral evidence while remaining aware of global measurement location.

Occupancy-Guided Doppler Query. Since Doppler responses outside occupied radar regions are unreliable, RaVF predicts a coarse occupancy prior before using Doppler. We first estimate a geometric ego cue $\mathbf{v}_{geom}$ from the coarsest paired features and encode it into an ego token $\mathbf{z}_{ego}$. The occupancy decoder uses paired features, inter-frame feature differences, and the ego token to predict $\mathcal{O}^0$. Then, instead of directly feeding dense folded Doppler, we query Doppler only through occupancy support:

$$\mathcal{O}^0 = \sigma(H_{occ}(\mathcal{F}_{k,k+1}, \mathbf{z}_{ego})), \bar{\mathcal{V}}_{occ} = \frac{\mathcal{A}(\mathcal{O}^0 \odot \tilde{\mathcal{V}})}{\mathcal{A}(\mathcal{O}^0) + \epsilon}, \mathcal{Q}_{dop} = \mathcal{O}^0 \odot \left[G \odot \tilde{\mathcal{V}} + (1 - G) \odot \bar{\mathcal{V}}_{occ} \right]. \tag{2}$$

Here, H_{occ} denotes an ego-conditioned occupancy decoder that aggregates paired multi-scale spectral features and their inter-frame discrepancies $\mathcal{F}_{k,k+1} = \{\mathbf{f}_k^s, \mathbf{f}_{k+1}^s, |\mathbf{f}_k^s - \mathbf{f}_{k+1}^s|\}_{s=1}^3$ to estimate the coarse occupancy prior $\mathcal{O}^0$. $G \in [0, 1]$ is a learned Doppler gate, $\mathcal{A}(\cdot)$ denotes local 3D average aggregation over a small neighborhood, $\bar{\mathcal{V}}_{occ}$ is a local occupied-Doppler prior, and $\mathcal{Q}_{dop}$ is the queried folded Doppler. This design suppresses invalid empty-space Doppler while preserving reliable occupied radial evidence.

Ego-Doppler Factorization. The key physical prior is that static radar returns should be explainable by ego-induced radial motion. RaVF first predicts a coarse ego velocity $\mathbf{v}_{geom}$ from paired high-level spectral features, which provides a geometry-conditioned motion prior before Doppler reasoning. Let $\mathbf{u}(x)$ denote the line-of-sight unit vector at voxel $x = (r, a, e)$. For a static return, the ego-induced radial velocity under $\mathbf{v}_{geom}$ is given by $-\mathbf{u}(x)^\top \mathbf{v}_{geom}$. Since Doppler measurements are folded by the Nyquist limit, RaVF does not treat the queried Doppler $\mathcal{Q}_{dop}$ as a direct dense velocity label. Instead, it introduces a differentiable aliasing inference module that models the Doppler cycle as a latent variable. For the m-th aliasing cycle, the candidate radial velocity is $v^{(m)}(x) = \mathcal{Q}_{dop}(x) + 2mv_{\max}$, where $m \in [-M, M]$. We define the ego-Doppler compatibility residual as $\varepsilon_m(x) = |v^{(m)}(x) + \mathbf{u}(x)^\top \mathbf{v}_{geom}|$, which measures how well each unfolded candidate agrees with the coarse ego-induced static motion. Therefore, the unfolded Doppler can be expressed as

$$v_{unfold}(x) = \sum_{m=-M}^{M} p_m(x) v^{(m)}(x) + \delta_v(x), \quad p_m(x) = \mathrm{Softmax}_m(-\varepsilon_m(x)/\tau + \ell_m(x)). \tag{3}$$

Here, $\mathrm{Softmax}_m(\cdot)$ normalizes over aliasing cycles, $\ell_m(x)$ denotes learned cycle evidence predicted from spectral-Doppler features, τ controls the strength of the geometry-conditioned prior, and $\delta_v(x)$ is a bounded in-cycle residual. This neural cycle inference converts folded Doppler into a soft, ego-consistent radial representation while preserving uncertainty over ambiguous velocity aliases.

The inferred radial evidence is further used to refine ego motion. RaVF aggregates reliability-weighted unfolded Doppler features into a Doppler-driven ego proposal $\mathbf{v}_{dop}$, and adaptively fuses it with the coarse estimate as $\mathbf{v}_{ego} = \alpha \mathbf{v}_{geom} + (1 - \alpha)\mathbf{v}_{dop}$, where α is a learned fusion gate. With the refined ego velocity, we compute the static radial field $v^*_{static}(x) = -\mathbf{u}(x)^\top \mathbf{v}_{ego}$ and the residual static inconsistency $\Delta_{static}(x) = |v_{unfold}(x) - v^*_{static}(x)|$. Static responses should be absorbed by the ego-consistent radial field, whereas independently moving objects remain as localized residual deviations. Therefore, RaVF predicts dynamic probability through an occupancy-constrained residual-motion head:

$$\mathcal{D}(x) = \mathrm{Sigmoid}\big(H_{dyn}\big(\mathbf{f}^1(x), \mathcal{O}^0(x), v_{unfold}(x), v^*_{static}(x), \Delta_{static}(x)\big)\big) \odot \mathcal{O}^0(x). \tag{4}$$

Here, H_{dyn} denotes the dynamic-motion prediction head and the multiplication by $\mathcal{O}^0(x)$ restricts dynamic reasoning to physically supported occupied regions. In this way, the dominant static response is explained by ego-induced radial motion, making residual dynamic cues more separable for subsequent velocity-field estimation.

3.2 Ego-Conditioned Bi-Flow Regularization

Motion-Conditioned Pyramidal Flow. After spatial-Doppler factorization, RaVF estimates velocity fields using motion-conditioned pyramidal features. For each scale s, the feature conditioner produces $\hat{\mathbf{f}}_t^s = \Phi_s(\mathbf{f}_t^s, \mathcal{O}^0, \mathcal{D}_t, \Delta_{static,t}, v_{unfold,t}, \mathbf{z}_{ego})$, where Φ_s injects occupancy, dynamic probability, unfolded Doppler, static residual, and the refined ego token into the spectral feature. This

conditioning enables the flow estimator to adaptively exploit different motion evidence: static regions are anchored by ego-Doppler consistency, while dynamic regions are emphasized through residual violations. The bidirectional velocity fields are then estimated by a shared pyramidal cost-volume network:

$$\mathbf{F}_{fwd} = \mathcal{P}(\{\hat{\mathbf{f}}_k^s\}_{s=1}^3, \{\hat{\mathbf{f}}_{k+1}^s\}_{s=1}^3), \quad \mathbf{F}_{rev} = \mathcal{P}(\{\hat{\mathbf{f}}_{k+1}^s\}_{s=1}^3, \{\hat{\mathbf{f}}_k^s\}_{s=1}^3). \tag{5}$$

Here, $\mathcal{P}$ denotes the coarse-to-fine pyramidal flow estimator with local 3D cost volumes and feature warping. Estimating both temporal directions provides two complementary views of the same physical motion field, enabling subsequent cross-directional consistency regularization.

Ego-Conditioned Flow Transfer. Forward and backward velocity fields are estimated from opposite temporal references, so they cannot be directly compared in the same voxel coordinate system. Instead of assuming an exact closed-form rigid transformation on the discretized radar grid, RaVF parameterizes a learnable ego-conditioned transfer operator $T_{k\to k+1}$ to map the backward flow into the forward frame. Specifically, H_s and H_A are lightweight MLP heads conditioned on the refined ego velocity $\mathbf{v}_{ego}$, where H_s predicts a global grid shift $\mathbf{s}_{ego} = H_s(\mathbf{v}_{ego})$ for spatial re-indexing, and H_A predicts a ζ bounded residual vector transform $\Delta\mathbf{A}_{ego} = \zeta \tanh(H_A(\mathbf{v}_{ego}))$, yielding $\mathbf{A}_{ego} = \mathbf{I} + \Delta\mathbf{A}_{ego}$. The transferred backward flow is then defined as

$$\bar{\mathbf{F}}_{rev} = T_{k\to k+1}(\mathbf{F}_{rev}) = \mathbf{A}_{ego}\, \mathrm{Warp}(\mathbf{F}_{rev}, \mathbf{s}_{ego}). \tag{6}$$

Here, $\mathrm{Warp}(\cdot, \mathbf{s}_{ego})$ aligns the spatial support of the backward flow to the forward frame, while $\mathbf{A}_{ego}$ adjusts the vector components under ego-conditioned motion. Thus, $T_{k\to k+1}$ should be understood as a learnable approximation to frame transfer on radar voxel grids, rather than a strict SE(3) transformation. This transferred flow provides a common reference for evaluating the anti-symmetric relation between forward and backward velocity fields.

Anti-Symmetric Flow Fusion. The second physical prior is that the forward velocity field should agree with the negative transferred backward field in a common ego-conditioned reference frame. Due to radar noise, voxel discretization, occlusion, and Doppler ambiguity, we use this relation as a soft constraint rather than a hard equality. Specifically, RaVF measures the anti-symmetric discrepancy by $e_{anti} = \|\mathbf{F}_{fwd} + \bar{\mathbf{F}}_{rev}\|_2$, and feeds it together with occupancy, dynamic probability, and Doppler residual cues into a lightweight fusion head. The final velocity field is obtained by adaptively combining the forward prediction and the transferred backward prediction:

$$\mathbf{F} = \beta \odot \mathbf{F}_{fwd} + (1-\beta) \odot (-\bar{\mathbf{F}}_{rev}) + \Delta\mathbf{F}. \tag{7}$$

Here, the gate β estimates the local reliability of the two temporal directions, while $\Delta\mathbf{F}$ provides a residual correction for imperfect transfer and radar-specific uncertainty. This design allows reliable motion evidence to be cross-validated across temporal directions, thereby suppressing implausible velocity predictions in weak or ambiguous regions.

Finally, RaVF refines occupancy using the estimated motion field. Instead of relying only on spectral appearance, the refinement decoder integrates paired fine-scale features, inter-frame feature discrepancy, coarse occupancy logits, dynamic probability, Doppler residual, and flow cues.

$$\mathcal{O} = \sigma\big(\mathbf{Z}_{occ}^0 + H_{ref}(\mathbf{f}_k^1, \mathbf{f}_{k+1}^1, |\mathbf{f}_k^1 - \mathbf{f}_{k+1}^1|, \mathbf{Z}_{occ}^0, \mathcal{D}, \Delta_{static}, \mathbf{F})\big), \quad \mathbf{F}_{out} = \mathrm{Up}(\mathbf{F}). \tag{8}$$

Here, H_{ref} denotes the motion-aware occupancy refinement head, and $\mathbf{Z}_{occ}^0$ is the coarse occupancy logit. This refinement closes the loop between spectral evidence, Doppler residuals, velocity estimation, and occupancy prediction, improving the consistency of the final scene-level outputs.

