UNIVERSITY OF CALIFORNIA SAN DIEGO

Hybrid Model/Data-Driven Solutions in ISAC: Improving Link Establishment, Channel Estimation, and User Localization in mmWave Vehicular Networks

A dissertation submitted in partial satisfaction of the
requirements for the degree
Doctor of Philosophy

in

Electrical Engineering
(Communication Theory & Systems)

by

Yun Chen

Committee in charge:

Professor Nuria González-Prelcic, Chair
Professor Jorge Cortes
Professor Robert W. Heath Jr.
Professor Florian Meyer

2026

The Dissertation of Yun Chen is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2026

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

My doctoral experience has been a combination of challenges, discoveries, and rewarding experiences. While exploring the frontiers of my research, I often felt like a kite, free to explore, yet always guided and supported by the people and environment around me. Their encouragement gave me the strength to persevere and the joy of seeing my ideas take shape. I am deeply grateful to everyone who has stood by me from the beginning to the end of this journey. All the struggles and uncertainties along the way became worthwhile in the moment of achieving these results. I may not be exceptionally gifted, but being able to contribute even a small part to the research community is a lifelong source of pride.

I am deeply grateful to Prof. Nuria for being the most impactful figure in my academic life. Her guidance has been instrumental in transforming my research since she became my supervisor. She provided me with abundant resources, support, and encouragement, creating a secure and nurturing environment throughout my doctoral program. Under her supervision, I learned how to approach problems, identify strategies and develop solutions, and ultimately turn initial challenges into meaningful achievements. Beyond being an exceptional academic advisor, she has been a true mentor. During difficult periods of struggle and setback, her counsel gave me the strength to persevere in the face of frustration. She taught me how to navigate challenges unique to women in academia and empowered me both professionally and personally. Under her guidance, I have grown not only as a researcher but also as a person. Her wisdom and support will continue to guide and inspire me throughout my career.

I am also sincerely grateful to Prof. Nuria for leading such a welcoming and supportive lab that I have come to regard as a second home, the WiSeCom Lab. It has been a pleasure to work alongside so many talented and dedicated colleagues. The lab offers a warm and open environment where we can have meaningful conversations about our work, travel experiences, and life, while also maintaining a focused and rigorous research atmosphere that enables us to engage deeply with cutting-edge topics. The time spent together, especially in the post-pandemic period, has enhanced my confidence in pursuing research projects and improved my ability to

collaborate effectively within a team. I am grateful to every colleague, whether within our lab or from collaborating groups, who offered their insights, feedback, or simply a fresh perspective when I was too immersed in a problem to see it clearly.

I was fortunate to have two internships at Ericsson in Santa Clara, CA, from 2019 to 2020, during which our team received a *Best Paper Award* at IEEE WCNC 2020. From 2021 to 2023, I completed three internships at Qualcomm, San Diego, CA. Through these experiences, I engaged in meaningful work that provided valuable insight into how academic research translates into industrial applications. I also had the opportunity to contribute to patent filings aimed at advancing the broader research and engineering community. These experiences have deepened my understanding of how academic knowledge can be applied in practice and how rigorous problem-solving approaches manifest in real-world outcomes.

I also thank the ECE Office and the administrative units at UCSD, including the International Students and Programs Office (ISPO) and the Division of Graduate Education and Postdoctoral Affairs (GEPA), for their efficient support throughout my program. Their professionalism and responsiveness greatly eased the administrative process, especially as a transfer student managing extensive paperwork and time-sensitive administrative challenges. Their assistance made a meaningful difference throughout my time at UCSD.

Finally, to my parents and all my relatives thousands of miles away in China, from whom I have been separated since early 2021: I deeply miss them. Their video calls have been a constant source of strength during difficult times, whether I was facing failed experiments, paper rejections, or moments of exhaustion. Their support has been like a spring in the desert, helping me persevere. Without them, I would have lost one of the most important anchors in my life. Knowing they are always with me has allowed me to stay strong, grow more independent, and continue moving forward.

Chapter 1, in part, is a reprint, with permission, of the material as it appears in the papers: Y. Chen, A. Graff, N. González-Prelcic, T. Shimizu, “Radar Aided mmWave Vehicle-to-Infrastructure Link Configuration Using Deep Learning”, in The 2021 IEEE Global

Communications Conference (IEEE GLOBECOM 2021), Madrid, Spain, Dec., 2021; A. Graff, Y. Chen, N. González-Prelcic, T. Shimizu, "Deep Learning-based Link Configuration for Radar-aided Multiuser mmWave Vehicle-to-Infrastructure Communication", in IEEE Trans. Veh. Technol., doi: 10.1109/TVT.2023.3239227, 2022; and N. González-Prelcic, A. Ali, Y. Chen, "Radar-aided Millimeter Wave Communication", in Signal Processing for Joint Radar Communications, Wiley-IEEE Press, 2024. The dissertation author was the primary investigator and author of these papers, which were supported in part by the National Science Foundation (NSF) under Grant No. 2147955 and by a gift from Toyota Motor North America, Inc., USA.

Chapter 2, in part, is a reprint, with permission, of the material as it appears in the papers: Y. Chen, J. Palacios, N. González-Prelcic, T. Shimizu, H. Lu, "Joint Initial Access and Localization in Millimeter Wave Vehicular Networks: a Hybrid Model/Data Driven Approach", in 2022 IEEE 12th Sensor Array and Multichannel Signal Processing Workshop (SAM 2022), Trondheim, Norway, Jun., 2022; Y. Chen, N. González-Prelcic, T. Shimizu, and H. Lu, "Learning to localize with attention: From sparse mmWave channel estimates from a single BS to high accuracy 3D location," IEEE Trans. Wireless Commun., March 2024 (second round of reviews); and Y. Chen, N. González-Prelcic, T. Shimizu, H. Lu, C. Mahabal, "Beamspace ESPRIT-D for Joint 3D Angle and Delay Estimation for Joint Localization and Communication at MmWave", IEEE International Conference on Communications (IEEE ICC 2024), Denver, USA, Jun., 2024. The dissertation author was the primary investigator and author of these papers, which were partially supported by the NSF under grant no. 2433782, no. 2147955, and in part by funds from the federal agency and industry partners as specified in the Resilient & Intelligent NextG Systems (RINGS) program, and by a gift from Toyota Motor North America, Inc., USA.

Chapter 3, in part, is a reprint, with permission, of the material as it appears in the papers: *[Best Student Paper]* Y. Chen, N. González-Prelcic, T. Shimizu, H. Lu, C. Mahabal, "Sparse Recovery with Attention: A Hybrid Data/Model Driven Solution for High Accuracy Position and Channel Tracking at mmWave", in the 2023 IEEE 24th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC 2023), Shanghai, China, Sep., 2023;

Y. Chen, N. González-Prelcic, T. Shimizu, C. Mahabal, "V2X-Enabled Collaborative Orientation Estimation and Position Tracking in mmWave Networks", 2024 58th Asilomar Conference on Signals, Systems, and Computers. IEEE, 2024, Oct., 2024; and Y. Chen, N. González-Prelcic, T. Shimizu, C. Mahabal, "A Hybrid Model/Data-Driven Solution to Channel, Position and Orientation Tracking in mmWave Vehicular Systems.", IEEE J. Sel. Areas Commun., vol. 44, pp. 927–941, Mar. 2026, doi: 10.1109/JSAC.2025.3612354. The dissertation author was the primary investigator and author of these papers, which were partially supported by the NSF under grant no. 2433782, no. 2147955, and in part by funds from the federal agency and industry partners as specified in the Resilient & Intelligent NextG Systems (RINGS) program, and by a gift from Toyota Motor North America, Inc., USA.

VITA

2017	B. E. in Information Engineering, Southeast University, Nanjing, China
2020	M.S. in Electrical and Computer Engineering The University of Texas at Austin, Austin, TX, USA
2021–2024	Graduate Research Assistant, North Carolina State University, Raleigh, NC, USA
2024–2025	Graduate Student Researcher University of California San Diego, La Jolla, CA, USA
2026	Ph. D. in Electrical Engineering (Communication Theory & Systems) University of California San Diego, La Jolla, CA, USA

PUBLICATIONS

Y. Chen, N. González-Prelcic, T. Shimizu, C. Mahabal, "A Hybrid Model/Data-Driven Solution to Channel, Position and Orientation Tracking in mmWave Vehicular Systems," *IEEE J. Sel. Areas Commun.*, vol. 44, pp. 927–941, Mar. 2026, doi: 10.1109/JSAC.2025.3612354.

S. Javid, **Y. Chen**, N. González-Prelcic, "Learning-Based Cross-Frequency Channel Prediction in the Upper Mid-band," in *2025 IEEE Global Communications Conference (GLOBECOM 2025)*, Dec. 2025.

Y. Chen, N. González-Prelcic, T. Shimizu, C. Mahabal, "V2X-Enabled Collaborative Orientation Estimation and Position Tracking in mmWave Networks," in *2024 58th Asilomar Conference on Signals, Systems, and Computers*, Pacific Grove, CA, USA, Oct. 2024.

B. Kumar, **Y. Chen**, N. González-Prelcic, T. Shimizu, A. Ganlath, "DeepSatLoc: A Multimodal Fusion Strategy for Enhanced Localization in Urban Canyons Exploiting GPS and LEO Satellite Communication Signals," in *2024 58th Asilomar Conference on Signals, Systems, and Computers*, Pacific Grove, CA, USA, Oct. 2024.

Y. Chen, N. González-Prelcic, T. Shimizu, H. Lu, C. Mahabal, "Beamspace ESPRIT-D for Joint 3D Angle and Delay Estimation for Joint Localization and Communication at mmWave," in *IEEE International Conference on Communications (IEEE ICC 2024)*, Denver, USA, Jun. 2024.

Y. Chen, N. González-Prelcic, T. Shimizu, H. Lu, "Learning to Localize with Attention: From Sparse mmWave Channel Estimates from a Single BS to High Accuracy 3D Location," *IEEE Trans. Wireless Commun.*, 2024, under second round of review.

[Best Student Paper] Y. Chen, N. González-Prelcic, T. Shimizu, H. Lu, C. Mahabal, "Sparse Recovery with Attention: A Hybrid Data/Model Driven Solution for High Accuracy Position and

Channel Tracking at mmWave," in *2023 IEEE 24th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC 2023)*, Shanghai, China, Sep. 2023.

N. González-Prelcic, A. Ali, **Y. Chen**, "Radar-aided Millimeter Wave Communication," in *Signal Processing for Joint Radar Communications*, Wiley-IEEE Press, 2023.

Y. Chen, J. Palacios, N. González-Prelcic, T. Shimizu, H. Lu, "Joint Initial Access and Localization in Millimeter Wave Vehicular Networks: A Hybrid Model/Data Driven Approach," in *2022 IEEE 12th Sensor Array and Multichannel Signal Processing Workshop (SAM 2022)*, Trondheim, Norway, Jun. 2022.

A. Graff, **Y. Chen**, N. González-Prelcic, "Deep Learning-based Link Configuration for Radar-aided Multiuser mmWave Vehicle-to-Infrastructure Communication," *IEEE Trans. Veh. Technol.*, doi: 10.1109/TVT.2023.3239227, 2023.

G. Geraci, A. Garcia-Rodriguez, M. Azari, A. Lozano, M. Mezzavilla, S. Chatzinotas, **Y. Chen**, S. Rangan, M. Renzo, "What Will the Future of UAV Cellular Communications Be? A Flight from 5G to 6G," *IEEE Commun. Surveys Tuts.*, vol. 24, no. 3, pp. 1304–1335, Third Quarter 2022.

Y. Chen, A. Graff, N. González-Prelcic, T. Shimizu, "Radar Aided mmWave Vehicle-to-Infrastructure Link Configuration Using Deep Learning," in *2021 IEEE Global Communications Conference (GLOBECOM 2021)*, Madrid, Spain, Dec. 2021.

Y. Chen, X. Lin, T. Khan, M. Afshang, M. Mozaffari, "5G Air-to-Ground Network Design and Optimization: A Deep Learning Approach," in *2021 IEEE 93rd Vehicular Technology Conference (IEEE VTC 2021)*, Helsinki, Finland, Apr. 2021.

Y. Chen, X. Lin, T. Khan, M. Mozaffari, "A Deep Learning Approach to Efficient Drone Mobility Support," in *2nd Workshop on Drone Assisted Wireless Communications for 5G and Beyond, co-located with ACM MobiCom 2020 (DroneCom 2020)*, London, United Kingdom, Sep. 2020.

Y. Chen, N. González-Prelcic, R. W. Heath, "Collision-free UAV Navigation with a Monocular Camera Using Deep Reinforcement Learning," in *2020 IEEE International Workshop on Machine Learning for Signal Processing*, Espoo, Finland, Sep. 2020.

[Best Paper] Y. Chen, X. Lin, T. Khan, M. Mozaffari, "Efficient Drone Mobility Support Using Reinforcement Learning," in *2020 IEEE Wireless Communications and Networking Conference (IEEE WCNC 2020)*, Seoul, South Korea, May 2020.

Y. Chen, W. Yan, C. Li, Y. Huang, and L. Yang, "Personalized Optimal Bicycle Trip Planning Based on Q-learning Algorithm", in *2018 IEEE Wireless Communications and Networking Conference (IEEE WCNC 2018)*, Barcelona, Spain, Apr. 2018.

Y. Wang, **Y. Chen**, H. Dai, Y. Huang, and L. Yang, "A Learning-Based Approach for Proactive

Caching in Wireless Communication Networks," in *the 9th International Conference on Wireless Communications and Signal Processing*, Nanjing, China, Oct. 2017.

PATENTS

X. Lin, **Y. Chen**, M. Mozaffari, T. Khan, "Efficient 3D Mobility Support Using Reinforcement Learning". US20240056933A1. Feb. 2024.

M. Mozaffari, X. Lin, T. Khan, M. Afshang, **Y. Chen**, "Network Design and Optimization Using Deep Learning." US20230388196A1. Nov. 2023.

ABSTRACT OF THE DISSERTATION

Hybrid Model/Data-Driven Solutions in ISAC: Improving Link Establishment, Channel Estimation, and User Localization in mmWave Vehicular Networks

by

Yun Chen

Doctor of Philosophy in Electrical Engineering
(Communication Theory & Systems)

University of California San Diego, 2026

Professor Nuria González-Prelcic, Chair

The sixth-generation (6G) cellular ecosystem is evolving toward integrated sensing and communication (ISAC), where the same infrastructure supports both high-rate data exchange and environmental perception. Vehicular networks are among the most demanding use cases: autonomous navigation demands ultra-reliable connectivity and centimeter-level positioning accuracy, yet conventional beam training protocols incur prohibitive overhead and global navigation satellite system (GNSS)-based localization fails in urban canyons. This dissertation develops a hybrid model/data-driven framework addressing these challenges through three contributions.

The first introduces a passive radar-aided framework for multiuser millimeter wave (mmWave) link configuration. A roadside unit senses existing automotive frequency modulated continuous wave (FMCW) transmissions from multiple vehicles via a mixing filter bank with constant false alarm rate (CFAR) detection, isolating individual vehicle signals from multiuser interference. Three deep neural network architectures then map estimated radar spatial covariances to communication covariances, compensating for frequency and geometry mismatches between the 76 GHz radar and 73 GHz communication bands. The covariance-prediction-assisted approach reduces beam search space by up to 32 times and improves the achievable rate by 21.9% over radar-only methods.

The second addresses high-accuracy 3D vehicle localization from a single base station snapshot. Two time-domain channel estimation algorithms, a two-stage multidimensional orthogonal matching pursuit (MOMP) variant and ESPRIT-D (an off-grid subspace method), extract multipath parameters while accounting for system filtering effects and unknown clock offsets. A lightweight network, PathNet, classifies estimated paths to select geometrically useful components, while a Transformer-based ChanFormer refines initial geometric position estimates through cross-attention, achieving 28 cm accuracy for 80% of users in line-of-sight (LOS) and sub-meter accuracy for 55% in non-line-of-sight (NLOS).

The third extends single-snapshot localization to continuous tracking and collaborative multi-vehicle positioning. F-MOMP enables efficient high-resolution channel tracking by exploiting temporal correlation and factored dictionaries. VO-ChAT and VP-ChAT leverage channel sequence history through spatial-temporal and cross-attention mechanisms to track vehicle orientation and refine position estimates. For NLOS vehicles, a model-based collaborative framework exploits vehicle-to-vehicle (V2V) sidelinks to geometrically estimate per-vehicle clock and orientation offsets, achieving 0.3 m average error with sub-meter accuracy for 90% of cases.

Together, these contributions demonstrate that embedding physical-layer signal models into deep learning architectures yields ISAC systems that are accurate, computationally efficient,

and interpretable, key properties for safety-critical vehicular deployments.

Introduction

The 6G cellular ecosystem is poised to transcend the traditional boundaries of wireless communications, evolving toward intelligent, sensing-aware networks that unify the physical and digital worlds [2]. As the standardization of 5G-advanced reaches maturity in 3GPP Release 18 and beyond [3], ISAC has emerged as a fundamental paradigm, enabling infrastructure to simultaneously support high-rate data exchange and high-resolution environmental perception [1, 4]. This convergence is particularly transformative for vehicular networks, where safety-critical applications ranging from autonomous navigation to cooperative platooning (where centimeter-level positioning accuracy is required for safe lane-keeping at highway speeds [5]) demand not only ultra-reliable low-latency communication (URLLC) but also precise, robust localization and tracking capabilities that remain resilient under severe channel conditions.

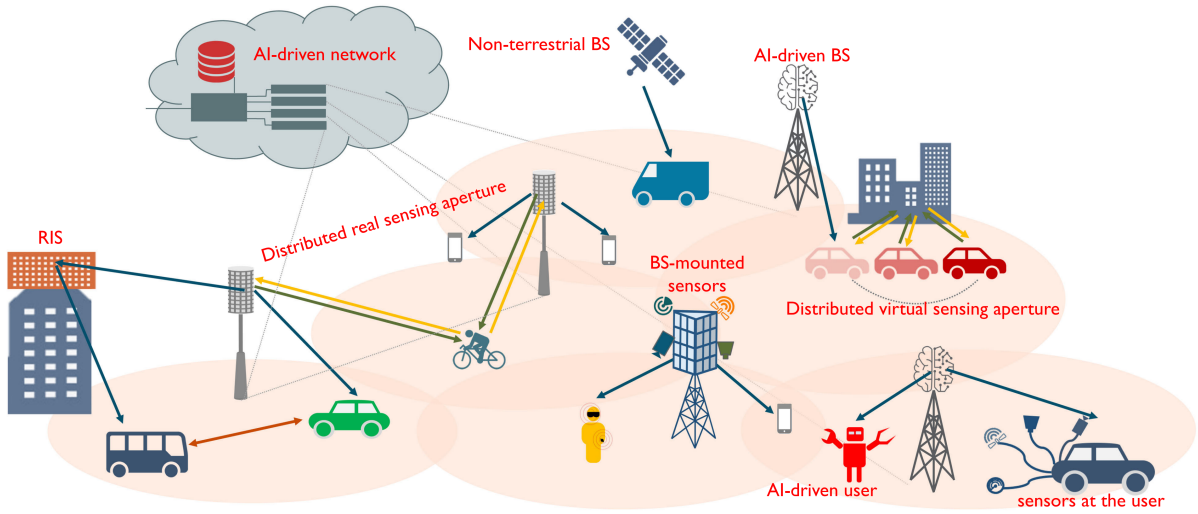


Figure 1. Illustration of the different elements in the infrastructure of an ISAC network. Image source [1].

In ISAC systems as illustrated in Fig. 1, sensing data can be used to acquire channel state information (CSI) and optimize beamforming and handover decisions, ensuring seamless connectivity for vehicles moving at high speeds or in complex environments [6–8]. The mmWave frequencies central to 6G vehicular systems present unique challenges: severe path loss and channel sparsity necessitate high-gain beamforming through large antenna arrays, yet the overhead associated with conventional beam training protocols ranging from hundreds of milliseconds to several seconds becomes prohibitive in highly dynamic vehicular environments [9]. To address this, Chapter 1 proposes a passive radar-aided communication framework wherein the roadside unit (RSU) opportunistically senses existing automotive FMCW radar transmissions to configure mmWave downlink links. An FMCW mixing filter bank with constant false alarm rate (CFAR) detection isolates individual vehicle signals from multiuser interference, while deep neural network (DNN) architectures map estimated radar spatial covariances to communication covariances, accounting for frequency and geometry mismatches between radar (76 GHz) and communication (73 GHz) bands, enabling up to $32\times$ reduction in beam search space and 21.9% rate improvement over traditional beam training methods.

Conversely, communication signals benefit sensing functions by enabling high-precision localization without relying on GNSS, which becomes unreliable in urban canyons due to signal blockage and multipath-induced ranging errors [10]. The geometric sparsity of mmWave channels facilitates 3D localization from high-resolution channel parameter estimates, yet existing methods either neglect realistic system impairments (such as pulse shaping, filtering effects, and unknown transmitter-receiver clock offsets) or suffer from prohibitive computational complexity with practical planar arrays. Chapter 2 addresses this through a hybrid model/data-driven localization framework, introducing two low-complexity time-domain channel estimation algorithms: a two-stage MOMP (an on-grid sparse recovery method) and ESPRIT-D (an off-grid subspace approach), both accounting for system filtering effects. A lightweight neural network (*PathNet*) classifies multipath components to identify LOS and first-order reflections, while a Transformer-based *ChanFormer* employs cross-attention mechanisms to refine geometric localization estimates,

achieving 28 cm accuracy for 80% of users in LOS and sub-meter accuracy for 55% of users in NLOS, representing up to 10× improvement over state-of-the-art (SOTA) fingerprinting and maximum likelihood methods.

Furthermore, the integration of artificial intelligence (AI) and deep learning methodologies significantly broadens the capabilities of vehicular ISAC networks [11, 12]. Rather than pure data-driven approaches that demand extensive training data and lack interpretability, this dissertation advocates for hybrid model/data-driven solutions that strategically embed physical layer knowledge (such as antenna array manifolds, geometric propagation models, and radar signal structures) into deep learning architectures. Chapter 3 exemplifies this through a robust vehicle-to-infrastructure (V2I) tracking system that employs *F-MOMP* (Factored MOMP) for efficient channel tracking with 0.1 ns delay accuracy, reducing dictionary search complexity by factoring the five-dimensional sparse recovery problem into independent per-dimension operations; *VO-ChAT* for orientation tracking via spatial-temporal attention mechanisms; and *VP-ChAT* for position refinement through cross-attention between channel evolution and trajectory history. This framework achieves 2D position errors below 26 cm for 80% of cases with orientation errors below 0.5°. Extensions to collaborative V2V positioning further demonstrate the framework's efficacy in NLOS conditions, achieving 0.2 m average localization error through sidelink exploitation.

By co-designing sensing and communication functions through this hybrid lens, intelligent ISAC systems can reduce hardware duplication and improve spectral efficiency while maintaining the interpretability and reliability required for safety-critical automotive applications [13–15]. These advancements hold significant promise for transforming autonomous driving and smart cities by enabling real-time, untethered user-environment interactions.

0.1 Dissertation layout

This work specifically presents radar-aided communication link establishment as a representative use case of sensing-aided communication (Chapter 1). Conversely, we also

explore how communication signals can support localization during initial access and tracking, representing a communication-aided sensing scenario (Chapters 2 & 3). Throughout the studies, hybrid model/data-driven approaches are employed, with the choice of model/data-based modules guided by the limitations and suitability of each approach.

In Chapter 1, we address the challenge of high communication overhead in configuring mmWave links using conventional beam training protocols. We propose leveraging a passive radar array at the RSU to sense automotive radar signals from multiple vehicles, using a radar processing chain that estimates radar timing, chirp rates, and spatial covariances, filters out interference via CFAR detection, and isolates individual radar signals. Three DNN designs are introduced to map radar spatial features to communication spatial covariances (azimuth power spectrum (APS), covariance eigenvector, and covariance vector, each exploiting a progressively richer representation of the covariance structure) in both LOS and NLOS scenarios. Simulation results on ray-tracing simulated multiuser (MU) channels show that our approach reliably isolates individual radar covariances and significantly outperforms traditional radar-aided methods in terms of beam training overhead reduction and communication rate improvement.

In Chapter 2, we tackle the challenge of accurate vehicle localization in mmWave multiple-input multiple-output (MIMO) systems. Two practical 3D time-domain channel estimation algorithms are proposed, both accounting for system filtering effects: (1) a low-complexity two-stage MOMP algorithm, and (2) estimation of signal parameters via rotational invariance techniques (ESPRIT) with dictionary (ESPRIT-D), a beamspace extension of ESPRIT. These algorithms estimate channel multipath parameters including azimuth and elevation angle-of-arrival (AoA), angle-of-departure (AoD), and time-difference-of-arrival (TDoA). To address path ordering, which is essential for applying geometric relationships among vehicles, RSUs, and scatterers, we introduce a DNN (*PathNet*) to classify paths and select only the LOS and first-order reflections for localization. We further develop a geometry-based 3D localization method robust to LOS and NLOS conditions and unknown clock offsets, and refine the initial position estimates using a Transformer-based attention network (*ChanFormer*). Ray-tracing simulations in realistic

vehicular scenarios demonstrate angular errors below 0.01° and delay errors below 3×10^{-10} s in LOS conditions; localization errors below 28 cm are achieved for 80% of users in LOS. In NLOS, sub-meter accuracy is achieved for 55% of users, representing up to $10\times$ improvement over SOTA methods.