3.3 Loss Function

RaVF is optimized with a compact multi-task objective that combines standard task supervision with two physics-induced regularizers:

$$\mathcal{L} = \mathcal{L}_{task} + \lambda_{dop}\mathcal{L}_{dop} + \lambda_{bi}\mathcal{L}_{bi}. \tag{9}$$

Here, $\mathcal{L}_{task}$ includes supervised ego-velocity regression, occupancy prediction, and dynamic velocity-field estimation. Specifically, ego velocity is trained with Smooth-ℓ_1, occupancy and dynamic probability are trained with class-balanced BCE/Dice losses, and velocity-field supervision uses a dynamic-aware robust regression loss to emphasize sparse moving regions while weakly regularizing static occupied voxels.

The Doppler consistency loss implements our folded ego-Doppler observation. It encourages the unfolded radial velocity to reconstruct the folded Doppler observation, while enforcing static occupied voxels to be explained by ego-induced radial motion:

$$\mathcal{L}_{dop} = \mathbb{E}_{x\in\mathcal{O}^{gt}}[|\,\mathrm{wrap}(v_{unfold}(x)) - \tilde{v}_r(x)|] + \mathbb{E}_{x\in\mathcal{S}^{gt}}[|\,\mathrm{wrap}(v^*_{static}(x)) - \tilde{v}_r(x)|]\,, \tag{10}$$

where $\mathcal{O}^{gt}$ denotes occupied voxels, $\mathcal{S}^{gt} = \mathcal{O}^{gt}(1 - \mathcal{D}^{gt})$ denotes static occupied voxels, and $\mathcal{D}^{gt}$ refers to the dynamic probability. This loss preserves compatibility with folded Doppler measurements while encouraging ego-consistent static-motion explanation.

The bi-flow loss implements our anti-symmetric motion prior. After transferring the backward flow into the forward frame, the two directions should cancel under temporal correspondence:

$$\mathcal{L}_{bi} = \mathbb{E}_{x\in\mathcal{O}^{gt}}\left[\|\mathbf{F}_{fwd}(x) + T_{k\to k+1}(\mathbf{F}_{rev})(x)\|_2\right]. \tag{11}$$

Together, $\mathcal{L}_{dop}$ and $\mathcal{L}_{bi}$ convert folded ego-Doppler consistency and ego-conditioned anti-symmetric bi-flow into learnable constraints, improving velocity-field estimation in weak, noisy, and Doppler-ambiguous radar regions. More implementation details are provided in the Supplementary Materials.

4 Experiments Setup

4.1 Datasets

We evaluate *RaVF* on two large-scale real-world 4D radar datasets, **RaDelft** [42] and **K-Radar** [36], which provide ADC data or FFT-processed radar tensors, together with synchronized LiDAR supervision for scene-level velocity-field and ego-motion evaluation.

1. **RaDelft Dataset** spans suburban roads, university campuses, old-town streets, and bridge infrastructures in Delft. It provides 16,975 synchronized radar frames at 10 Hz from a TI MMWCAS-RF-EVM 2243 imaging radar, together with a RoboSense Ruby Plus 128-beam LiDAR, a front-view camera, and high-rate odometry. We train/validate on Scenes 1, 2, 3, 6, and 7, and reserve Scenes 4 and 5 for scene-disjoint testing.
2. **K-Radar Dataset** spans urban roads, suburban roads, alleyways, and highways across day/night and adverse-weather conditions, including fog, rain, sleet, and snow. It provides 35K 4D radar tensor frames, amounting to roughly 12 TB of 4DRT data, along with 64-/128-channel LiDARs, surround stereo cameras, RTK-GPS, and IMUs. Detailed dataset splits and preprocessing protocols are provided in the supplementary material.

For point-level baselines on RaDelft, we apply conventional Doppler velocity compensation, whereas on K-Radar we use the provided point-level Doppler directly because raw ADC data are unavailable, and the Doppler values are folded.

4.2 Radar Point Cloud Enhancement

Baselines. To evaluate the efficacy of our *RaVF* point cloud enhancement module, we perform comparative experiments on the K-Radar dataset against three representative baselines: the traditional statistical **CFAR** algorithm [5], the deep learning-based **RPDNet** [10], and the high-resolution enhancement framework **RadarHD** [41].

Metrics. We assess the generated point cloud quality using **Chamfer Distance (CD) and F1-Score (FS)**. CD measures overall geometric alignment via the average nearest-neighbor Euclidean distance, with lower values indicating better performance. Conversely, FS computes the harmonic mean of precision and recall at a set threshold to robustly evaluate local detail reconstruction against outliers.

Results. As shown in Tab. 1, *RaVF* consistently achieves the lowest CD and the highest FS across all evaluated sequences, demonstrating clear superiority in radar occupancy estimation. Compared with the strongest baseline, RadarHD, *RaVF* improves the average CD/FS from 1.14/0.70 to 0.78/0.80, indicating substantially better geometric fidelity and occupancy completeness. The performance gain remains stable on more challenging sequences, such as Seq18, Seq24, and Seq26, which suggests stronger robustness to sparse, noisy, and incomplete radar observations. Qualitative results in Fig. 3 further confirm that *RaVF* recovers more complete object structures and less noisy clusters. In particular, the refined occupancy prediction better preserves fine-grained structures around vehicles and vulnerable road users, producing point clouds that are visually the most faithful to the LiDAR reference.

Detection Methods	Seq2		Seq4		Seq6		Seq10		Seq12		Seq18		Seq22		Seq24		Seq26	
	CD ↓	FS ↑	CD ↓	FS ↑	CD ↓	FS ↑	CD ↓	FS ↑	CD ↓	FS ↑	CD ↓	FS ↑	CD ↓	FS ↑	CD ↓	FS ↑	CD ↓	FS ↑
OS-CFAR [5]	2.56	0.13	2.83	0.10	2.53	0.06	2.60	0.13	2.40	0.08	2.35	0.06	2.20	0.17	2.59	0.14	2.20	0.05
RPDNet [10]	2.33	0.15	2.45	0.11	2.18	0.10	2.33	0.14	2.16	0.10	1.99	0.08	1.91	0.20	2.42	0.15	1.74	0.11
RadarHD [41]	1.23	0.76	0.83	0.77	1.14	0.54	0.64	0.82	0.73	0.84	1.49	0.40	1.08	0.74	1.75	0.72	1.34	0.68
RaVF (Ours)	0.80	0.84	0.67	0.85	0.71	0.73	0.53	0.85	0.45	0.90	0.87	0.63	0.77	0.83	1.24	0.81	0.94	0.79

Table 1: Comparison of spatial detection performance across different scenes.

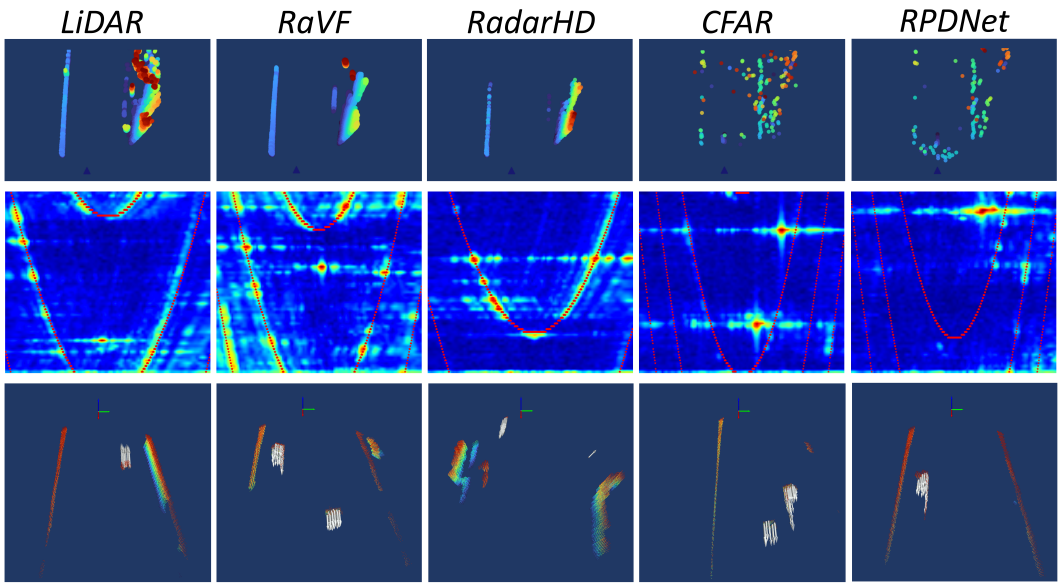


Figure 3: Qualitative results including the occupancy, static and dynamic field.

4.3 Static Field Estimation

Baselines. To rigorously evaluate the performance of our static field estimation module, we benchmark our approach against a comprehensive spectrum of fundamental and robust estimators traditionally employed for point cloud registration and ego-velocity estimation. Specifically, our selected baselines include **ICP** [3], **GICP** [23], **DICP** [20], **RANSAC** [21], **TEASER++** [58], and the learning-based **RadarEVE** [38].

Results. Tab. 2 demonstrates that *RaVF* provides the most accurate ego-induced static-field estimation on both RaDelft and K-Radar. The full model achieves the lowest average error, reducing it to 0.24 on RaDelft and 0.12 on K-Radar, outperforming the strongest baselines by a clear margin. This advantage is consistent across speed ranges and distance intervals, suggesting that dense spatial-Doppler reasoning is more reliable than sparse point-level registration for radar motion estimation. Ablation studies show that both Physics-Aware Spectral Encoding, *w.r.t.* P. and Ego-Conditioned Bi-Flow Model *w.r.t.* E. contribute to the final performance, with their combination yielding the most stable results. As visualized in the second row of Fig. 3, the predicted static motion produces a Doppler distribution that closely matches the measured one, qualitatively validating the Doppler consistency of the learned static velocity field.

4.4 Dynamic Field Estimation

Baselines. Our perception capabilities for dynamic field estimation are validated against several established scene flow estimation architectures, **FastNSF** [28], **VoteFlow** [31], and **ICPFlow** [30].