In Chapter 3, we present a robust mmWave vehicle tracking system for practical V2I MIMO communication, explicitly accounting for clock offsets, filtering effects, and antenna orientation misalignments. The system starts with a channel tracking algorithm, *F-MOMP* (Factored MOMP), that uses reduced sparsifying dictionaries and prior frame estimates for efficient high-resolution channel tracking. A geometric solver then provides an initial 2D position estimate, adaptable to both orientation-known and orientation-unknown scenarios. For the former, a Transformer-based network, *VP-ChAT*, refines position estimates after a geometric single-shot localization. For the latter, two Transformer-based networks, *VO-ChAT* for vehicle orientation tracking and *VP-ChAT* for position refinement, leverage historical channel and trajectory data through cross-attention mechanisms. Evaluated on ray-tracing data in urban canyon environments, the proposed system achieves a 90th-percentile 2D position error below 20 cm under the orientation-known condition. In the general case, it attains 2D position errors below 26 cm for 80% of cases, with orientation errors maintained below 0.5°. Furthermore, we propose a collaborative position tracking solution that leverages both V2I and V2V link parameters to enhance localization accuracy for vehicles in NLOS conditions, treating them as targets localized via sidelink channels. A Kalman filter (KF) is applied for temporal tracking. Simulations show orientation offset errors below 0.05°, clock offset errors under 10 ns, and average localization errors of 0.2 m, with 90% of cases below 0.7 m.

In Chapter 4, the main contributions of this work are summarized, highlighting the key methodological advances and performance gains. This is followed by a discussion of potential future research directions to further enhance system robustness, scalability, and applicability in more diverse vehicular environments.

Chapter 1

DL-based Link Configuration for Radar-Aided MU mmWave V2I Communication

1.1 Introduction

The automotive industry is incorporating advanced sensing and communication technologies to produce vehicles that are both more aware of their surroundings and able to communicate with others. This perceptual awareness has been achieved through the use of several onboard sensors, most notably automotive radars. The ability to communicate with infrastructure, referred to as V2I communication, allows for the sharing of sensor data, navigation information, multimedia, and more. These applications need high data rates to operate seamlessly.

Wireless communications at mmWave bands can achieve such high data rates. Communication at mmWave bands, however, typically requires gains through beamforming to reach acceptable signal-to-noise ratio (SNR), often requiring large antenna arrays. The configuration of these large arrays following standard beam training protocols introduces a large overhead (time required to configure the link). For example, for initial access (initial configuration of the link when the user enters the cell), when the number of antennas at the base station (BS) and the user are 64 and 16 respectively, and considering an analog MIMO architecture at both ends, the overhead of the beam training protocol associated with the 3GPP 5G NR standard varies in the range of 750 ms to 5 s as a function of the number of synchronization signal blocks (SSBs) per burst and the UE grid of beams [9]. This training overhead becomes even more demanding in

multiuser V2I links, because the channel coherence times are short and users are rapidly entering and exiting the communication cell, resulting in a high probability that users are in initial access.

The onboard radars on many of these next-generation vehicles provide a unique signal of opportunity that contains out-of-band information that can be leveraged to reduce training overhead during initial access [16]. Since many automotive radars operate in a mmWave band adjacent to mmWave communications bands, the spatial covariance obtained at a passive radar receiver can be sufficiently similar to the spatial covariance of the communication link, and therefore, it can be utilized to reduce the training overhead required to configure the link, as established in prior work [7].

This chapter proposes the use of a RSU equipped with a passive radar to aid in establishing MU-MIMO communication links at mmWave. The passive radar senses the transmitted FMCW signals from multiple automotive radars at once, isolates the signal from each individual vehicle, estimates the spatial covariance of the individual FMCW signals, and then predicts the mmWave communication spatial covariance. This predicted communication covariance is then used to extract the main channel directions and select the beams that will act as analog precoders and combiners at the RSU, significantly reducing training overhead.

1.1.1 Contributions

The main contributions of this chapter are as follows:

- We propose to leverage the spatial covariance information obtained with a passive radar receiver at the RSU to configure the different mmWave communication links between the RSU and different vehicles on the road.

- We propose a passive radar processing chain that uses a filter bank architecture to isolate the individual FMCW signals from a reception containing multiple interfering FMCW signals from different vehicles. These isolated signals are then used to estimate the individual spatial covariances corresponding to the radar transmissions coming from different vehicles.

- We design multiple deep learning architectures to predict the communication spatial covariance from an estimated noisy radar spatial covariance. These neural networks learn the intricate relations and differences between the spatial covariances in the radar and communication bands. Three variations of neural networks are proposed to predict different functions useful for beamformer design: APS prediction, eigenvector prediction, and covariance vector prediction.

- We create a ray tracing simulation of the radar-aided MU vehicular communication scenario to evaluate the performance of the proposed systems in terms of both the sum-rates and outage probabilities of different beam training strategies. This setup is consistent with the 3GPP V2X evaluation methodology for vehicular communication systems, and emulates the deployment of automotive radars in commercial vehicles. The obtained radar-communication channels dataset is made available to the research community to enable further work on radar-aided communication at mmWave [17].

1.1.2 Prior Work

Out-of-band information to aid mmWave communication [16] can come from several sources, including sub-6 GHz systems [18–20], or sensors such as radar [7, 21–23], light detection and ranging (LiDAR) [24, 25], inertial-measurement-units [26, 27], or position information [28–30].

Different approaches that exploit sub-6 GHz signals have been proposed to reduce the beam training time or to estimate the spatial covariance, which is later used to design the beamformers. mmWave link configuration assisted by sub-6 GHz systems has, however, many limitations. The use of sub-6 GHz information in [18] is restricted to LOS channels. The strategies in [19, 20] are applicable to NLOS channels, but require that both the mmWave and sub-6 GHz channels have identical states (both LOS or both NLOS), which does not hold in general.

Position information extracted from a global positioning system (GPS) has also been used in different ways to reduce the overhead of mmWave link configuration. For example, inverse fingerprinting learns a subset of location-dependent beam-pairs based on past measurements in similar locations, such that with a high probability at least one of the vectors in the subset works well [31–33]. A more sophisticated version of this idea exploits the sparsity of the MIMO channel to also provide beam recommendations in new positions where channel measurements are not available [30]. Further reductions in overhead can happen if there is also knowledge of other connected vehicles (which may have different sizes and act as blockages) [34] or other context [35] information. Beam-tracking for automotive vehicles aided by inertial measurement unit (IMU) was proposed in [26, 27]. The common limitation of all these approaches is that they only target LOS scenarios. A mmWave communication system aided by LiDAR is described in [24]. It considers two scenarios: the LiDAR is located at the BS, or LiDAR data from two neighboring vehicles are fused. In [25, 36], different learning strategies predict V2I beam selections using precomputed spatial information collected from LiDAR sensors. The main limitation of these systems is that they are designed to operate only in LOS propagation. The first work that proposed leveraging a radar sensor to aid millimeter wave link configuration considers an active radar at the RSU [21] that transmits RF signals toward vehicles to estimate their spatial covariance from the reflected echoes. This is also the first study that experimentally shows there is a similarity between the angular information extracted from the radar and the communication spatial covariances, even when the center frequencies of operation are different. In [22], a dual function radar and communication system was proposed for simultaneously sensing vehicles and establishing the communication link aided by the sensing information. Position information obtained with a radar unit at the road infrastructure was also used in [37] to reduce the overhead of the beam training protocol. The accuracy of position information provided by radar is higher than that provided by GPS, which leads to a larger reduction in communication overhead when exploiting position information provided by a radar sensor than when leveraging GPS-based position, as shown in the field measurements provided in [38]. The high accuracy of radar

locations was also exploited in [23] in the context of vehicle-to-vehicle (V2V) communication.

Although all these approaches based on an active radar provide an interesting reduction of the link configuration overhead, they only perform well in LOS scenarios. An additional limitation is that the allocation of power to active radar sensing may be prohibitive given a power budget at the roadside unit. Alternatively, a passive radar approach was taken in [7], where the RSU senses signals transmitted from automotive radars onboard the vehicles themselves. This solves the power consumption issue and allows NLOS estimation, but the study was restricted to a single-user case without interference from multiple radars. Furthermore, there is an inherent mismatch between the estimated radar covariance and the true communication covariance due to different operation frequencies or different locations of the radars and communication transceivers in the vehicles.

In this chapter, we overcome these limitations by building upon our preliminary work in [39] to add multiuser capabilities and to further refine the covariance estimate by translating the radar covariance to the communication domain. To this aim, we use neural networks which effectively learn the mismatches between radar and communication channels. Although the work in [39] already explores the idea of learning mismatches, only a single user scenario and the estimation of the APS are considered.

Notation: We use the following notation throughout this chapter. Bold lowercase $\mathbf{x}$ is used for column vectors, bold uppercase $\mathbf{X}$ is used for matrices, non-bold letters x, X are used for scalars. $[\mathbf{x}]_i$ and $[\mathbf{X}]_{i,j}$, denote i-th entry of $\mathbf{x}$ and entry at the i-th row and j-th column of $\mathbf{X}$, respectively. We use the serif font, e.g., $\mathbf{x}$, for the frequency-domain variables. Superscript $^{\mathsf{T}}$, $*$ and $\dagger$ represent the transpose, conjugate transpose and pseudo inverse, respectively. $\mathbf{0}$ and $\mathbf{I}$ denote the zero vector and identity matrix respectively. $\mathcal{CN}(\mathbf{x},\mathbf{X})$ denotes a complex circularly symmetric Gaussian random vector with mean $\mathbf{x}$ and covariance $\mathbf{X}$, and $\mathcal{U}[a,b]$ is a Uniform random variable with support $[a,b]$. We use $\mathbb{E}[\cdot]$ and $\|\cdot\|_{\mathrm{F}}$ to denote expectation and Frobenius norm, respectively. $\mathcal{N}_{\mathrm{APS}}(\cdot;\mathbf{w})$, $\mathcal{N}_{\mathrm{eig}}(\cdot;\mathbf{w})$, and $\mathcal{N}_{\mathrm{col}}(\cdot;\mathbf{u})$ denote the DNNs for APS, dominant eigenvector, and covariance column prediction, where $\mathbf{w}$ and $\mathbf{u}$ are the respective trainable

parameters. $S_{\mathrm{RSU},u}$ and $S_{\mathrm{V},u}$ denote the beam search spaces at the RSU and vehicle for user u. $S_{1\to 2}(L)$ denotes the similarity metric between two APSs within window L.

1.2 System Model

We consider the MU-MIMO V2I communication system represented in Fig. 1.1, where the RSU is located on the side of a roadway and U ego-vehicles are driving along the road with other non-connected vehicles. The RSU is equipped with a passive radar uniform linear array (ULA) and a communications ULA. Both the radar and the communication system operate at mmWave bands. The ego vehicles have 4 ULAs for communications and 4 single-antenna automotive radars. The communication arrays in the vehicles are placed in accordance with 3GPP proposals [40], and the radar arrays are placed at the 4 corners of the vehicle as in many commercial models. The passive radar array at the RSU will use receptions of the automotive radar signals to estimate the radar spatial covariances for each link. These covariances will then be used to configure the MU-MIMO mmWave communication link.

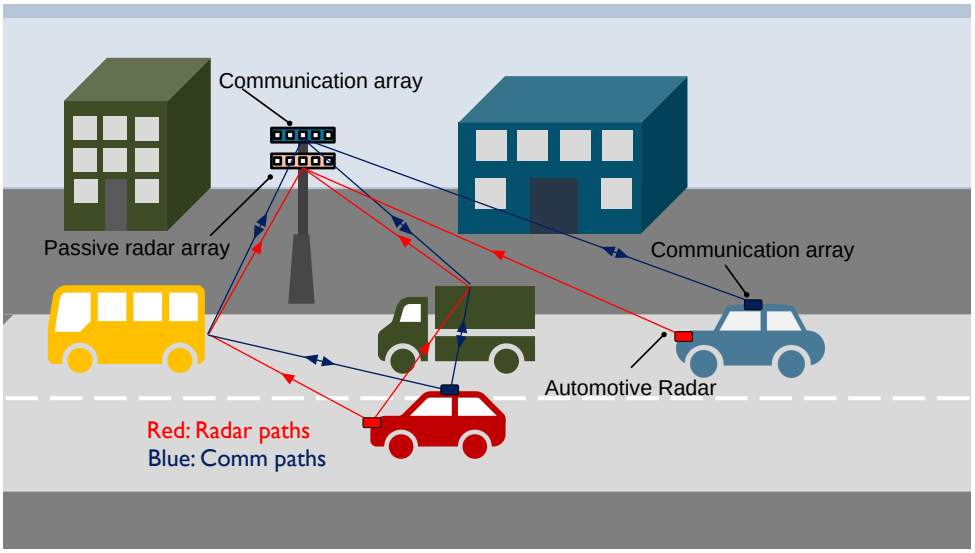


Figure 1.1. A multiuser V2I communication system with the RSU equipped with a passive radar receiver and a communication array operating at mmWave in different bands. The radar array receives automotive radar transmissions (red), while the RSU and vehicles exchange data over the communication channel (blue).

1.2.1 Communication System Model

The communication array on the RSU is equipped with N_{RSU} antennas and $M_{\mathrm{RSU}} \leq N_{\mathrm{RSU}}$ radio frequency (RF)-chains. We let A denote the number of communication arrays at the ego vehicle. Each vehicle array has N_{V} antenna elements and $M_{\mathrm{V}} \leq N_{\mathrm{V}}$ RF-chains. This hybrid architecture supports $N_{\mathrm{s}} \leq \min\{M_{\mathrm{RSU}}, M_{\mathrm{V}}\}$ data streams. The communication link is based on a K sub-carrier orthogonal frequency-division multiplexing (OFDM) system, with modulated symbols $\mathbf{s}[k] \in \mathbb{C}^{N_{\mathrm{s}} \times 1}$ such that $\mathbb{E}[\mathbf{s}[k]\mathbf{s}^*[k]] = \frac{P_{\mathrm{c}}}{KN_{\mathrm{s}}}\mathbf{I}_{N_{\mathrm{s}}}$ and P_{c} denotes the total average transmitted power. The baseband precoder $\mathbf{F}_{\mathrm{BB}}[k] \in \mathbb{C}^{M_{\mathrm{RSU}} \times N_{\mathrm{s}}}$ and RF precoder $\mathbf{F}_{\mathrm{RF}} \in \mathbb{C}^{N_{\mathrm{RSU}} \times M_{\mathrm{RSU}}}$ are combined to form the hybrid precoder $\mathbf{F}[k] = \mathbf{F}_{\mathrm{RF}}\mathbf{F}_{\mathrm{BB}}[k] \in \mathbb{C}^{N_{\mathrm{RSU}} \times N_{\mathrm{s}}}$ on sub-carrier k. The RF precoder is realized using quantized phase shifters and is the same across all subcarriers. Letting $\zeta_{i,j}$, $i = 1, \ldots, N_{\mathrm{RSU}}$, $j = 1, \ldots, M_{\mathrm{RSU}}$, be the quantized phase shift, the RF precoder is described as $[\mathbf{F}_{\mathrm{RF}}]_{i,j} = \frac{1}{\sqrt{N_{\mathrm{RSU}}}} e^{\mathrm{j}\zeta_{i,j}}$. The total power constraint is enforced as $\sum_{k=1}^{K} \|\mathbf{F}[k]\|_{\mathrm{F}}^2 = KN_{\mathrm{s}}$. The baseband combiner $\mathbf{W}_{\mathrm{BB}}^{(a)}[k] \in \mathbb{C}^{M_{\mathrm{V}} \times N_{\mathrm{s}}}$ and RF combiner $\mathbf{W}_{\mathrm{RF}}^{(a)} \in \mathbb{C}^{N_{\mathrm{V}} \times M_{\mathrm{V}}}$, $a = 1, \ldots, A$, are multiplied to form the hybrid combiner $\mathbf{W}^{(a)}[k] = \mathbf{W}_{\mathrm{RF}}^{(a)}\mathbf{W}_{\mathrm{BB}}^{(a)}[k] \in \mathbb{C}^{N_{\mathrm{V}} \times N_{\mathrm{s}}}$ on sub-carrier k. In the simulations we will assume that the number of streams and RF chains is equal to the number of vehicles supported by the RSU, and that the analog precoder/combiner corresponding to the link between the RSU and the uth vehicle is denoted as $[\mathbf{F}_{\mathrm{RF}}]_u / [\mathbf{W}_{\mathrm{RF}}]_u$, $u = 1, ..., U$.

The $N_{\mathrm{V}} \times N_{\mathrm{RSU}}$ frequency-domain MIMO channel at array $a \in A$ in vehicle u is denoted as $\mathbf{H}_u^{(a)}[k]$. Assuming perfect synchronization, the received signal on sub-carrier k after processing is

$$\mathbf{y}_u^{(a)}[k] = \mathbf{W}^{(a)_u *}[k]\mathbf{H}_u^{(a)}[k]\mathbf{F}[k]\mathbf{s}[k] + \mathbf{W}_u^{(a)*}[k]\mathbf{n}^{(a)}[k], \tag{1.1}$$

where $\mathbf{n}^{(a)} \sim \mathcal{CN}(\mathbf{0}, \sigma_{\mathbf{n}}^2\mathbf{I})$ is additive white Gaussian noise.

1.2.2 Channel Model

The wideband channel is modeled geometrically with C clusters. Each of the clusters experiences a mean time delay $\tau_c \in \mathbb{R}$, mean AoA $\theta_c \in [0, 2\pi)$, and mean AoD $\phi_c \in [0, 2\pi)$.

Assuming there are R_c paths in each cluster, each path $r_c \in [R_c]$ has complex gain α_{r_c}, relative time-delay τ_{r_c}, relative arrival angle shift ϑ_{r_c}, and relative departure angle shift φ_{r_c}. The array response vectors are $\mathbf{a}_{\mathrm{RSU}}(\phi)$ at the RSU and $\mathbf{a}_{\mathrm{V}}(\theta)$ at the ego-vehicle. The uniform spacing between array elements is Δ, normalized to units of wavelength. The RSU response vector and ego-vehicle response vectors are defined as

$$\mathbf{a}_{\mathrm{RSU}}(\theta) = [1, e^{\mathrm{j}2\pi\Delta\sin(\theta)}, \cdots, e^{\mathrm{j}(N_{\mathrm{RSU}}-1)2\pi\Delta\sin(\theta)}]^{\mathsf{T}}. \tag{1.2}$$

$$\mathbf{a}_{\mathrm{V}}(\phi) = [1, e^{\mathrm{j}2\pi\Delta\sin(\phi)}, \cdots, e^{\mathrm{j}(N_{\mathrm{V}}-1)2\pi\Delta\sin(\phi)}]^{\mathsf{T}}. \tag{1.3}$$

We remove the notation (a) in the channel $\mathbf{H}$ for the following equations. We define the analog filtering and pulse shaping effect at delay τ as $p(\tau)$. T_{c} denotes the signaling interval. The delay-d MIMO channel matrix $\mathbf{H}_u[d]$ is [41]

$$\mathbf{H}_u[d] = \sum_{c=1}^{C} \sum_{r_c=1}^{R_c} \alpha_{rc,u} p(dT_{\mathrm{c}} - \tau_{c,u} - \tau_{r_c,u}) \times \mathbf{a}_{\mathrm{V}}(\phi_{c,u} + \varphi_{r_c,u}) \mathbf{a}^*_{\mathrm{RSU}}(\theta_{c,u} + \vartheta_{r_c,u}). \tag{1.4}$$

If there are D delay-taps in the channel, the channel at sub-carrier k, $\mathbf{H}[k]$ is [41]

$$\mathbf{H}_u[k] = \sum_{d=0}^{D-1} \mathbf{H}_u[d] e^{-\mathrm{j}\frac{2\pi k}{K}d}. \tag{1.5}$$

1.2.3 Covariance Model

We define the spatial covariance at the RSU for the link with the uth vehicle on sub-carrier k as $\mathbf{R}_{\mathrm{RSU},u}[k] = \frac{1}{N_{\mathrm{V}}}\mathbb{E}[\mathbf{H}^*_u[k]\mathbf{H}_u[k]]$. Assuming the covariance does not change across sub-carriers [42], we can create an estimate by averaging over all sub-carriers $\hat{\mathbf{R}}_{\mathrm{RSU},u} = \frac{1}{K}\sum_{k=1}^{K} \hat{\mathbf{R}}_{\mathrm{RSU},u}[k]$. As we will describe in Sec. 1.2.5, our proposed system uses the covariance estimates to design the RF precoder at the RSU, while training symbols are used to design the baseband precoder, which can account for sub-carrier-dependent covariance variations [7] and multiuser interference.

1.2.4 Radar System Model

Each automotive radar in the environment is assumed to operate in the mmWave band, transmiting a unique FMCW signal. We will assume there are M radars transmitting. The mth radar, for $m \in [M]$, has a chirp rate of β_m, a time offset of Δt_m, and a phase offset of $\Delta\phi_m$. We will assume all radars operate with the same bandwidth B. Then the chirp period is defined as $T_m = \frac{B}{\beta_m}$.

Then the transmitted signal can be defined as

$$s_m(t) = \sqrt{P_{\mathrm{r}}} \exp\left(j2\pi \left(f_r t + \frac{\beta_m t^2}{2} \right) + j\phi_m \right) \quad \text{for } t \in [\Delta t_m, \Delta t_m + T_m]. \tag{1.6}$$

This transmitted signal repeats every T_m seconds. The received signal on the N_{r} element antenna array on the RSU will be denoted as a vector $\mathbf{x}(t) \in \mathbb{C}^{N_{\mathrm{r}}}$. Assume that due to multipath effects, the radar transmission propagates along R_m paths. Each path $r_m \in [R_m]$ experiences an attenuation of α_{r_m} and a time delay of τ_{n,r_m} during propagation to the nth antenna. The received signal at antenna n is

$$[\mathbf{x}(t)]_n = \sum_{m=1}^{M} \sum_{r_m=1}^{R_m} \alpha_{r_m} s(t - \tau_{n,r_m}). \tag{1.7}$$

We can model the propagation delay as the sum of two components: one accounting for common distance τ and another accounting for the difference among antenna elements at the ULA τ'_n. This delay at antenna n is described as $\tau_{n,r_m} = \tau_{r_m} + \tau'_{n,r_m}$ [43]. We assume our ULA has half-wavelength spacing and that the signal from radar m and path r_m arrives at an angle of θ_{r_m}, so

$$\tau'_{n,r_m} = \frac{\sin\theta_{r_m}(n-1)}{2f_{\mathrm{r}}}. \tag{1.8}$$

Then we collect the I samples of the signal into a matrix $\mathbf{Y} \in \mathbb{C}^{N_{\mathrm{r}} \times I}$. Let $i \in \{1, 2, \cdots, I\}$ denote the sample index, and T_{r} denote the sampling time. Then the ith sample on the nth antenna is $[\mathbf{Y}]_{n,i} = [\mathbf{x}(iT_{\mathrm{r}})]_n$. The spatial covariance of the received radar signal could then be estimated as $\hat{\boldsymbol{R}} = \frac{1}{I}\mathbf{Y}\mathbf{Y}^*$.

However, this covariance estimate is not particularly useful when multiple vehicular radars are transmitting. The covariance will contain all of the interfering signals from all M vehicular radars that are transmitting. As a result, the APS computed from such an estimate may be dominated by the contributions from higher SNR signals while the contributions from low SNR signals are undetectable due to the interference and sidelobes. A more useful covariance would be the isolated covariance of each transmitting radar signal. Let us define the true spatial covariance for signal m. This will be done by propagating a unit power Dirac-delta signal through the radar channel from radar m, i.e.

$$[\check{\mathbf{x}}_m(t)]_n = \sum_{r_m=1}^{R_m} \alpha_{r_m} \delta(t - \tau_{n,r_m}). \tag{1.9}$$

Then perform the same sampling and covariance estimation as before:

$$[\check{\mathbf{Y}}_m]_{n,i} = [\check{\mathbf{x}}_m(iT_\mathrm{r})]_n, \tag{1.10}$$

$$\boldsymbol{R}_m = \frac{1}{I} \check{\mathbf{Y}}_m \check{\mathbf{Y}}_m^*. \tag{1.11}$$

$\boldsymbol{R}_m$ is our ideal isolated spatial covariance from the radar signal m. Section 1.3.1 describes our designed signal processing chain that creates estimates $\hat{\boldsymbol{R}}_m$ of the isolated spatial covariance from the total received signal $\mathbf{x}(t)$.