Metrics. We evaluate dynamic scene flow using five metrics: 1) EPE_{dyn} (m/s), the mean Euclidean distance between predicted and ground-truth dynamic flows; 2) CosSim, measuring their directional alignment; 3) Strict Accuracy (AccS, %), the percentage of points satisfying $\text{EPE}_{\text{dyn}} < 1.0$ m/s or relative error $< 10\%$; 4) Relaxed Accuracy (AccR, %), relaxing the thresholds to 2.0 m/s or 20%;

Dataset	Point Set	Methods	Scene 4				Scene 5				Avg.
			[0,4)	[4,8)	[8,12)	[12,16]	[0,4)	[4,8)	[8,12)	[12,16]	
RaDelft	CFAR	ICP [3]	0.24	0.49	0.65	3.71	0.62	0.57	1.45	2.60	1.33
		GICP [23]	0.55	1.13	2.25	6.24	0.98	1.82	3.85	4.19	2.44
		DICP [20]	0.28	0.20	0.84	1.95	0.27	0.15	0.62	0.94	0.64
		RANSAC [21]	0.46	0.16	0.62	1.00	0.20	0.18	0.49	1.08	0.67
		TEASER++ [58]	0.27	0.48	0.52	0.60	0.64	0.95	0.61	0.71	0.52
		RadarEVE [38]	0.16	0.28	0.42	0.48	0.22	0.27	0.24	**0.35**	0.30
	RPDNet	ICP [3]	0.30	0.47	0.61	1.08	0.39	0.36	0.68	0.94	0.63
		GICP [23]	0.68	0.86	1.30	1.74	0.62	0.63	1.19	1.50	1.12
		TEASER++ [58]	0.62	0.61	0.50	0.74	0.66	0.40	0.46	0.61	0.61
	Ours	w/o P. & E.	**0.14**	0.17	0.23	0.13	0.28	0.16	0.51	0.92	0.34
		w/o E.	0.21	0.22	0.37	0.16	0.28	0.20	0.61	1.29	0.45
		w/o P.	0.29	**0.12**	0.36	0.26	0.20	0.16	0.37	0.62	0.32
		w P. & E.	0.15	0.14	**0.11**	**0.06**	**0.13**	**0.14**	**0.13**	0.42	**0.24**

Dataset	Point Set	Methods	Distance Intervals (m)								Avg.
			[0,5)	[5,10)	[10,15)	[15,20)	[20,25)	[25,30)	[30,35)	[35,40]	
K-Radar	CFAR	ICP [3]	1.29	1.60	4.06	7.36	12.49	19.09	18.29	20.24	10.55
		GICP [23]	1.76	2.44	5.39	9.52	15.65	22.02	19.99	19.01	11.97
		TEASER++ [58]	1.08	2.37	5.33	13.43	19.22	23.74	29.80	33.74	16.09
		RadarEVE [38]	4.99	4.44	3.93	4.02	3.72	4.08	5.37	7.68	4.78
	RPDNet	ICP [3]	1.73	1.49	4.21	7.96	12.01	19.51	20.12	16.67	10.46
		GICP [23]	2.18	2.30	5.08	9.15	14.86	21.49	22.47	16.92	11.81
		TEASER++ [58]	2.90	2.15	4.34	12.11	21.21	24.46	24.49	30.09	15.22
	Ours	w/o P. & E.	0.28	0.28	0.31	0.29	0.28	0.31	0.63	0.69	0.31
		w/o E.	0.27	0.27	0.29	0.29	0.26	0.29	0.61	0.74	0.30
		w/o P.	0.18	0.18	0.20	0.19	0.19	0.19	0.21	0.20	0.19
		w P. & E.	**0.07**	**0.08**	**0.15**	**0.16**	**0.17**	**0.17**	**0.18**	**0.16**	**0.12**

Table 2: Quantitative comparison of ego-velocity estimation (MAE ↓, m/s).

Flow Model	Point	K-Radar					RaDelft				
		EPE_{dyn} ↓	CosSim↑	AccS↑	AccR↑	OutlierR↓	EPE_{dyn} ↓	CosSim↑	AccS↑	AccR↑	OutlierR↓
FastNSF [28]	CFAR	0.79	0.66	77.72	91.50	46.94	7.91	0.74	17.83	43.77	62.48
	RPDNet	0.70	0.72	82.49	93.40	43.12	7.25	0.67	20.82	47.35	58.73
	RaVF	0.96	0.84	63.95	95.51	85.53	4.58	0.77	38.81	63.27	40.63
VoteFlow [31]	CFAR	0.30	0.88	89.93	98.84	28.10	15.23	0.35	1.81	8.36	96.46
	RPDNet	0.30	0.88	89.69	98.43	26.09	13.18	0.43	2.80	13.93	92.34
	RaVF	0.34	0.90	93.32	99.28	27.49	5.10	0.72	13.78	42.26	67.20
ICPFlow [30]	CFAR	0.35	0.89	85.04	97.30	34.32	18.63	-0.10	1.53	6.27	99.96
	RPDNet	0.34	0.92	86.92	97.62	32.45	18.28	-0.11	1.45	7.52	100.00
	RaVF	0.29	0.96	89.67	97.48	25.43	15.71	-0.05	4.34	12.60	98.99
RaVF (Ours)	w/o P.&E.	0.15	0.98	99.60	**100.00**	6.36	1.53	0.77	20.03	75.17	35.31
	w/o E.	0.14	**0.99**	99.69	**100.00**	8.91	1.47	0.81	24.14	78.70	31.67
	w/o P.	0.09	**0.99**	99.67	**100.00**	4.27	1.47	0.82	24.18	79.08	12.35
	w P.&E.	**0.06**	**0.99**	**99.80**	**100.00**	**3.93**	**0.79**	**0.98**	**89.87**	**95.79**	**9.83**

Table 3: Dynamic scene flow evaluation on the K-Radar and RaDelft datasets. Color intensity indicates per-column ranking, where darker denotes better performance.

and 5) Severe Outlier Rate (OutlierR, %), the fraction of extreme deviations ($EPE_{dyn} > 4.0$ m/s or relative error $> 40\%$).

Results. Tab. 3 shows that *RaVF* achieves the best dynamic velocity-field performance on both datasets, with consistent improvements across error, accuracy, and reliability metrics. Compared with the strongest task-specific baseline, *RaVF* reduces EPE_{dyn} by 79.3% and OutlierR by 84.5% on K-Radar, while improving AccS by 6.5 percentage points. On RaDelft, the advantage is more pronounced: *RaVF* reduces EPE_{dyn} by 82.8% and OutlierR by 75.8%, while improving CosSim by 27.3%, AccS by 51.1 percentage points, and AccR by 32.5 percentage points. These gains indicate that voxel-level spatial-Doppler reasoning is substantially more effective than point-level flow estimation for sparse and noisy radar observations. Qualitatively, Fig. 3 further shows that *RaVF* produces coherent object-level flow with consistent direction and magnitude.

4.5 Validation of Spatial-Doppler Guidance

We further examine bidirectional alignment and occupancy-supported Doppler querying on K-Radar, using the same frozen full RaVF checkpoint for both diagnostics without any adaptation.

Ego-conditioned bidirectional alignment. We compare three transfers before fusion. On the half-resolution grid, we compute $E(\mathcal{T}) = \mathrm{mean}_{x \in \Omega_{\mathrm{occ}}} \|F_{\mathrm{fwd}}(x) + \mathcal{T}(F_{\mathrm{rev}})(x)\|_2$. All operators act on identical predicted fields, probing compatibility within the jointly trained representation. Learned

Evaluation regime	Pairs	Consistency residual (m/s)↓			Reduction↑	
		Identity	SE(3)	Learned	vs. Identity	vs. SE(3)
All pairs	1,111	0.842	0.724	**0.344**	59.1%	52.4%
Large ego motion	278	1.535	1.286	**0.401**	73.9%	68.8%
High support turnover	278	0.983	0.808	**0.422**	57.0%	47.7%
High dynamic ratio	278	1.085	0.919	**0.425**	60.8%	53.7%

Table 4: Bidirectional alignment on 1,111 frame pairs. Residuals are averaged over occupied voxels within each pair and then across pairs. Stress groups are upper-quartile subsets with ties included and may overlap. Reductions use the corresponding reference operator within each row.

transfer yields the lowest mean residual in all four regimes. Its overall residual is 0.344 m/s, versus 0.842 for identity and 0.724 for odometry-based SE(3) (Table 4).

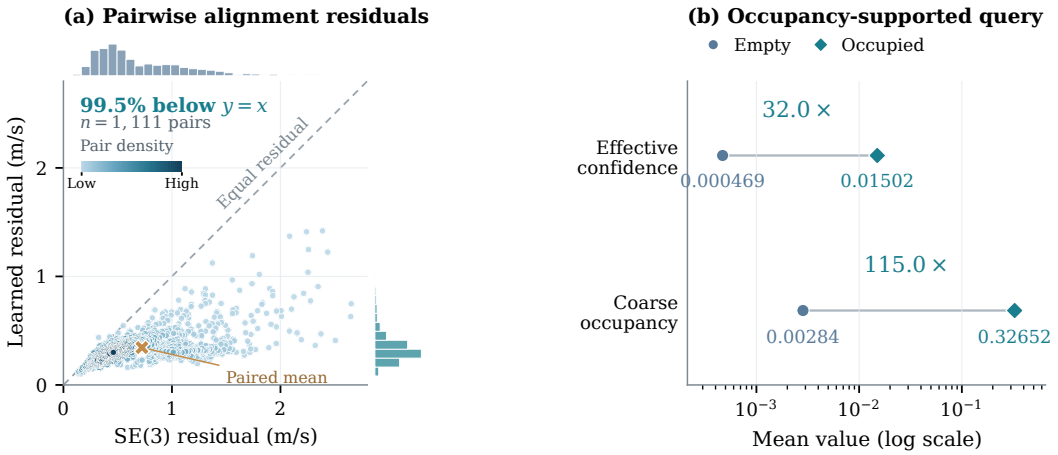


Figure 4: Spatial-Doppler validation on all 1,111 pairs. (a) Learned versus SE(3) residuals colored by pair density, with marginal histograms, an equality line (dashed), and the paired mean (cross). (b) Voxel-pooled confidence and coarse occupancy on LiDAR-defined occupied/empty support (logarithmic axis).

Alignment across motion regimes. Overall, learned transfer reduces the mean residual by 59.1% and 52.4% relative to identity and SE(3), respectively. Stress-group thresholds are 23.60 m/s for ego speed, 0.898 for occupancy turnover, and 1.23% for dynamic occupancy ratio. Turnover is $1 - \mathrm{IoU}(O_t, \mathrm{Warp}_{\mathrm{SE(3)}}(O_{t+1}))$. Per-pair residuals are lower than for SE(3) in 99.5% of the registered cases (Fig. 4(a)).