For this chapter, we assume that the communication links and automotive radars operate in separate bands where mutual interference between the two bands is negligible. This assumption is supported by the study in [44], which describes experimental results that show that the radar-to-communication interference can be neglected in this context.

1.2.5 Link Configuration

We assume that the link configuration system at the RSU receives the estimated radar covariance for the link with each vehicle from the passive radar receiver. Then, an additional

processing stage translates these estimated radar covariances into the averaged communication covariance $\hat{\mathbf{R}}_{\text{RSU},u}$ for the link with the uth vehicle in initial access. Note that the radar and communication covariances are similar, but there are mismatches due to, for example, the different operation frequencies or different locations of the radar and communication antenna arrays. Section 1.3.2 includes our proposed solutions for this stage.

Once the estimate of the averaged communication covariance is available, the RF precoder phase shifts are configured based on a standard beam search over a restricted codebook. The subset of beams in this restricted codebook is selected as those around the direction of maximum power in the azimuth power spectrum (related to the communication covariance through the Fourier transform). The number of beams in the restricted codebook is a parameter of the system. The RF combiner is defined at the vehicle using a beam search over a full codebook. Note that overhead reduction in the beam training stage is achieved by the significant reduction in the size of the restricted codebook used at the RSU.

To obtain the digital precoders and combiners after the RF stage has been designed, a practical system could use different algorithms that can provide an estimation of the beamformed channel (propagation channel plus RF stage) based on the transmission of a sequence of training symbols. From the estimated channels, different state-of-the-art solutions for the frequency selective digital precoders and combiners are also available [45, 46]. In our simulations in Sec. 1.4, we show the performance of the analog counterpart of the proposed system, since our only goal is to show the effectiveness of the radar-aided beam training strategy (the innovation proposed in this work) independently of the approach for digital beamforming.

1.3 Radar-aided Multiuser Link Configuration

1.3.1 Multiuser Separation and Radar Covariance Estimation

As mentioned, the received signal $\mathbf{Y}$ contains contributions from all transmitting radars in the environment. However, the RSU must now attempt to estimate the individual radar

covariances for each vehicle, which will later be used to assist in RF precoder design for downlink communications. We assume that for vehicles in initial access, the RSU has no knowledge of the automotive radar chirp rate or timing. Furthermore, the RSU may have no knowledge whether or not a new vehicle has entered its area-of-coverage and must be able to detect radar transmissions from new vehicles. As such, we propose a FMCW mixing filter bank processing chain to detect vehicles, suppress interference from other automotive radar signals, and estimate the spatial covariance for each vehicle in the area-of-coverage individually. The FMCW mixing filter bank correlates the received passive radar signals with several FMCW chirps, each with a different chirp rate. This correlation is effectively a matched filter, maximizing the output power of the correlator when the chirp rate of a received signal matches the chirp rate of the filter bank. Individual FMCW signals can be detected from the peaks of the filter bank correlator output. These detections provide an estimate of each FMCW signal's chirp rate and time delay. With knowledge of the chirp rate and time delay, the mixed signals corresponding to each detection can be multiplied by a time-delay correction signal, concentrating the power of the matched signal near DC. A lowpass filter is then applied, filtering out interference from other FMCW signals, thereby isolating the individual radar signals corresponding to each detection. These isolated signals are then used to estimate the spatial covariance of each detected radar.

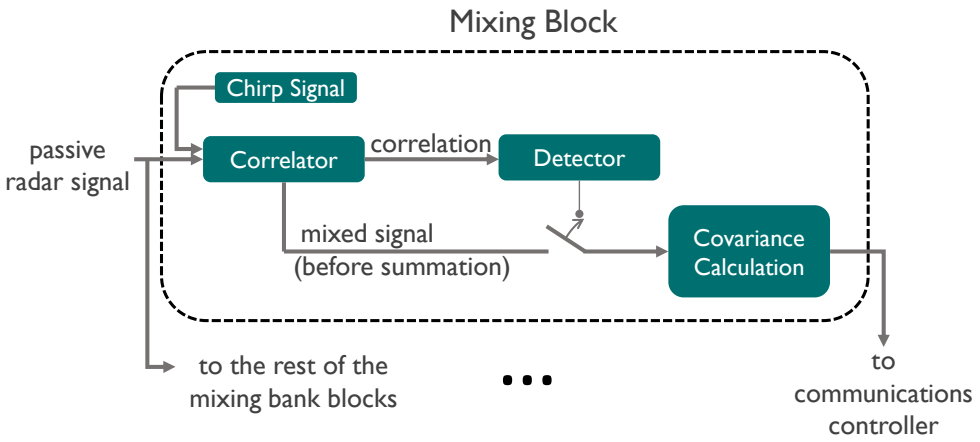


Figure 1.2. Visualization of a single mixing block within the mixing bank. The received signal is mixed with a chirp signal, sampled, and correlated; the detector then identifies any FMCW signal matching the block's chirp rate, after which the covariance is estimated from the mixed signal.

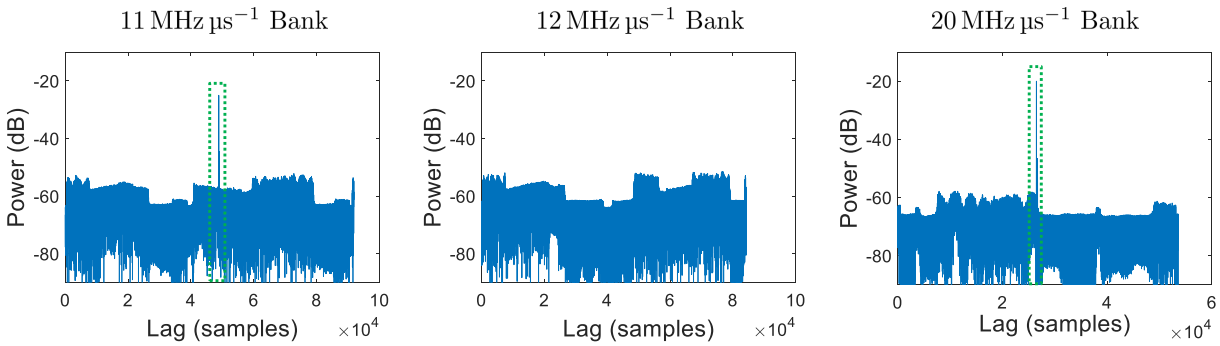


Figure 1.3. The correlator output of 3 mixing blocks with reference FMCW chirp rates of 11, 12, and 19 MHz/μs. Sharp peaks are shown in the first and last output corresponding to passive radar receptions of signals with a matching chirp rate. The second output has no such peak.

Let us consider a single block within the mixing bank, which is visualized in Fig. 1.2. The received signal is mixed with a reference FMCW signal $s_{\text{ref}}(t)$ with the desired chirp rate β_{mix}. Much like the transmitted FMCW signals, the reference FMCW signal has a chirp period of T_{mix} and repeats every T_{mix} seconds,

$$s_{\text{ref}}(t) = \exp\left(-j2\pi\left(f_r t + \frac{\beta_r t^2}{2}\right)\right) \quad \text{for } t \in [0, T_m]. \tag{1.12}$$

The output mixed signal is denoted as $\mathbf{x}_{\text{mix}}(t)$, and can be written as

$$\mathbf{x}_{\text{mix}}(t) = \mathbf{x}(t)s_{\text{ref}}(t). \tag{1.13}$$

This signal after sampling is then called $\mathbf{Y}_{\text{mix}}$, and can be expressed as

$$[\mathbf{Y}_{\text{mix}}]_{n,i} = [\mathbf{x}_{\text{mix}}(iT_r)]_{n,i}. \tag{1.14}$$

Note that the received chirp signals are not time-aligned with the reference chirp signal. The output of this mixer is then sampled and processed digitally.

The next stage in the processing block is the correlator. The sampled signal is then digitally mixed with an offset-correction signal before the I samples are summed together. This

digital mixing and summation is repeated for every lag. The output of the correlator is then passed to a detector. The digital correction signal for lag l is defined as

$$[S_{\text{corr}}]_{l,i} = exp\left(j2\pi \left(\frac{\beta_m (lT_r)^2}{2} iT_r \right) \right). \tag{1.15}$$

The corrected sampled mixed signal at lag l is defined as

$$[\mathbf{Y}_{\text{mix},l}]_{n,i} = ([\mathbf{Y}_{\text{mix}}]_{n,i})([S_{\text{corr}}]_{l,i}). \tag{1.16}$$

And finally, the correlator output is defined as

$$[\mathbf{C}]_{n,l} = \sum_{i=1}^{I} [\mathbf{Y}_{\text{mix},l}]_{n,i}. \tag{1.17}$$

The digital correction signal accounts for the frequency of the reference FMCW signal at the start of lag l. This makes the correlator output equivalent to mixing and summing with an FMCW signal that starts at frequency f_r at every lag l. This saves on hardware complexity by allowing each mixing block to only require mixing with a single FMCW reference signal. The digital correction is then an element-wise complex multiplication. Assuming that the FMCW reference signal is controlled by a voltage-controlled oscillator (VCO), the digital processor can have knowledge of the reference signals frequency at each sample time. When the chirp start and chirp rate of the block align with the chirp start and chirp rate of a received signal, the magnitude of the output of the correlator will exhibit a sharp peak. For chirp rates that do not align, the output of the correlator will have its power spread out over the lag domain. To show this, consider a case where 51 mixing blocks are used in the filter bank, each with a uniformly spaced chirp rate between 10 and 60 MHz/μs. This is visualized in Fig. 1.3, where the correlator output of the same received signals is shown for 3 of the 51 mixing blocks. In the 11 MHz/μs bank and 20 MHz/μs bank, sharp peaks exist because the received signal contains FMCW pulses matched to their mixing chirp rates of 11 and 20 MHz/μs. The 12 MHz/μs bank has no such

peak, because the received signal contained no FMCW pulses with a chirp rate of 12 MHz/μs. In the 12 MHz/μs bank, the power of the chirp detected in the 11 MHz/μs bank becomes spread out over the lag domain. This spreading effects grow as the signal and mixing chirp rates become further separated. In the 20 MHz/μs bank, the chirp detected in the 11 MHz/μs bank is spread out significantly, allowing for the individual FMCW signal matching 20 MHz/μs to be detected. The remaining interference from the spread out signal is not present across the entire lag domain, however, so the detector of these peaks must be able to adapt to changing interference and noise levels. The objective of the detector is to determine whether such a peak exists in the output of the correlator and to remain robust against the interference power from other signals that have different chirp starts and chirp rates.

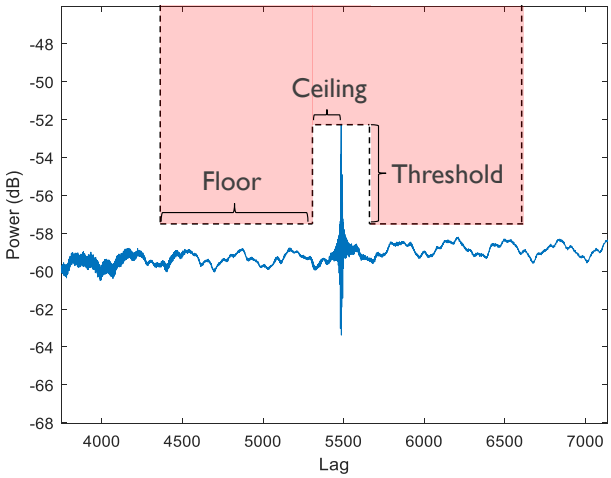


Figure 1.4. A visualization of the max CFAR detector. For each lag, the boundary box defined by the guard cells and floor cells is defined relative to the power at the lag being tested. If the signal does not conflict with the boundary box, a detection is marked at that particular lag.

For our purposes, we propose the use of a max CFAR (constant false alarm rate) detector. The interference power at any given time delay and mixing block may not be known beforehand, especially as vehicles enter and exit the area-of-coverage. As such, an adaptive detection algorithm that estimates the interference power and determines a threshold dynamically is desirable. The max CFAR detector is visualized in Fig. 1.4. Max CFAR estimates the noise and interference power around a cell-under-test (CUT) by taking the maximum power of a set of cells that neighbor

the CUT. A set of guard cells close to the CUT are ignored in this estimation, because power from the CUT may leak into close-by cells. However, we enforce that the power at the CUT is greater than all the powers in the guard cells. In our application, the set of guard cells must account for the delay spread of the radar channel. Once the noise and interference power has been estimated, a detection threshold is determined by multiplying the power by some scaling factor. This scaling factor can be tuned by the system designer to achieve a desired false alarm rate. If the power in the CUT exceeds this threshold, the system considers this a detection. Max CFAR can be implemented efficiently digitally with an FPGA (see [47] for example).