Spatial support for Doppler reasoning. Effective confidence in Eq. (2) is $c = \mathcal{A}(O^0) \odot G$. Its voxel-pooled mean is 0.01502 on occupied support and 0.000469 on empty support (32.0× contrast). Coarse occupancy yields a 115.0× contrast (Fig. 4(b)). These contrasts show that Doppler queries concentrate on observed scene structure.

5 Conclusion

We introduced *RaVF*, a physically grounded framework for radar velocity-field learning from spatial-Doppler spectra. RaVF explicitly decomposes scene motion into an ego-motion-induced static field and a residual dynamic field, enabling Doppler-consistent motion reasoning beyond sparse radial velocity regression. Through Ego-Conditional Bi-Flow fusion, RaVF robustly estimates both ego velocity and dynamic motion. Experiments on large-scale real-world 4D radar datasets demonstrate its effectiveness for velocity field estimation.

Acknowledgments and Disclosure of Funding

This work was supported in part by the Ministry of Education of China under grant number 24YJC630240, National Natural Science Foundation of China with Grant 62522214, 62471194, Key Research and Development Program of Hubei Province under grant number 2025BAB023, and the Fundamental Research Funds for the Central Universities.

References

[1] D. Adolfsson, M. Magnusson, A. Alhashimi, A. J. Lilienthal, and H. Andreasson. Cfear radarodometry-conservative filtering for efficient and accurate radar odometry. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5462–5469. IEEE, 2021.

[2] R. Ahuja, C. Baker, and W. Schwarting. Optflow: Fast optimization-based scene flow estimation without supervision. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3161–3170, 2024.

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

[4] O. Bialer and Y. Haitman. Radsimreal: Bridging the gap between synthetic and real data in radar object detection with simulation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15407–15416. IEEE, 2024.

[5] S. Blake. Os-cfar theory for multiple targets and nonuniform clutter. *IEEE transactions on aerospace and electronic systems*, 24(6):785–790, 1988.

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

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

[8] S. H. Cen and P. 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.

[9] Y. Chae, H. Park, H. Kim, and K.-J. Yoon. Doppler-aware lidar-radar fusion for weather-robust 3d detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 27197–27208, 2025.

[10] Y. Cheng, J. Su, M. Jiang, and Y. Liu. A novel radar point cloud generation method for robot environment perception. *IEEE Transactions on Robotics*, 38(6):3754–3773, 2022.

[11] F. Ding, Z. Pan, Y. Deng, J. Deng, and C. X. Lu. Self-supervised scene flow estimation with 4-d automotive radar. *IEEE Robotics and Automation Letters*, 7(3):8233–8240, 2022.

[12] F. Ding, A. Palffy, D. M. Gavrila, and C. X. Lu. Hidden gems: 4d radar scene flow learning using cross-modal supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9340–9349, 2023.

[13] F. Ding, Z. Luo, P. Zhao, and C. X. Lu. milliflow: Scene flow estimation on mmwave radar point cloud for human motion sensing. In *European Conference on Computer Vision*, pages 202–221. Springer, 2024.

[14] Z. Duan, X. Li, J. Pan, D. Xiaofei, H. Ning, and X. Song. Data-adaptive mahalanobis metric learning for cross-head attention in transformers, 2026. URL `https://arxiv.org/abs/2610.13261`.

[15] J. Fu, Z. Xiang, and N. Zhao. Raliflow: Scene flow estimation with 4d radar and lidar point clouds. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 4012–4021, 2026.

[16] P. Gao, S. Zhang, W. Wang, and C. X. 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.

[17] P. Gao, S. Zhang, W. Wang, and C. X. Lu. Robust metric localization in autonomous driving via doppler compensation with single-chip radar. *IEEE Transactions on Intelligent Transportation Systems*, 25(1): 491–502, 2023.

[18] K. Haggag, S. Lange, T. Pfeifer, and P. Protzel. A credible and robust approach to ego-motion estimation using an automotive radar. *IEEE Robotics and Automation Letters*, 7(3):6020–6027, 2022.

[19] K. Harlow, H. Jang, T. D. Barfoot, A. Kim, and C. Heckman. A new wave in robotics: Survey on recent mmwave radar applications in robotics. *IEEE Transactions on Robotics*, 40:4544–4560, 2024.

[20] B. Hexsel, H. Vhavle, and Y. Chen. DICP: Doppler Iterative Closest Point Algorithm. In *Proceedings of Robotics: Science and Systems*, New York City, NY, USA, June 2022. doi: 10.15607/RSS.2022.XVIII.015.

[21] D. Kellner, M. Barjenbruch, J. Klappstein, J. Dickmann, and K. Dietmayer. Instantaneous ego-motion estimation using doppler radar. In *16th International IEEE Conference on Intelligent Transportation Systems (ITSC 2013)*, pages 869–874. IEEE, 2013.

[22] D. Kellner, M. Barjenbruch, J. Klappstein, J. Dickmann, and K. Dietmayer. Instantaneous ego-motion estimation using multiple doppler radars. In *2014 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1592–1597. IEEE, 2014.

[23] K. Koide, M. Yokozuka, S. Oishi, and A. Banno. Voxelized gicp for fast and accurate 3d point cloud registration. In *2021 IEEE international conference on robotics and automation (ICRA)*, pages 11054–11059. IEEE, 2021.

[24] J. Kopp, D. Kellner, A. Piroli, V. Dallabetta, and K. Dietmayer. Simultaneous clutter detection and semantic segmentation of moving objects for automotive radar data. In *2023 IEEE 26th International Conference on Intelligent Transportation Systems (ITSC)*, pages 1078–1085. IEEE, 2023.

[25] P.-C. Kung, S. Harisha, R. Vasudevan, A. Eid, and K. A. Skinner. Radarsplat: Radar gaussian splatting for high-fidelity data synthesis and 3d reconstruction of autonomous driving scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 27596–27606, 2025.

[26] B. Li, Y. Shen, Y. Liu, Y. Xu, J. Liu, X. Li, Z. Li, J. Zhu, Y. Zhong, F. Lan, et al. Toward cognitive supersensing in multimodal large language model. *arXiv preprint arXiv:2602.01541*, 2026.

[27] C. Li, S. Wang, Z. Ye, and W. Wang. Mitigating radar-inertial calibration ambiguities via so (3) manifold steering. In *European Conference on Computer Vision*, pages 585–602. Springer, 2026.

[28] X. Li, J. Zheng, F. Ferroni, J. K. Pontes, and S. Lucey. Fast neural scene flow. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9878–9890, 2023.

[29] Y.-J. Li, S. Hunt, J. Park, M. O'Toole, and K. Kitani. Azimuth super-resolution for fmcw radar in autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17504–17513, 2023.

[30] Y. Lin and H. Caesar. Icp-flow: Lidar scene flow estimation with icp. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15501–15511, 2024.

[31] Y. Lin, S. Wang, L. Nan, J. Kooij, and H. Caesar. Voteflow: Enforcing local rigidity in self-supervised scene flow. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 17155–17164, 2025.

[32] Z. Lin, Z. Liu, Z. Xia, X. Wang, Y. Wang, S. Qi, Y. Dong, N. Dong, L. Zhang, and C. Zhu. Rcbevdet: Radar-camera fusion in bird's eye view for 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14928–14937, 2024.

[33] D. Liu, B. Zhuang, D. N. Metaxas, and M. Chandraker. Instantaneous perception of moving objects in 3d. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19573–19583, 2024.

[34] X. Liu, C. R. Qi, and L. J. Guibas. Flownet3d: Learning scene flow in 3d point clouds. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 529–537, 2019.

[35] Y. Liu, J. Zhu, Y. Mo, G. Li, X. Cao, J. Jin, Y. Shen, Z. Li, T. Yu, W. Yuan, et al. Palm: Progress-aware policy learning via affordance reasoning for long-horizon robotic manipulation. *arXiv preprint arXiv:2601.07060*, 2026.

[36] D.-H. Paek, S.-H. Kong, and K. T. Wijaya. K-radar: 4d radar object detection for autonomous driving in various weather conditions. *Advances in Neural Information Processing Systems*, 35:3819–3829, 2022.

[37] Z. Pan, F. Ding, H. Zhong, and C. X. Lu. Ratrack: moving object detection and tracking with 4d radar point cloud. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 4480–4487. IEEE, 2024.

[38] C. Pang, X. Chen, Y. Liu, H. Lu, and Y. Cheng. Radarmoseve: A spatial-temporal transformer network for radar-only moving object segmentation and ego-velocity estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 4424–4432, 2024.

[39] X. Peng, M. Tang, H. Sun, B. Kay, L. Servadei, and R. Wille. Elmar: Enhancing lidar detection with 4d radar motion awareness and cross-modal uncertainty. In *2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4939–4945. IEEE, 2025.

[40] X. Peng, Y. Wang, M. Tang, B. Kay, L. Servadei, and R. Wille. Moral: Motion-aware multi-frame 4d radar and lidar fusion for robust 3d object detection. *arXiv preprint arXiv:2505.09422*, 2025.

[41] A. Prabhakara, T. Jin, A. Das, G. Bhatt, L. Kumari, E. Soltanaghai, J. Bilmes, S. Kumar, and A. Rowe. High resolution point clouds from mmwave radar. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 4135–4142. IEEE, 2023.

[42] I. Roldan, A. Palffy, J. F. P. Kooij, D. M. Gavrila, F. Fioranelli, and A. Yarovoy. A deep automotive radar detector using the radelft dataset. *IEEE Transactions on Radar Systems*, 2:1062–1075, 2024. doi: 10.1109/TRS.2024.3485578.

[43] C. Ruiqi, D. Huijun, L. Jian, L. Feng, and L. Wei. Radarmp: Motion perception for 4d mmwave radar in autonomous driving. In *The Fortieth AAAI Conference on Artificial Intelligence (AAAI-26)*, pages 3282–3290, 2026.

[44] A. Segal, D. Haehnel, and S. Thrun. Generalized-icp. In *Robotics: science and systems*, volume 2, page 435. Seattle, WA, 2009.

[45] Y. Shen, B. Li, M. Huang, Y. Liu, X. Cao, J. Jin, Z. Li, A. Liu, J. Kim, J. Zhu, et al. Decoding children's gait behavior. *arXiv preprint arXiv:2608.00371*, 2026.

[46] Y. Shen, Y. Liu, J. Zhu, X. Cao, X. Zhang, Y. He, W. Ye, J. Rehg, and I. Lourentzou. Fine-grained preference optimization improves spatial reasoning in vlms. *Advances in Neural Information Processing Systems*, 38:17929–17960, 2026.