Let the N_{guard} be the number of guard cells, N_{floor} be the number of cells beyond the guard cells that are used to estimate the noise and interference power, and P_{det} be the power threshold for detection. Define the set of guard cell offsets as $\Delta L_{\text{guard}} = [-N_{\text{guard}}, 1] \bigcup [1, N_{\text{guard}}]$. Define the set of floor cell offsets as $\Delta L_{\text{floor}} = [-N_{\text{guard}} - N_{\text{floor}}, -N_{\text{guard}}] \bigcup [N_{\text{guard}}, N_{\text{guard}} + N_{\text{floor}}]$. Assume that the correlator output at lag l is our CUT. We decide to take the maximum power across antennas in these detection steps as well. The CUT power P_{CUT} is defined as

$$P_{\text{CUT}} = \max_{n} |[\mathbf{C}]_{n,l}|^2. \tag{1.18}$$

The noise and interference estimate is defined as

$$P_{\text{floor}} = \max_{n} \max_{\Delta l \in \Delta L_{\text{floor}}} |[\mathbf{C}]_{n,l+\Delta l}|^2. \tag{1.19}$$

The power in the guard cells is also defined as

$$P_{\text{guard}} = \max_{n} \max_{\Delta l \in \Delta L_{\text{guard}}} |[\mathbf{C}]_{n,l+\Delta l}|^2. \tag{1.20}$$

The CUT is then marked as a detection if the following conditions are met:

$$\begin{aligned} P_{\text{CUT}} &> P_{\text{guard}}, \\ P_{\text{CUT}} &> P_{\text{det}} P_{\text{floor}}. \end{aligned} \tag{1.21}$$

Let D be the set of lags where detections are found. The sampled output of the mixer corresponding to these detected lags, $[\mathbf{Y}_{\text{mix},l}]_{n,i} \forall l \in D$, is passed to the covariance estimator. It should be clarified that the samples passed to the covariance estimator are samples before the summation operation in the correlator. By detecting the correct time delay and chirp rate of a particular FMCW reception and then mixing it with a reference chirp corresponding to these exact parameters, we experience a power gain in this particular FMCW signal and a suppression of the other interfering FMCW signals. Furthermore, this power is concentrated in spikes near DC since the correct lag and time delay have been estimated. Therefore, a lowpass filter with a bandwidth B_f greater than the multipath spread of the propagation channel can be applied to the detected signals to filter out the interfering radar transmissions.

$$\hat{\mathbf{Y}}_l = \text{lowpass}\{\mathbf{Y}_{\text{mix},l}\} \tag{1.22}$$

Assuming the detection is accurate, this process isolates the received FMCW signals from each vehicle, allowing the RSU to estimate each vehicle's radar spatial covariance separately. These spatial covariances are then subsequently passed to the communications controller, which will establish a link with the newly detected vehicles in the area-of-coverage.

1.3.2 Radar-to-Communication Covariance Mapping

Background and Motivation

While radar and communication channels are inherently different due to variations in operating frequencies, antenna geometries, antenna array sizes, and physical antenna placement on both the BS and vehicles, they share a similar propagation environment. Consequently, their

spatial covariance exhibits similarities, particularly as both operate in the mmWave band with receivers typically separated by distances smaller than the vehicle size. Previous research, such as in [7], directly utilized radar spatial covariance as an estimate for the communication covariance without explicitly addressing the mismatch between radar and communication channels. This mismatch presents substantial challenges, and developing an analytical model of the mismatch between radar and communication covariances is unrealistic, particularly in NLOS scenarios, due to the complexity and a large number of parameters involved.

To mitigate this mismatch and efficiently establish communication links, we propose leveraging DNNs that explicitly learn the relationship between radar and communication spatial covariances from a suitably constructed dataset. Specifically, the neural network takes the radar spatial covariance estimates as input and outputs predicted communication spatial covariances. The communication controller subsequently identifies the direction of maximum power in the APS derived from the predicted covariance. It then selects a reduced set of beams from its codebook, focusing on those closest to this direction of maximum power. A standard beam search is then conducted only within this restricted subset. With this approach, the RF precoder phase shifts are configured according to the selected subset of beams, while the baseband precoder is adjusted using training symbols. By significantly reducing the size of the beam search codebook, fewer training blocks need to be transmitted, substantially lowering the beam-training overhead and enhancing system efficiency.

Proposed Covariance DNN Estimators

As per the previous section, the detector outputs a set of sampled post-mixing signals $[\mathbf{Y}_{\text{mix},l}]_{n,i} \forall l \in D$ that are filtered out to obtain $\hat{\mathbf{Y}}_l$. The radar spatial covariance can be estimated independently for each of these signals

$$\hat{\boldsymbol{R}}_l = \frac{1}{I} \hat{\mathbf{Y}}_l \hat{\mathbf{Y}}_l^*. \tag{1.23}$$

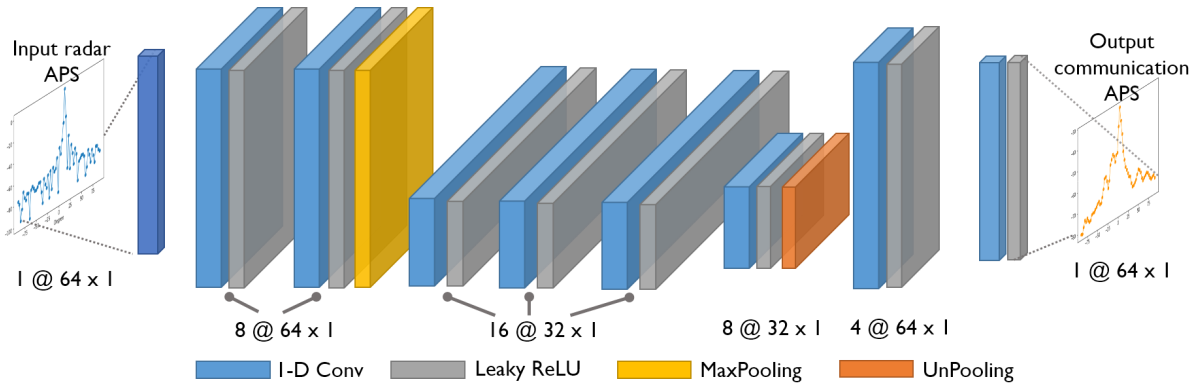


Figure 1.5. The neural network architecture for APS predictions.

The filtering and mixing results in a large gain in the SIR, but some interference power may still remain. Assuming that detection l corresponds to the vehicle our system is establishing a communication link with, The controller must now estimate $\mathbf{R}_{\text{RSU}}$ from the noisy radar covariance estimate $\hat{\boldsymbol{R}}_l$. A neural network will be trained and applied to handle this mapping between the radar and communication domains. However, direct prediction of the complete covariance matrices ignores much of the spatial structure and requires unreasonably high dimensional inputs and outputs. Instead, we train neural networks to input and predict three lower-dimensional features related to the covariance: 1) the APS, 2) the dominant eigenvector, and 3) the covariance vector, which is obtained as the first column of the covariance matrix.

APS prediction: The APS gives the distribution of power as a function of the azimuth angle. The beam search space can be reduced by referring to the angles corresponding to the peaks in the communication APS. In this work, we first acquire both the radar and communication spatial covariance matrices and then extract the APS from the covariance matrix [7]. Specifically, we first construct a DFT matrix $\mathbf{F} = [\mathbf{f}(\theta_1), ..., \mathbf{f}(\theta_N)]$, where $\mathbf{f}(\theta_i) = [e^{-j\cdot 0\cdot 2\pi\Delta\sin(\theta_i)}, ..., e^{-j\cdot(N_{RSU}-1)\cdot 2\pi\Delta\sin(\theta_i)}]^{\mathsf{T}}$ and θ_i is uniformly spaced from $-\frac{\pi}{2}$ to $\frac{\pi}{2}$. Then, the APS can be obtained by extracting the diagonal elements of $\mathbf{F}^*\mathbf{R}\mathbf{F}$, i.e., $\mathbf{d} = diag(\mathbf{F}^*\mathbf{R}\mathbf{F})$. A DNN is designed to predict the communication APS $\hat{\mathbf{d}}_c$ based on the radar APS $\mathbf{d}_r$ to assist beam search. As the APS has obvious local features and the APS exhibits an identifiable shape reflecting the power distribution over directions, and the DNN for APS prediction is mainly

based on 1-D convolutional layers for feature extraction. In this work, we use $N = N_{\text{RSU}}$, which satisfies the Nyquist sampling of the array response, and accordingly $\mathbf{d}_r(\mathbf{d}_c) \in \mathbb{R}^{1\times N_{\text{RSU}}}$. We design a DNN, denoted as $\mathcal{N}_{\text{APS}}(\cdot)$, for R2C APS prediction, which aims to approximate the true communication APS with the known radar APS, i.e.,

$$\hat{\mathbf{d}}_c = \mathcal{N}_{\text{APS}}(\mathbf{d}_r; \mathbf{w}), \tag{1.24}$$

where $\mathbf{d}_r$ and $\hat{\mathbf{d}}_c$ are the given radar APS and predicted communication APS, and $\mathbf{w}$ represents the network parameters of $\mathcal{N}_{\text{APS}}(\cdot)$ to be trained. As the APS is the power distribution over all the directions of arrival, we can observe peaks at directions where the energy is highly concentrated. This constitutes a local feature that we want to capture when linking the radar and communication APS through $\mathcal{N}_{\text{APS}}(\cdot)$, based on the fact that radar and communication channels share the same propagation environment. We apply 1-D convolutional layers in the network, since convolutional layers are well suited for capturing local features. In addition, as the components in the APS are real and non-negative values, we use LeakyReLU [48] as the activation function in the network, which helps to avoid both dying ReLU and vanishing gradient problems during the training process. The detailed architecture of the network is shown in Fig. 1.5. It is like an Encoder-Decoder network [49] where the MaxPooling layers reduce the input dimension and extract the local features of the radar APS efficiently, while the UpPooling layers are used to reconstruct the communication APS based on the extracted features. As the target for the prediction is to approximate the true communication APS for beam training, the mean squared error (MSE) between the predicted communication APS and the true one is considered as the loss function. The loss on the validation dataset is monitored during the training process for the purpose of early stopping against overfitting. In implementation, the radar APS will be computed and input into the trained neural network, which outputs a prediction of the communication APS. If the DFT beamforming matrix is not oversampled, this reduces the dimensionality of the neural network input and output from N^2 to N.

Eigenvector prediction: An alternative approach consists of using the dominant eigenvector of the estimated radar covariance to predict the dominant eigenvector of the communication covariance. Similar to APS prediction, it reduces the dimensionality of the neural network input and output from N^2 to N, simplifying the training and implementation of the neural network. Second, it can help further isolate the signal of interest from the remaining interference after isolation and filtering. In our communication protocol, which will be described in detail in Section 1.4.1, only a single stream will be transmitted to each target to ensure each link has the highest possible SNR. Therefore, predicting the dominant eigenvector of the communication covariance will naturally approximate the spatial weights corresponding to this stream. Consider the eigen decompositions $\hat{\boldsymbol{R}}_l = \mathbf{Q}_l \boldsymbol{\Lambda}_l \mathbf{Q}_l^{-1}$ and $\mathbf{R}_{\text{RSU}} = \mathbf{Q}_{\text{RSU}} \boldsymbol{\Lambda}_{\text{RSU}} \mathbf{Q}_{\text{RSU}}^{-1}$, where the columns of $\mathbf{Q}_l$ and $\mathbf{Q}_{\text{RSU}}$ are the eigenvectors of each covariance, and $\boldsymbol{\Lambda}_l$ and $\boldsymbol{\Lambda}_{\text{RSU}}$ are diagonal matrices containing the eigenvalues of each covariance. Let $\mathbf{v}_l$ be the eigenvector in $\mathbf{Q}_l$ corresponding to the greatest eigenvalue in $\boldsymbol{\Lambda}_l$. Let $\mathbf{v}_{\text{RSU}}$ be the eigenvector in $\mathbf{Q}_{\text{RSU}}$ corresponding to the greatest eigenvalue in $\boldsymbol{\Lambda}_{\text{RSU}}$. The neural network $\mathcal{N}_{\text{eig}}(\cdot)$ will take $\mathbf{v}_l$ as input and predict $\hat{\mathbf{v}}_{\text{RSU}}$. Just as the covariance has an APS, we can define the APS d of an eigenvector $\boldsymbol{v}$ as

$$\boldsymbol{d} = |\boldsymbol{F}^* \boldsymbol{v}|^2. \tag{1.25}$$

Fig. 1.6 illustrates the network design for dominant eigenvector prediction. Unlike the APS network, the eigenvector prediction neural network has to handle complex data. The input to the network is a 64 element vector containing the complex values of the input eigenvector. This complex input vector is then converted from complex values to magnitude and phase components. The magnitude and phase components are then stacked together in a single vector of 128 elements. After this restructuring, the vectorized data passes through 5 fully-connected layers containing 128, 256, 512, 256, and 128 activation units each. Each activation unit uses a LeakyReLU activation function with a shape parameter of $\alpha = 0.1$. A dropout layer is also placed after the activations of the 3rd fully-connected layer, with a dropout rate of 50%. The output of the

last layer is then scaled to unit norm before being output. The output data is a 128-element real-valued vector where the first 64 elements correspond to the real components of the predicted eigenvector and the last 64 elements correspond to the imaginary components of the predicted eigenvector. For optimizing this network, the loss function shown in Algorithm 1 is used. This loss function computes the mean squared error between the predicted eigenvector's APS and the true eigenvector's APS. Each APS is computed using a fast Fourier transform (FFT) with a 35 dB Chebyshev windowing function applied.

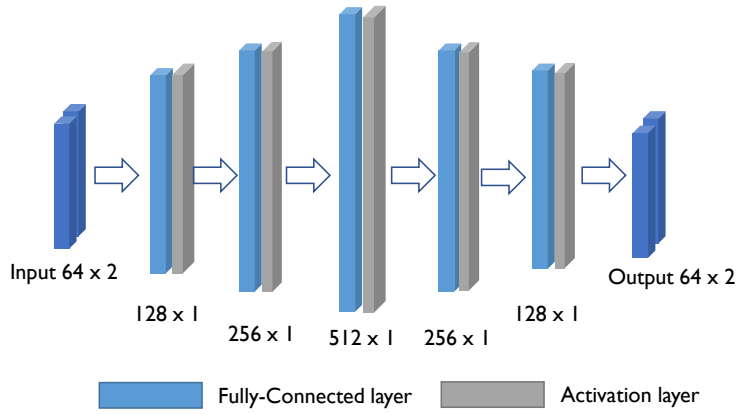


Figure 1.6. The neural network architecture that is trained to predict communication link spatial eigenvectors from radar spatial eigenvector estimates.

Algorithm 1. APS Loss.

Inputs:
$\boldsymbol{v}_{\text{pred}} \in \mathbb{C}^{N\times 1}$ - Predicted eigenvector
$\boldsymbol{v}_{\text{true}} \in \mathbb{C}^{N\times 1}$ - True eigenvector

Begin:
$\boldsymbol{c} \leftarrow$ chebwin(N,35)
$\boldsymbol{z}_{\text{pred}} \leftarrow |\text{FFT}(c \odot \boldsymbol{v}_{\text{pred}})|^2$
$\boldsymbol{z}_{\text{true}} \leftarrow |\text{FFT}(c \odot \boldsymbol{v}_{\text{true}})|^2$
$\text{loss} \leftarrow \frac{1}{N}\sum_{n=1}^{N} |[\boldsymbol{z}_{\text{pred}}]_n - [\boldsymbol{z}_{\text{true}}]_n|^2$
return loss

Covariance prediction: The spatial covariance matrix contains redundant information and it can be characterized by a featured column after Toeplitz completion [7]. The idea is to project the measured covariance matrix to a Toeplitz, Hermitian and positive semi-definite cone.

Thus, the projected matrix $\widetilde{\mathbf{R}}$ can be fully described by its first column $\mathbf{r} \in \mathbb{C}^{N_{\text{RSU}} \times 1}$, i.e.,

$$\widetilde{\mathbf{R}}_l(\widetilde{\mathbf{R}}_{\text{RSU}}) = \arg \min_{\mathbf{X} \in \mathbf{T}_+^N} ||\mathbf{X} - [\hat{\mathbf{R}}_l(\widetilde{\mathbf{R}}_{\text{RSU}}) - \sigma_n^2 \mathbf{I}]||_{\text{F}}, \tag{1.26}$$

where $\widetilde{\mathbf{R}}_l$ and $\widetilde{\mathbf{R}}_{\text{RSU}}$ are the projected covariance matrix for the radar and communication channels, which could be fully represented by their first columns, denoted as $\widetilde{\mathbf{r}}_l, \widetilde{\mathbf{r}}_{\text{RSU}} \in \mathbb{C}^{N_{\text{RSU}} \times 1}$ and called covariance vectors. The DNN designed for covariance prediction outputs the estimation for the column of the communication channel $\hat{\mathbf{r}}_c$ when the input is the column of the radar covariance $\mathbf{r}_r$. Considering that the column is structure agnostic (no special assumption needs to be made about the input), we choose fully connected (FC) layers as the basis for the network. The real and imaginary parts of the column are treated as a 2-channel input for the network, and the output is also 2-channel containing the real and imaginary parts of the predicted communication column. A DNN denoted as $\mathcal{N}_{\text{col}}(\cdot)$ is designed for R2C covariance column prediction. Thus, our proposed covariance column prediction can be described as

$$\hat{\mathbf{r}}_c = \mathcal{N}_{\text{col}}(\mathbf{r}_r; \mathbf{u}), \tag{1.27}$$

where $\mathbf{u}$ contains the network parameters to be trained. The structure of $\mathcal{N}_{\text{col}}(\cdot)$ is different from $\mathcal{N}_{\text{APS}}(\cdot)$. First, the covariance column contains both real and imaginary parts, thus we treat the two parts of $\mathbf{r}_r$ as a 2-channel input for the network. Second, as the covariance column is not like the APS, which has obvious local spatial coherence, the network is mainly built with FC layers, which offer learning features from all the combinations of the features embedded in the input. There are three hidden FC layers in the network, which are sufficient considering the complexity of our prediction problem while making the DL process computational-friendly. Tanh is selected as the activation function, which constrains the values to $[-1, 1]$, since the input can be either positive or non-positive values. The network architecture is shown in Fig. 1.7, where the 2-channel output contains the real and imaginary parts of $\hat{\mathbf{r}}_c$. This projection keeps most

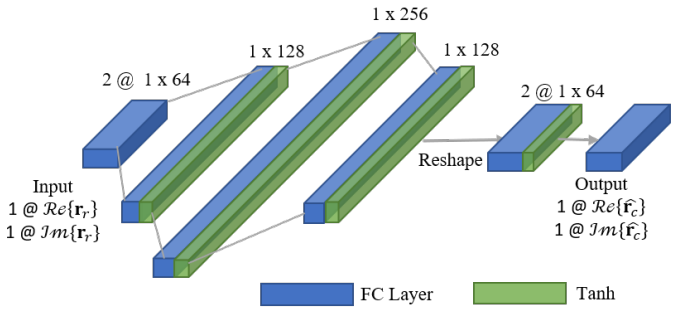


Figure 1.7. An illustration of the architecture of $\mathcal{N}_{\text{col}}$.

of the information of a covariance matrix as it approximates the closest Toeplitz matrix. This method also reduces the dimensionality of the neural network input and output from N^2 to N.

The loss function $\mathcal{L}(\mathbf{u})$ is computed by transforming the column back to the Toeplitz matrix $\widetilde{\mathbf{R}}(\mathbf{r})$, calculating the APS, and evaluating the MSE between the predicted and true APS, i.e.,

$$\mathcal{L}(\mathbf{u}) = \mathbb{E}\{||\mathbf{F}^*\widetilde{\mathbf{R}}(\hat{\mathbf{r}}_c)\mathbf{F} - \mathbf{d}_c||^2\}. \tag{1.28}$$

The training process minimizes the loss on the training dataset, and we monitor the loss on the validation dataset for early stopping.

1.4 Experimental Results

We will now present simulation data demonstrating the utility of multiuser covariance separation, machine-learning based covariance prediction, and our downlink MU-MIMO communication scheme using these new methods to reduce training overhead.

1.4.1 Simulation Setup

To generate our V2I communication and radar channels, ray-tracing simulations were conducted in Wireless Insite [50]. The simulated communication channels operate at 73 GHz and the simulated radar channels operate at 76 GHz. The ray-tracing environment models an urban roadway, which is visualized in Fig. 1.8. In this environment, a mix of medium and large

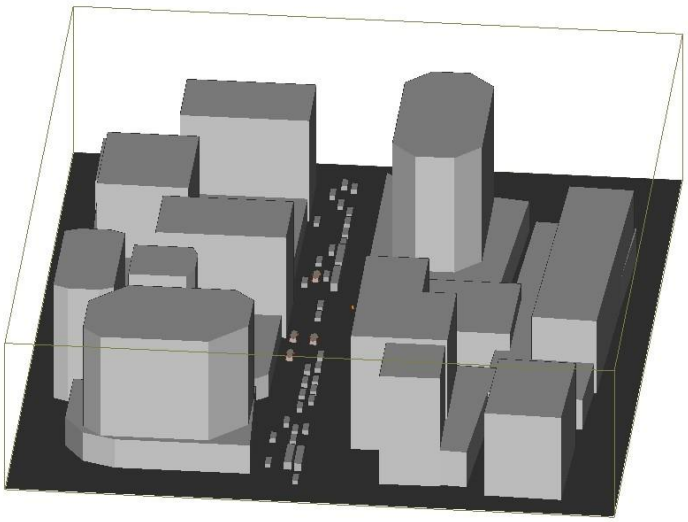


Figure 1.8. The ray-tracing propagation environment. This models an urban roadway with 4 lanes.

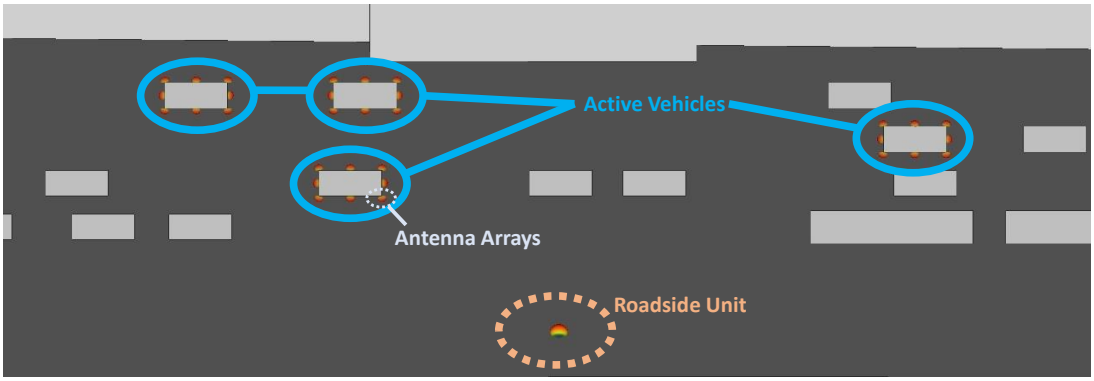


Figure 1.9. An aerial view of the ray-tracing environment that shows the location of the RSU and multiple active vehicles equipped with both radar and communication transceivers.

buildings are located along both sides of the roadway. For simulation purposes, the material of the buildings is assumed to be concrete with a relative permittivity of 5.31, conductivity of $1.0509\,\mathrm{S\,m^{-1}}$ in the communication band at 73 GHz, and a conductivity of $1.0858\,\mathrm{S\,m^{-1}}$ in the radar band at 76 GHz [51, Table 3]. The surface of the roadway is assumed to be asphalt that has a relative permittivity of 3.18, a conductivity of $0.4061\,\mathrm{S\,m^{-1}}$ at 73 GHz, and a conductivity of $0.4227\,\mathrm{S\,m^{-1}}$ at 76 GHz [52]. Root-mean-square (RMS) surface roughness is also modeled as 0.2 mm for concrete and 0.34 mm for asphalt [52, Table 1]. In Wireless Insite, diffuse scattering is

parameterized by a scattering coefficient in the range $[0,1]$, which we select to be 0.4 for concrete and 0.5 for asphalt [53]. The fraction of diffuse reflections that experience cross-polarization is also parameterized with another coefficient in the range $[0,0.5]$, which we select to be 0.5 for both concrete and asphalt. The material for the vehicles is assumed to be a perfect electric conductor metal.