[47] Y. Shen, J. Xu, B. Li, Y. Zhang, T. Yu, B. Li, H. Yang, R. Wang, and X. Cao. Chronovision: Temporal reasoning via latent state reconstruction. *arXiv preprint arXiv:2608.05631*, 2026.

[48] D. Wang, D. C. Herraez, S. May, and A. Nüchter. Dynamic-icp: Doppler-aware iterative closest point registration for dynamic scenes. *IEEE Robotics and Automation Letters*, 2026.

[49] N. Wang, H. Dong, M. Cui, L. Qiu, J. Xiong, and W. Wang. Vesta: In-vehicle sensing with gnss satellite signals. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 10(3): 1–28, 2026.

[50] S. Wang, X. Luo, Y. Xie, and W. Wang. Sddiff: boosting radar perception via spatial-doppler diffusion. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, pages 8805–8813, 2025.

[51] S. Wang, Y. Xie, Q. Liao, and W. Wang. S3e: Self-supervised state estimation for radar-inertial system. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 26686–26695, 2025.

[52] S. Wang, K. Wang, and W. Wang. Rauf: Learning the spatial uncertainty field of radar. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 42474–42483, June 2026.

[53] S. Wang, K. Wang, and W. Wang. Rauf: Learning the spatial uncertainty field of radar. *arXiv preprint arXiv:2603.01026*, 2026.

[54] W. Wang, S. Wang, R. Zhang, and Z. Luo. Research on millimeter-wave radar-assisted multi-agent collaborative localization and mapping. *Chinese Journal on Internet of Things*, 8(3):66–75, 2024. doi: 10.11959/j.issn.2096-3750.2024.00416.

[55] W. Wang, C. Li, B. Zeng, L. Chen, L. Sun, K. Luo, and D. Chen. Mitigating interference for automotive millimeter-wave radar perception in dense traffic scenarios. *IEEE Transactions on Mobile Computing*, 2025.

[56] Y. Wang, Z. Jiang, X. Gao, J.-N. Hwang, G. Xing, and H. Liu. Rodnet: Radar object detection using cross-modal supervision. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 504–513, 2021.

[57] Y. Xu, Q. Huang, S. Shen, and H. Yin. Incorporating point uncertainty in radar slam. *IEEE Robotics and Automation Letters*, 10(3):2168–2175, 2025.

[58] H. Yang, J. Shi, and L. Carlone. Teaser: Fast and certifiable point cloud registration. *IEEE Transactions on Robotics*, 37(2):314–333, 2020.

[59] T. Yu, X. Li, Y. Shen, Y. Liu, and I. Lourentzou. Core3d: Collaborative reasoning as a foundation for 3d intelligence. *arXiv preprint arXiv:2512.12768*, 2025.

[60] T. Yu, X. Li, Y. Shen, Y. Shen, K. A. Nguyen, A. S. Juvekar, and I. Lourentzou. Silsa: Sliding-window slice latents for topology-preserving high-resolution 3d generation. *arXiv preprint arXiv:2610.02201*, 2026.

[61] T. Yu, X. Li, Y. Shen, O. Susladkar, Y. Liu, X. Zhou, and I. Lourentzou. Elsa3d: Elastic semantic anchoring for unified 3d understanding and generation. *arXiv preprint arXiv:2607.06565*, 2026.

[62] M. Zeller, V. S. Sandhu, B. Mersch, J. Behley, M. Heidingsfeld, and C. Stachniss. Radar instance transformer: Reliable moving instance segmentation in sparse radar point clouds. *IEEE Transactions on Robotics*, 40:2357–2372, 2023.

[63] M. Zeller, V. S. Sandhu, B. Mersch, J. Behley, M. Heidingsfeld, and C. Stachniss. Radar instance transformer: Reliable moving instance segmentation in sparse radar point clouds (abstract reprint). In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 39906–39906, 2026.

[64] Q. Zhang, Y. Yang, H. Fang, R. Geng, and P. Jensfelt. Deflow: Decoder of scene flow network in autonomous driving. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 2105–2111. IEEE, 2024.

[65] Q. Zhang, Y. Yang, P. Li, O. Andersson, and P. Jensfelt. Seflow: A self-supervised scene flow method in autonomous driving. In *European Conference on Computer Vision*, pages 353–369. Springer, 2024.

[66] Y. Zhang, J. Edstedt, B. Wandt, P.-E. Forssén, M. Magnusson, and M. Felsberg. Gmsf: Global matching scene flow. *Advances in Neural Information Processing Systems*, 36:64415–64427, 2023.

[67] S. Zhu, A. Yarovoy, and F. Fioranelli. Deepego: Deep instantaneous ego-motion estimation using automotive radar. *IEEE Transactions on Radar Systems*, 1:166–180, 2023.

[68] S. Zhu, S. Ravindran, L. Chen, A. G. Yarovoy, and F. Fioranelli. Deepego+: Unsynchronized radar sensor fusion for robust vehicle ego-motion estimation. *IEEE Transactions on Radar Systems*, 3:483–497, 2025.

[69] S. Zhu, S. Ravindran, A. G. Yarovoy, and F. Fioranelli. Redefining radar segmentation: Simultaneous static–moving segmentation and ego-motion estimation using radar point clouds. *IEEE Transactions on Radar Systems*, 4:771–786, 2026.

Contents

1 Introduction **1**

2 Related Work **3**

3 Method **3**

3.1 Aliasing-Aware Spatial-Doppler Factorization 4

3.2 Ego-Conditioned Bi-Flow Regularization 5

3.3 Loss Function 6

4 Experiments Setup **7**

4.1 Datasets 7

4.2 Radar Point Cloud Enhancement 7

4.3 Static Field Estimation 8

4.4 Dynamic Field Estimation 8

4.5 Validation of Spatial-Doppler Guidance 9

5 Conclusion **10**

A Additional Analysis **16**

A.1 Additional Analysis on Doppler Velocity Consistency 16

A.2 Additional Analysis on Physical Radar Spectral Bias 17

A.3 Additional Analysis on Non-uniform Uncertainty 17

B Additional Experimental Details **18**

B.1 Dataset Splits and Preprocessing 18

B.2 Implementation Details 19

B.3 Training Objective 19

B.4 Evaluation Protocol 20

B.5 Ablation Analysis on Static and Dynamic Velocity Fields 21

B.6 Broader Impacts 22

B.7 Asset Licenses. 22

B.8 Compute Resources. 22

A Additional Analysis

A.1 Additional Analysis on Doppler Velocity Consistency

Doppler measurements provide a physically grounded one-dimensional projection of the underlying scene-level velocity field. For a radar voxel centered at polar coordinate (r, α, β), we denote its unit line-of-sight vector by

$$\mathbf{n}(\alpha, \beta) = [\cos\alpha\cos\beta \quad \sin\alpha\cos\beta \quad \sin\beta]^\top . \tag{12}$$

In RaVF, the velocity field is decomposed into an ego-induced static component and an independently moving dynamic residual:

$$\mathbf{F}(\mathbf{p}) = \mathbf{F}_{\text{sta}}(\mathbf{p}; \boldsymbol{\xi}) + \mathbf{F}_{\text{dyn}}(\mathbf{p}), \qquad \boldsymbol{\xi} = (\mathbf{v}, \boldsymbol{\omega}), \tag{13}$$

where $\mathbf{p} = r\mathbf{n}(\alpha, \beta)$ is the voxel center in the radar frame, and $\boldsymbol{\xi}$ denotes the ego-motion twist. For a stationary scatterer in the world, its apparent radar-frame velocity is fully determined by ego motion:

$$\mathbf{F}_{\text{sta}}(\mathbf{p}; \boldsymbol{\xi}) = -\mathbf{v} - \boldsymbol{\omega} \times \mathbf{p}. \tag{14}$$

Therefore, the Doppler velocity observed at this voxel should satisfy the following radial consistency:

$$v_D(\mathbf{p}) = s_D\, \mathbf{n}(\alpha, \beta)^\top \left(\mathbf{F}_{\text{sta}}(\mathbf{p}; \boldsymbol{\xi}) + \mathbf{F}_{\text{dyn}}(\mathbf{p})\right) + \epsilon_D, \tag{15}$$

$$= s_D\, \mathbf{n}(\alpha, \beta)^\top \left(-\mathbf{v} - \boldsymbol{\omega} \times \mathbf{p}\right) + s_D\, \mathbf{n}(\alpha, \beta)^\top \mathbf{F}_{\text{dyn}}(\mathbf{p}) + \epsilon_D, \tag{16}$$

where $s_D \in \{-1, 1\}$ accounts for the Doppler sign convention and ϵ_D denotes Doppler noise.

This relation reveals an important structural property of radar velocity fields. For static background voxels, $\mathbf{F}_{\text{dyn}}(\mathbf{p}) = \mathbf{0}$, and the Doppler field is a low-dimensional function governed by the global ego motion. In contrast, independently moving objects violate this global static pattern and induce a residual radial component:

$$\Delta v_D(\mathbf{p}) = v_D(\mathbf{p}) - s_D\, \mathbf{n}(\alpha, \beta)^\top \mathbf{F}_{\text{sta}}(\mathbf{p}; \boldsymbol{\xi}) \approx s_D\, \mathbf{n}(\alpha, \beta)^\top \mathbf{F}_{\text{dyn}}(\mathbf{p}). \tag{17}$$

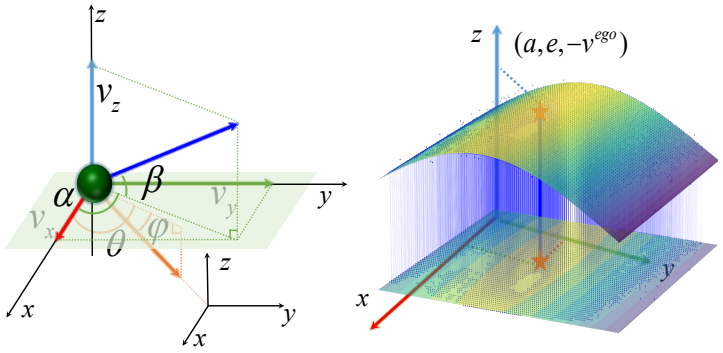


Figure 5: Doppler-Consistency for stationary positive scatters.

Visually, the static Doppler distribution is shown in fig. 5. The shape of surface is determined by the ego-motion parameters, and the Doppler values on this surface are consistent with the static background. Deviations from this surface indicate potential dynamic motion or unreliable measurements. Thus, Doppler consistency naturally separates the scene into an ego-motion-induced static field and a dynamic residual field. This motivates the design of RaVF: instead of directly regressing unconstrained motion at every voxel, RaVF first models the physically regular static field induced by ego motion and then learns the object-centric residual field associated with independently moving targets.

When the radar Doppler is folded by the maximum unambiguous velocity $v_{\max}$, the measured Doppler $\widetilde{v}_D$ belongs to an equivalence class:

$$\mathcal{U}(\widetilde{v}_D) = \{\widetilde{v}_D + 2mv_{\max} \mid m \in \mathbb{Z}\}\,. \tag{18}$$

RaVF uses the static-field prediction as a physical prior to select a Doppler-consistent unfolded candidate:

$$v_D^\star(\mathbf{p}) = \arg\min_{u \in \mathcal{U}(\widetilde{v}_D)} \left| u - s_D\, \mathbf{n}(\alpha,\beta)^\top \mathbf{F}_{\mathrm{sta}}(\mathbf{p};\boldsymbol{\xi}) \right|. \tag{19}$$

This formulation is particularly useful for RaVF, since the static field provides a global physical reference, while large residuals after unfolding indicate potential dynamic motion or unreliable spectral responses.

A.2 Additional Analysis on Physical Radar Spectral Bias

Radar spectra are not uniformly informative across the spatial domain. Even before learning, the FFT-processed radar tensor contains strong physical and algorithmic biases caused by range-dependent power attenuation and non-uniform angular sampling.

First, the received radar energy decays rapidly with range. For a scatterer with radar cross section σ_{rcs}, the expected spectral power at range r can be abstracted as

$$\mathbb{E}\left[|\mathcal{S}(r,\alpha,\beta,d)|^2\right] \approx \frac{C G_t(\alpha,\beta) G_r(\alpha,\beta) \sigma_{\mathrm{rcs}}}{r^4} + \sigma_n^2, \tag{20}$$

where $\mathcal{S}$ denotes the spatial-Doppler spectrum, G_t and G_r are the transmit and receive antenna gains, C absorbs waveform-dependent constants, and σ_n^2 is the noise floor. Consequently, distant moving objects may produce much weaker spectral responses than nearby static structures, even when their motion is more relevant for velocity-field estimation. This range-dependent spectral imbalance biases a network toward short-range high-energy regions unless explicitly compensated or regularized.

Second, the azimuth dimension is biased by the direction-of-arrival estimation process. For a uniform linear array, the phase progression is linear in spatial frequency $\mu = \sin\alpha$ rather than in the physical angle α. The beamforming response can be written as

$$P(\mu) = \left| \sum_{m=0}^{M-1} y_m \exp(-j\pi m \mu) \right|^2, \qquad \mu = \sin\alpha, \tag{21}$$

where y_m is the received signal at the m-th virtual antenna. Since FFT beamforming samples μ approximately uniformly, the corresponding angular grid is obtained by

$$\alpha_i = \arcsin(\mu_i), \qquad \Delta\alpha_i \approx \frac{\Delta\mu}{\sqrt{1-\mu_i^2}} = \frac{\Delta\mu}{\cos\alpha_i}. \tag{22}$$

Therefore, the azimuth bins are inherently non-uniform in the physical angular domain. The effective angular resolution becomes worse near large azimuth angles, leading to broader spatial uncertainty and angle-dependent spectral distortion. For RaVF, this implies that the same Cartesian motion may induce different spectral patterns at different azimuths, making physically aware modeling important for robust velocity-field learning.

A.3 Additional Analysis on Non-uniform Uncertainty

Radar measurements are naturally represented in polar coordinates, whereas scene flow and ego-motion are usually evaluated in Cartesian space. This coordinate transformation introduces non-uniform and anisotropic uncertainty, especially for far-range objects.

Let the radar measurement be

$$\mathbf{q} = [r \quad \alpha \quad \beta]^\top, \qquad \mathbf{p} = \Phi(\mathbf{q}) = r \begin{bmatrix} \cos\alpha\cos\beta \\ \sin\alpha\cos\beta \\ \sin\beta \end{bmatrix}. \tag{23}$$

Assume the polar measurement perturbation is $\delta\mathbf{q} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_q)$. A first-order propagation gives

$$\delta\mathbf{p} \approx \mathbf{J}_\Phi(\mathbf{q})\delta\mathbf{q}, \qquad \boldsymbol{\Sigma}_p \approx \mathbf{J}_\Phi(\mathbf{q})\boldsymbol{\Sigma}_q \mathbf{J}_\Phi(\mathbf{q})^\top. \tag{24}$$

Using the decomposition

$$\mathbf{J}_\Phi(\mathbf{q}) = \left[\mathbf{n} \quad r\frac{\partial \mathbf{n}}{\partial\alpha} \quad r\frac{\partial \mathbf{n}}{\partial\beta} \right], \tag{25}$$

we can see that angular uncertainty is amplified by the range r. For the common diagonal case $\mathbf{\Sigma}_q = \mathrm{diag}(\sigma_r^2, \sigma_\alpha^2, \sigma_\beta^2)$, the total Cartesian uncertainty satisfies

$$\mathrm{tr}(\mathbf{\Sigma}_p) = \sigma_r^2 + r^2\sigma_\alpha^2 \left\| \frac{\partial \mathbf{n}}{\partial \alpha} \right\|_2^2 + r^2\sigma_\beta^2 \left\| \frac{\partial \mathbf{n}}{\partial \beta} \right\|_2^2 \tag{26}$$

$$= \sigma_r^2 + r^2\sigma_\alpha^2 \cos^2\beta + r^2\sigma_\beta^2. \tag{27}$$

Hence, even if the angular uncertainty is constant in polar coordinates, its Cartesian effect grows quadratically with range. This explains why far-range radar voxels exhibit larger spatial ambiguity and why uniform Cartesian losses may over-penalize distant predictions.

For velocity-field estimation, this uncertainty further propagates to the estimated inter-frame displacement. Let the velocity field be approximated by two radar-frame positions:

$$\mathbf{F}_{k\to k+1} = \frac{\mathbf{p}_{k+1} - \mathbf{T}_{k\to k+1}\mathbf{p}_k}{\Delta t}. \tag{28}$$

The corresponding uncertainty can be approximated as

$$\mathbf{\Sigma}_F \approx \frac{1}{\Delta t^2}\left(\mathbf{\Sigma}_{p_{k+1}} + \mathbf{R}_{k\to k+1}\mathbf{\Sigma}_{p_k}\mathbf{R}_{k\to k+1}^\top + \mathbf{J}_T\mathbf{\Sigma}_T\mathbf{J}_T^\top\right), \tag{29}$$

where $\mathbf{R}_{k\to k+1}$ is the rotation component of the ego transformation, $\mathbf{\Sigma}_T$ denotes ego-motion uncertainty, and $\mathbf{J}_T$ is the Jacobian with respect to the transformation parameters. Therefore, the reliability of a radar velocity vector is jointly determined by range, angular resolution, Doppler quality, and ego-motion uncertainty. This non-uniform uncertainty motivates RaVF to exploit physically grounded spatial-Doppler constraints rather than treating all voxels as equally reliable observations.

B Additional Experimental Details

B.1 Dataset Splits and Preprocessing

We evaluate *RaVF* on two real-world 4D radar datasets, **RaDelft** [42] and **K-Radar** [36], both of which provide low-level radar measurements and synchronized LiDAR observations. For RaDelft, we adopt a scene-disjoint split: Scenes 1, 2, 3, 6, and 7 are used for training and validation, while Scenes 4 and 5 are reserved for testing. This split prevents temporally adjacent frames from appearing across training and testing and better reflects cross-scene generalization. For K-Radar, we select the sequence 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 18, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30 considering the odometry quality and flow supervision reliability instead of some high GPS variance scenes like passing through tunnels or travels in adverse weather conditions. The selected sequences are split into training and testing by the 8:2 ratio, with the first 80% of frames used for training and the remaining 20% for testing.

For each training sample, we use two consecutive radar frames at 10 Hz, denoted by frame k and frame $k+1$. The input consists of FFT-processed radar tensors and folded Doppler maps from both frames:

$$\mathcal{X}_{k,k+1} = \{\mathcal{R}_k, \mathcal{D}_k, \mathcal{R}_{k+1}, \mathcal{D}_{k+1}\}, \tag{30}$$

where $\mathcal{R}$ denotes the radar power cube and $\mathcal{D}$ denotes the Doppler map derived from the radar spectrum. LiDAR measurements are used only to construct supervision signals, including occupancy, scene-level velocity field, and dynamic masks when available. Unless otherwise specified, the radar tensor is normalized and cropped to the region of interest

$$r \in [0, 58.88]\mathrm{m}, \qquad \alpha \in [-53^\circ, 53^\circ], \qquad \beta \in [-18^\circ, 18^\circ]. \tag{31}$$

All predictions are represented on a spherical voxel grid of size

$$R \times A \times E = 128 \times 128 \times 64. \tag{32}$$

The LiDAR point cloud is converted into the radar coordinate system and voxelized on the same radar grid. The occupancy label is defined as

$$\mathcal{O}^\star(\mathbf{u}) = \mathbb{1}\left[\exists \mathbf{p}^{\mathrm{lidar}} \text{ s.t. } \mathbf{p}^{\mathrm{lidar}} \in \mathcal{V}(\mathbf{u})\right], \tag{33}$$

where $\mathbf{u}$ indexes a radar voxel and $\mathcal{V}(\mathbf{u})$ denotes its corresponding spatial cell. The scene-level velocity target $\mathbf{F}^{\star}_{k\to k+1}$ is generated on occupied voxels by associating LiDAR-derived geometry across adjacent frames. We define dynamic supervision according to the magnitude of the ground-truth velocity field:

$$\mathcal{M}_{\mathrm{dyn}}(\mathbf{u}) = \mathbb{1}\left[\left\|\mathbf{F}^{\star}_{k\to k+1}(\mathbf{u})\right\|_2 > \tau_{\mathrm{dyn}}\right], \qquad \tau_{\mathrm{dyn}} = 0.3\mathrm{m/s}. \tag{34}$$

This definition allows the model to focus velocity-field learning on independently moving regions while still using occupancy and Doppler constraints to model the ego-induced static field.

Ground-Truth Velocity Construction. We construct velocity-field supervision using official odometry, calibrated LiDAR, annotated 3D boxes, and persistent object IDs. Each source-frame LiDAR point is assigned to either the static background or a tracked object. Background motion is derived from the official ego-motion transform, while object motion is obtained by associating consecutive box poses through persistent IDs. The object rigid transform is cross-validated using Kabsch alignment, retaining the candidate with lower Chamfer error.

All point motions are ego-compensated and expressed in the source radar frame before voxel-wise aggregation to obtain $\mathbf{F}^{\star}_{k\to k+1}$. Static background points consequently have zero residual velocity. Points with missing IDs, ambiguous assignments, or invalid spatial support are excluded. Label uncertainty mainly arises from synchronization, calibration, odometry and box-pose errors, sparse LiDAR observations, and rigid-motion assumptions.

B.2 Implementation Details

RaVF is implemented in PyTorch Lightning. We train the model with distributed data parallelism on multiple GPUs. The number of data-loading workers is set to 12. The maximum number of training epochs is 150, with validation performed after every epoch. We use gradient clipping with a maximum norm of 35. The model checkpoint with the lowest validation loss is selected for evaluation, and the top 5 checkpoints together with the last checkpoint are saved. Early stopping is applied with a patience of 15 epochs.

We use AdamW as the optimizer:

$$\theta_{t+1} = \theta_t - \eta_t \frac{\widehat{\mathbf{m}}_t}{\sqrt{\widehat{\mathbf{v}}_t} + \epsilon} - \eta_t \lambda_{\mathrm{wd}} \theta_t, \tag{35}$$

where the initial learning rate is $\eta_0 = 10^{-4}$ and the weight decay is $\lambda_{\mathrm{wd}} = 5 \times 10^{-4}$. The learning rate follows cosine annealing with warm restarts:

$$\eta_t = \eta_{\min} + \frac{1}{2}\left(\eta_0 - \eta_{\min}\right)\left(1 + \cos\frac{\pi T_{\mathrm{cur}}}{T_i}\right), \tag{36}$$

where $\eta_{\min} = 10^{-6}$, $T_0 = 10$, and $T_{\mathrm{mult}} = 2$. Unless otherwise stated, the final model is trained jointly. For curriculum experiments, we use three phases: occupancy/ego warm-up, Doppler-static refinement, and full velocity-field learning.

B.3 Training Objective

The full training objective combines ego-motion estimation, occupancy prediction, dynamic velocity-field supervision, bidirectional consistency, Doppler unfolding, and static-field regularization:

$$\begin{aligned}\mathcal{L} = {} & \lambda_{\mathrm{ego}}\mathcal{L}_{\mathrm{ego}} + \lambda_{\mathrm{occ}}\mathcal{L}_{\mathrm{occ}} + \lambda_{\mathrm{flow}}\mathcal{L}_{\mathrm{flow}} + \lambda_{\mathrm{bi}}\mathcal{L}_{\mathrm{bi}} \\ & + \lambda_{\mathrm{fold}}\mathcal{L}_{\mathrm{fold}} + \lambda_{\mathrm{sta}}\mathcal{L}_{\mathrm{sta}} + \lambda_{\mathrm{egoD}}\mathcal{L}_{\mathrm{egoD}} + \lambda_{\mathrm{dyn}}\mathcal{L}_{\mathrm{dyn}}.\end{aligned} \tag{37}$$

In our default configuration, we use $\lambda_{\mathrm{ego}} = 20, \lambda_{\mathrm{occ}} = 60, \lambda_{\mathrm{flow}} = 2, \lambda_{\mathrm{bi}} = 0.15, \lambda_{\mathrm{fold}} = 0.2, \lambda_{\mathrm{sta}} = 0.2, \lambda_{\mathrm{egoD}} = 0.2, \lambda_{\mathrm{dyn}} = 1$.

Ego-motion loss. The ego-velocity head is supervised using a Smooth-ℓ_1 loss:

$$\mathcal{L}_{\mathrm{ego}} = \mathrm{SmoothL1}\left(\widehat{\mathbf{v}}_{\mathrm{ego}}, \mathbf{v}^{\star}_{\mathrm{ego}}\right). \tag{38}$$

Occupancy loss. Occupancy supervision uses a weighted binary cross-entropy and Dice loss:

$$\mathcal{L}_{\mathrm{occ}} = \mathcal{L}_{\mathrm{WBCE}}\left(\widehat{\mathcal{O}}, \mathcal{O}^{\star}\right) + \left(1 - \frac{2\sum_{\mathbf{u}} \widehat{\mathcal{O}}(\mathbf{u})\mathcal{O}^{\star}(\mathbf{u}) + \epsilon}{\sum_{\mathbf{u}} \widehat{\mathcal{O}}(\mathbf{u}) + \sum_{\mathbf{u}} \mathcal{O}^{\star}(\mathbf{u}) + \epsilon}\right). \tag{39}$$

The positive occupancy weight is set to 6 to compensate for the sparsity of radar-occupied voxels.

Velocity-field loss. Since the static field is largely determined by ego motion, we mainly supervise the residual velocity field on dynamic voxels:

$$\mathcal{L}_{\mathrm{flow}} = \frac{\sum_{\mathbf{u}} w_{\mathrm{flow}}(\mathbf{u})\rho\left(\widehat{\mathbf{F}}_{k\to k+1}(\mathbf{u}) - \mathbf{F}^{\star}_{k\to k+1}(\mathbf{u})\right)}{3\sum_{\mathbf{u}} w_{\mathrm{flow}}(\mathbf{u}) + \epsilon}, \tag{40}$$

$$w_{\mathrm{flow}}(\mathbf{u}) = \mathcal{M}_{\mathrm{dyn}}(\mathbf{u}) + \lambda^{\mathrm{flow}}_{\mathrm{stat}}\mathcal{O}^{\star}(\mathbf{u})\left(1 - \mathcal{M}_{\mathrm{dyn}}(\mathbf{u})\right). \tag{41}$$

Here $\rho(\cdot)$ is Smooth-ℓ_1 with $\beta = 0.5$. We set $\lambda^{\mathrm{flow}}_{\mathrm{stat}} = 0$ in the main experiments, which avoids over-constraining static occupied voxels and makes the flow supervision focus on independently moving objects.

Bidirectional velocity consistency. RaVF predicts both forward and backward velocity fields. The backward field is transformed to the forward coordinate system using an ego-conditioned transfer operator $\mathcal{T}_{k\to k+1}$. The anti-symmetric consistency loss is

$$\mathcal{L}_{\mathrm{bi}} = \frac{\sum_{\mathbf{u}} \mathcal{O}^{\star}(\mathbf{u})\left\|\widehat{\mathbf{F}}_{k\to k+1}(\mathbf{u}) + \mathcal{T}_{k\to k+1}\left(\widehat{\mathbf{F}}_{k+1\to k}\right)(\mathbf{u})\right\|_2}{\sum_{\mathbf{u}} \mathcal{O}^{\star}(\mathbf{u}) + \epsilon}. \tag{42}$$

Doppler folding and static consistency. Let $\Pi_{v_{\max}}(\cdot)$ denote the Doppler wrapping operator:

$$\Pi_{v_{\max}}(v) = \mathrm{mod}(v + v_{\max}, 2v_{\max}) - v_{\max}, \qquad v_{\max} = 1.92\mathrm{m/s}. \tag{43}$$

The unfolded Doppler prediction $\widehat{v}_D^{\mathrm{unf}}$ is constrained to reconstruct the observed folded Doppler:

$$\mathcal{L}_{\mathrm{fold}} = \frac{\sum_{\mathbf{u}} \mathcal{O}^{\star}(\mathbf{u})\left|\Pi_{v_{\max}}\left(\widehat{v}_D^{\mathrm{unf}}(\mathbf{u})\right) - v_D^{\mathrm{fold}}(\mathbf{u})\right|}{\sum_{\mathbf{u}} \mathcal{O}^{\star}(\mathbf{u}) + \epsilon}. \tag{44}$$

For static occupied voxels, the unfolded Doppler should agree with the radial projection of the ego-induced static field:

$$\mathcal{L}_{\mathrm{sta}} = \frac{\sum_{\mathbf{u}} \mathcal{M}_{\mathrm{sta}}(\mathbf{u})\left|\widehat{v}_D^{\mathrm{unf}}(\mathbf{u}) - \widehat{v}_D^{\mathrm{sta}}(\mathbf{u})\right|}{\sum_{\mathbf{u}} \mathcal{M}_{\mathrm{sta}}(\mathbf{u}) + \epsilon}, \tag{45}$$

where

$$\mathcal{M}_{\mathrm{sta}}(\mathbf{u}) = \mathcal{O}^{\star}(\mathbf{u})\left(1 - \mathcal{M}_{\mathrm{dyn}}(\mathbf{u})\right). \tag{46}$$

We further align the folded static Doppler with the observed Doppler:

$$\mathcal{L}_{\mathrm{egoD}} = \frac{\sum_{\mathbf{u}} \mathcal{M}_{\mathrm{sta}}(\mathbf{u})\left|\Pi_{v_{\max}}\left(\widehat{v}_D^{\mathrm{sta}}(\mathbf{u})\right) - v_D^{\mathrm{fold}}(\mathbf{u})\right|}{\sum_{\mathbf{u}} \mathcal{M}_{\mathrm{sta}}(\mathbf{u}) + \epsilon}. \tag{47}$$

Dynamic classification loss. Dynamic voxels are extremely sparse. We therefore use a weighted BCE loss with an adaptive positive weight, capped at 500, together with a Dice term:

$$\mathcal{L}_{\mathrm{dyn}} = \mathcal{L}_{\mathrm{WBCE}}\left(\widehat{\mathcal{M}}_{\mathrm{dyn}}, \mathcal{M}_{\mathrm{dyn}}\right) + 1 - \frac{2\sum_{\mathbf{u}} \widehat{\mathcal{M}}_{\mathrm{dyn}}(\mathbf{u})\mathcal{M}_{\mathrm{dyn}}(\mathbf{u}) + \epsilon}{\sum_{\mathbf{u}} \widehat{\mathcal{M}}_{\mathrm{dyn}}(\mathbf{u})\mathcal{O}^{\star}(\mathbf{u}) + \sum_{\mathbf{u}} \mathcal{M}_{\mathrm{dyn}}(\mathbf{u}) + \epsilon}. \tag{48}$$

B.4 Evaluation Protocol

We evaluate *RaVF* on ego-velocity estimation, dynamic scene-flow estimation, and radar occupancy quality. For ego motion, we report the mean absolute error (MAE) of the estimated ego velocity:

$$\mathrm{MAE}_{\mathrm{ego}} = \frac{1}{N}\sum_{i=1}^{N}\left\|\widehat{\mathbf{v}}^{(i)}_{\mathrm{ego}} - \mathbf{v}^{\star(i)}_{\mathrm{ego}}\right\|_1. \tag{49}$$

For dynamic scene flow, metrics are computed on valid dynamic voxels:

$$\Omega_{\mathrm{dyn}} = \{\mathbf{u} \mid \mathcal{O}^{\star}(\mathbf{u}) > 0.5,\ \|\mathbf{F}^{\star}(\mathbf{u})\|_2 > \tau_{\mathrm{dyn}}\}, \qquad \tau_{\mathrm{dyn}} = 0.3\ \mathrm{m/s}. \tag{50}$$

The dynamic endpoint error is defined as

$$\mathrm{EPE}_{\mathrm{dyn}} = \frac{1}{|\Omega_{\mathrm{dyn}}|} \sum_{\mathbf{u}\in\Omega_{\mathrm{dyn}}} e(\mathbf{u}), \qquad e(\mathbf{u}) = \left\| \widehat{\mathbf{F}}(\mathbf{u}) - \mathbf{F}^{\star}(\mathbf{u}) \right\|_2. \tag{51}$$

We further report strict accuracy, relaxed accuracy, and severe outlier rate:

$$\mathrm{AccS} = \frac{1}{|\Omega_{\mathrm{dyn}}|} \sum_{\mathbf{u}\in\Omega_{\mathrm{dyn}}} \mathbb{1}\left[e(\mathbf{u}) < 1.0 \ \vee\ \frac{e(\mathbf{u})}{\|\mathbf{F}^{\star}(\mathbf{u})\|_2 + \epsilon} < 0.10 \right], \tag{52}$$

$$\mathrm{AccR} = \frac{1}{|\Omega_{\mathrm{dyn}}|} \sum_{\mathbf{u}\in\Omega_{\mathrm{dyn}}} \mathbb{1}\left[e(\mathbf{u}) < 2.0 \ \vee\ \frac{e(\mathbf{u})}{\|\mathbf{F}^{\star}(\mathbf{u})\|_2 + \epsilon} < 0.20 \right], \tag{53}$$

$$\mathrm{OutlierR} = \frac{1}{|\Omega_{\mathrm{dyn}}|} \sum_{\mathbf{u}\in\Omega_{\mathrm{dyn}}} \mathbb{1}\left[e(\mathbf{u}) > 4.0 \ \vee\ \frac{e(\mathbf{u})}{\|\mathbf{F}^{\star}(\mathbf{u})\|_2 + \epsilon} > 0.40 \right]. \tag{54}$$

Cosine similarity measures the directional consistency between predicted and ground-truth dynamic flow:

$$\mathrm{CosSim} = \frac{1}{|\Omega_{\mathrm{ang}}|} \sum_{\mathbf{u}\in\Omega_{\mathrm{ang}}} \frac{\widehat{\mathbf{F}}(\mathbf{u})^{\top}\mathbf{F}^{\star}(\mathbf{u})}{\left\|\widehat{\mathbf{F}}(\mathbf{u})\right\|_2 \|\mathbf{F}^{\star}(\mathbf{u})\|_2 + \epsilon}, \tag{55}$$

where

$$\Omega_{\mathrm{ang}} = \left\{ \mathbf{u} \in \Omega_{\mathrm{dyn}} \mid \left\|\widehat{\mathbf{F}}(\mathbf{u})\right\|_2 > 0.1,\ \|\mathbf{F}^{\star}(\mathbf{u})\|_2 > 0.1 \right\}. \tag{56}$$

For radar occupancy evaluation, we convert the predicted occupancy volume into a point cloud and compare it with the LiDAR-derived reference. We report the Chamfer metric (CM) for geometric fidelity and the F1 score (FS) for occupancy reliability. During testing, we save the predicted ego velocity, scene occupancy, and scene velocity field for each frame; qualitative point-cloud and flow visualizations are used only for analysis and do not affect the quantitative metrics.

B.5 Ablation Analysis on Static and Dynamic Velocity Fields

We provide additional ablation studies to analyze the contribution of the two core designs in *RaVF*: **P.** denotes the Physics-Aware Spectral Encoding, and **E.** denotes the Ego-Conditioned Bi-Flow Model. For ego-induced static-field estimation, **P.** improves the physical fidelity of radar feature extraction by compensating range-dependent attenuation and angular sampling bias, while **E.** provides an ego-aware temporal constraint that stabilizes motion estimation across frames. As shown in Tab. 2, the full model achieves the best average MAE on K-Radar, reducing the error by 61.3% over the variant without both modules and by 36.8% over the variant without **P.**. On RaDelft, the full model obtains the best results in most speed intervals, although the average MAE is slightly lower for the variant without **P.**; this suggests that physical spectral encoding is particularly beneficial under the longer-range and more diverse K-Radar setting, while RaDelft is less sensitive to this component due to its more constrained acquisition setup.

For dynamic velocity-field estimation, the benefit of both modules becomes more consistent. As shown in Tab. 3, the full model achieves the best $\mathrm{EPE}_{\mathrm{dyn}}$ and OutlierR on both datasets. Compared with removing both modules, the full model reduces $\mathrm{EPE}_{\mathrm{dyn}}$ by 60.0% on K-Radar and 48.4% on RaDelft, while reducing OutlierR by 38.2% and 72.2%, respectively. The large improvement on RaDelft, especially the 69.8 percentage-point gain in AccS, indicates that ego-conditioned bi-flow reasoning is crucial for suppressing noisy dynamic residuals and enforcing temporally coherent object motion. Overall, **P.** mainly improves the reliability of radar spectral evidence, whereas **E.** strengthens bidirectional motion consistency; their combination yields the most robust static and dynamic velocity-field estimation.

B.6 Broader Impacts

This work advances radar-centric motion perception by enabling scene-level velocity-field estimation from 4D radar spectra, which may improve the robustness of autonomous systems under adverse weather, low-light, and visually degraded conditions. Such capability can benefit safety-critical applications including autonomous driving, robotics, and intelligent transportation systems, especially when camera or LiDAR measurements become unreliable. However, improved radar perception may also be misused for surveillance or deployed in safety-critical systems without sufficient validation, potentially leading to privacy concerns or unsafe decisions under out-of-distribution scenarios. Therefore, we emphasize that RaVF should be evaluated carefully across diverse environments and used as a complementary perception module rather than a standalone safety guarantee.

B.7 Asset Licenses.

We use publicly available datasets, including RaDelft and K-Radar, following their official terms of use and properly citing the original papers and dataset sources. For baseline methods, we cite the corresponding papers and use publicly released implementations when available, respecting their licenses and usage terms. No proprietary or restricted assets are redistributed in this work.

B.8 Compute Resources.

All experiments were conducted on a Linux server running Ubuntu 22.04.5 LTS with kernel 6.8.0. The server is equipped with two AMD EPYC 9555 64-core CPUs, providing 256 CPU threads in total, 755 GB of system memory, and eight NVIDIA L40S GPUs. The storage system includes an ext4-based 3.44 TB system disk, a 72.47 TB data disk, and a 3.44 TB SSD workspace for intermediate files and experiment outputs. Unless otherwise specified, *RaVF* was trained using distributed data parallelism on all eight L40S GPUs with a per-GPU batch size of 2. Training the full model for 150 epochs took approximately 7 days.

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: The abstract and introduction accurately summarize the paper's main contributions, including radar velocity-field learning from spatial-Doppler spectra, ego-induced static field and dynamic residual disentanglement, occupancy-guided Doppler reasoning, and bidirectional velocity fusion. The claimed scope is supported by the method design and experiments on large-scale real-world 4D radar datasets.

   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 limitations arising from tangential motion, multipath reflections, and LiDAR-derived supervision. We also acknowledge the challenges of cross-sensor generalization and evaluation under severe domain shifts.

   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: [Yes]

   Justification: The paper provides the assumptions and derivations for the theoretical analyses in the appendix, including Doppler velocity consistency, spectral bias induced by radar propagation and beamforming, and non-uniform uncertainty under polar-to-Cartesian transformation. These results are used to motivate the design of RaVF rather than to claim formal optimality or convergence guarantees.

   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: The paper provides the key information needed to reproduce the main experimental results, including dataset splits, preprocessing protocols, model architecture, loss functions, training hyperparameters, and evaluation metrics. We will further release the code, configurations, and processed evaluation scripts to facilitate exact reproduction.

   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: The experiments are conducted on publicly available datasets, and the manuscript provides the necessary implementation details, training settings, and evaluation protocols. However, the project code and processed scripts are not publicly released at submission time; we plan to release the code and configurations upon publication.

   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: The paper specifies the experimental settings needed to understand and reproduce the main results, including dataset splits, preprocessing protocols, model architecture, optimizer, learning-rate schedule, loss functions, training hyperparameters, and evaluation metrics. Additional implementation and evaluation details are provided in the supplemental material.

   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: The paper reports results on fixed dataset splits and provides per-sequence or per-condition breakdowns to show consistency across scenarios, but it does not report error bars or statistical significance tests from multiple independent runs.

   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: The supplemental material reports the compute resources used for the main experiments, including the distributed training setup, number of GPUs, per-GPU batch size, number of data-loading workers, approximate training/evaluation time, and peak GPU memory usage.

   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: The research conforms to the NeurIPS Code of Ethics. It uses publicly available autonomous-driving radar datasets under their intended research usage, does not involve human-subject experiments or personal identification, and focuses on radar perception for scene-level motion understanding. We disclose experimental protocols and discuss limitations and potential safety-relevant considerations for deployment in autonomous-driving scenarios.

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: The paper discusses both positive impacts, such as improving robust radar-based perception for autonomous driving and robotics under adverse conditions, and potential negative impacts, including misuse for surveillance and risks from deployment without sufficient validation in safety-critical scenarios.

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: The work does not release high-risk generative models, scraped personal datasets, or sensitive data that would require dedicated misuse safeguards. The experiments are conducted on public autonomous-driving radar datasets and the released code/configurations are intended for research reproduction of radar perception experiments.

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: The paper properly credits the original creators of the datasets and baseline methods through citations, and uses publicly available assets under their official terms of use. We also state in the supplemental material that dataset and code assets are used according to their licenses and that no proprietary or restricted assets are redistributed.

    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: [N/A]

    Justification: The paper does not introduce new datasets, benchmarks, or pretrained model assets. It proposes a new method and evaluates it on existing public datasets; code and configurations are planned for release upon publication with documentation for reproduction.

    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: The paper does not involve crowdsourcing experiments or research with human subjects. We use existing public autonomous-driving radar datasets and do not collect new human-subject data or conduct participant-based studies.

    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: The paper does not involve research with human subjects, crowdsourcing experiments, or newly collected participant data. We use existing public autonomous-driving radar datasets, so IRB approval or equivalent human-subject review is not applicable.

    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: The paper does not use LLMs as an important, original, or non-standard component of the core methodology. Any LLM assistance, if used, was limited to writing, editing, or formatting and did not affect the scientific content, method design, experiments, or conclusions.

    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.