The placement of the vehicles along the roadway is in accordance with option B for Urban scenarios as suggested by 3GPP [40, 6.1.2]. In this setup, we simulate 80% of vehicles being cars of size $5 \times 2 \times 1.6$m and the remaining 20% of vehicles being trucks of size $13 \times 2.6 \times 3$m. We assume the speed of vehicles is dependent upon the lane they are positioned in, and that the spacing between vehicles is exponentially distributed with a mean dependent upon the vehicle speed. The speed of vehicles in each lane are 60, 50, 25, and $15\,\mathrm{km\,h^{-1}}$, denoted s_l. For each vehicle, let $d_l \sim \mathrm{Exp}(0.5/s_l)$. Then the distance from the previous vehicle is given by $\max(2, d_l)$ [40, 6.1.2]. Note that despite ascribing a velocity to each vehicle, each ray-tracing simulation operates in a static environment frozen in time. For each simulation, the type and placement of vehicles are generated randomly and independently according to the above distributions. After generating the vehicle placements, M cars are selected within the area-of-coverage to be active vehicles equipped with radar and communication arrays. In our simulations, $M = 4$ and the area-of-coverage is defined as the 60m section of roadway centered around the RSU.

The active cars are equipped with 4 communication arrays and 4 radar arrays. The communication arrays are placed at the front, sides, and rear of the car at a height of 1.6m. The radar arrays are placed at the 4 corners of the car with 10° rotation toward the front or rear of the car and a height of 0.75m. The antenna patterns for both the communication and radar antenna elements are chosen to have a half-power beamwidth of 120°. The arrays are assumed to be ULA with an inter-element spacing of half a wavelength. The radar and communication channels generated this way are available at [17].

We simulate our communication link with a transmit power of 24 dBm. We assume rectangular pulse-shaping for simplicity. We use $K = 2048$ subcarriers with a subcarrier spacing

of 240 kHz. $D = 512$ time-domain channel taps are used to capture the delay spread of the ray-traced channels. Furthermore, a canonical polyadic (CP) of $D-1$ samples is included in each OFDM symbol. At the RSU, the system has 4 RF-chains and 64 antenna elements. At the vehicle, the system has 1 RF-chain and 16 antenna elements per array. The downlink protocol is visualized in Fig. 1.10. This protocol is modeled after the standalone-downlink (SA-DL) scheme proposed in the beam management tutorial for 3GPP NR. In our protocol, synchronization (SS) blocks are transmitted every channel coherence time. Each SS block spans 100% of the available subcarriers, uses 4 OFDM symbols, and supports up to 4 simultaneous beams in accordance with the number of RF-chains. Tracking (CSI-RS) blocks are transmitted 4 times per coherence time. Each CSI-RS block spans 25% of the available subcarriers, uses 1 OFDM symbol, and also supports up to 4 simultaneous beams. All other resources not occupied with SS or CSI-RS blocks are used to transmit data to the UE vehicles. We evaluate three protocol variants for initial access: exhaustive search, assisted narrow search, and assisted wide search. The exhaustive variant requires all possible beam configurations to be searched over when establishing a link. This corresponds to a typical approach unaided by out-of-band or radar information. The assisted variants use either the direct radar covariance estimates or any of the three neural network predictions. In these assisted variants, the search space of beam configurations is reduced to a set of beams significantly smaller than the full codebook which the exhaustive search uses. Table 1.1 lists the number of RSU beams that are searched over for each variant, as well as the number of SS blocks required to search over that quantity of beams. This can be determined since each SS block supports 4 beams and the UE vehicle needs to search over 16 beams for every RSU. The narrow search only uses 4 beams, while the wide search uses 12 beams. For too small of a search size, the best beam may not be selected. For too large of a search size, the training overhead may become too costly and leave little to no resources available for data transmission.

The FMCW radars are capable of transmitting at chirp rates in the range $[10,60]\,\mathrm{MHz\,s^{-1}}$. Each vehicle independently selects a chirp rate $\beta \sim \mathcal{U}[10,60]\,\mathrm{MHz\,s^{-1}}$. Regardless of the chirp rate, each radar transmits an FMCW waveform with bandwidth of $B_r = 1\,\mathrm{GHz}$ without downtime.

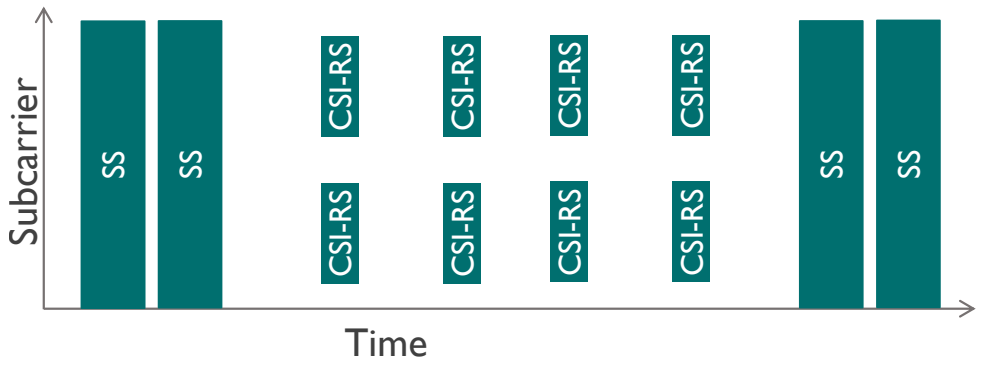


Figure 1.10. The downlink-based protocol. SS blocks use 100% of the subcarriers for training data. The CSI-RS blocks use 25% of the subcarriers for training data to track the channel.

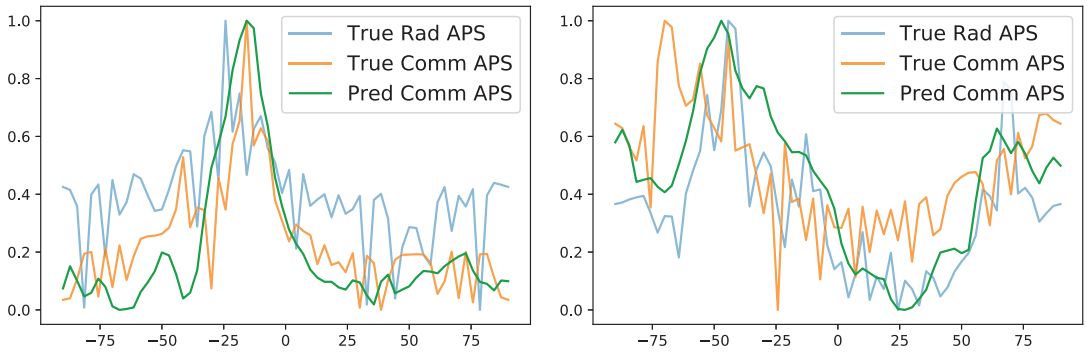
Table 1.1. Required search parameters for the three protocols.

Protocol	**Beam Search Size**	**SS Blocks**
Exhaustive Search	64	256
Assisted Narrow Search	4	16
Assisted Wide Search	12	48

The chirp period is defined as $T_p = \frac{B_r}{\beta}$. A random timing offset $\Delta t \sim \mathcal{U}[0, T_p]$s is also selected for each vehicle. Each vehicle uses this chirp rate and timing offset for all 4 of its radars, assuming that all 4 radars transmit in a synchronized manner. An additional random phase offset $\varepsilon \sim \mathcal{U}[0, 2\pi]$ is added to each transmitted radar signal.

1.4.2 DNN Training and Covariance Predictions

A learning dataset was generated using the true covariances of the communication channels and the estimated radar covariances after detection and filtering. 3000 independent environments were generated, each with 4 covariance pairs corresponding to the 4 active vehicles. This resulted in 12000 unique covariance training pairs. Of the 12000 pairs, 208 pairs were not detected from their radar transmissions and were discarded from the learning dataset, leaving 11792 properly detected pairs. This learning dataset was isolated and kept independent from the evaluation set used for the results shown later. The learning dataset was randomly sampled into a training dataset of 9434 entries and a validation dataset of 2358 entries. For each of the 3



(a) Good APS prediction **(b)** Bad APS prediction

Figure 1.11. Good and bad R2C APS (dB scale) prediction examples with the similarity value of (a) 0.9958 and (b) 0.7290 between the predicted and true communication APS.

neural networks, these covariances were pre-processed into APS's, dominant eigenvectors, and 1st columns of the covariance matrices.

The neural network was trained using the Adam optimizer. Early-stopping was used to halt training after a plateau of 16 training epochs and to restore the best weights that minimized the validation set. The learning rate was also halved after plateaus of 6 training epochs, down to a minimum learning rate of $1e-6$. The training was run using Tensorflow and accelerated using an Nvidia GTX 1060 GPU.

Similarity Performance of the DNN Predictions

The similarity metric [7] compares two APS's $\mathbf{d}_1$ and $\mathbf{d}_2$ within a given window L, i.e., $S_{1\to 2}(L) = \frac{\sum_{i\in\mathcal{I}_1}\mathbf{d}_2[i]}{\sum_{i\in\mathcal{I}_2}\mathbf{d}_2[i]}$, where $\mathcal{I}_1$ ($\mathcal{I}_2$) contains indices of L largest components of $\mathbf{d}_1$ ($\mathbf{d}_2$). This similarity metric rather than the MSE is used to evaluate the prediction accuracy, because the angles corresponding to the largest components of the APS are of particular interest for the beam search process. We evaluate the similarity between the estimated communication APS from the two prediction methods and the true communication APS. As an example shown in Fig. 1.11 with L set to 5, higher similarity (Fig. 1.11a) indicates that the predicted APS and the true APS align better. While the APS prediction method cannot always capture the mismatching in the angle shift between the radar and communication APS (Fig. 1.11b), covariance column predictions

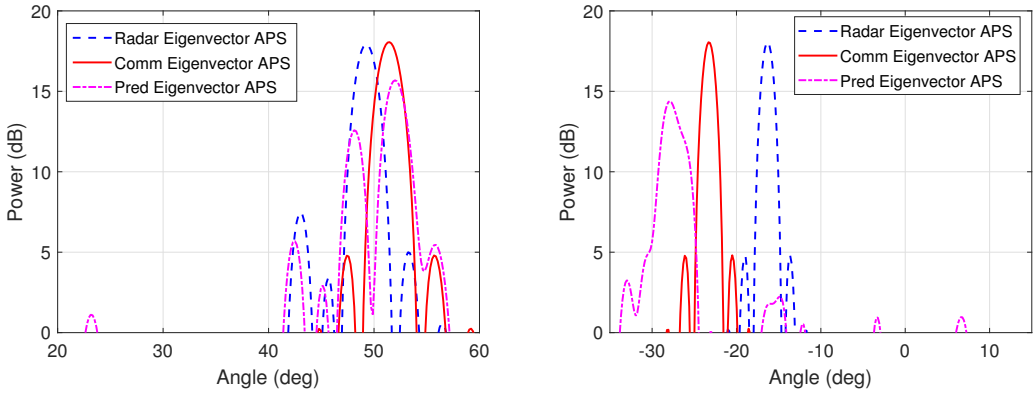


Figure 1.12. Examples of the trained network predicting an eigenvector input from the evaluation dataset: (a) good alignment between the main beam of the communication APS and the predicted eigenvector's APS; (b) some angular error between the peaks of the predicted and communication APS.

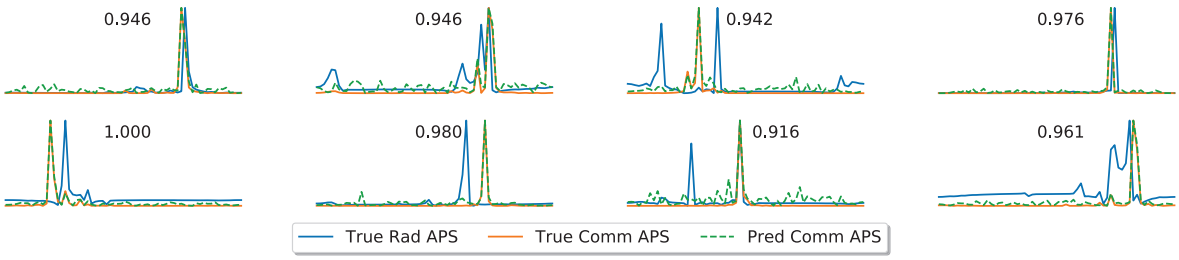


Figure 1.13. Examples of resulted communication APS from R2C covariance column prediction with samples randomly chosen from the test dataset. Column prediction captures the angle shift information, yielding high similarity with the true APS (shown on each subplot).

capture the mismatching and result in accurate communication APS estimations that align with the true APS, as shown in Fig. 1.13, where all the samples are randomly selected from the test dataset.

Fig. 1.14 shows the cumulative distribution function (CDF) of similarity values ($L = 5$) on the test dataset for the two proposed methods, with the similarity between the original radar APS and the true communication APS as the benchmark. The benchmark shows that the radar and the communication APS do not necessarily have high similarity, while our proposed APS prediction improves the similarity performance where the values are generally ≥ 0.6. The covariance

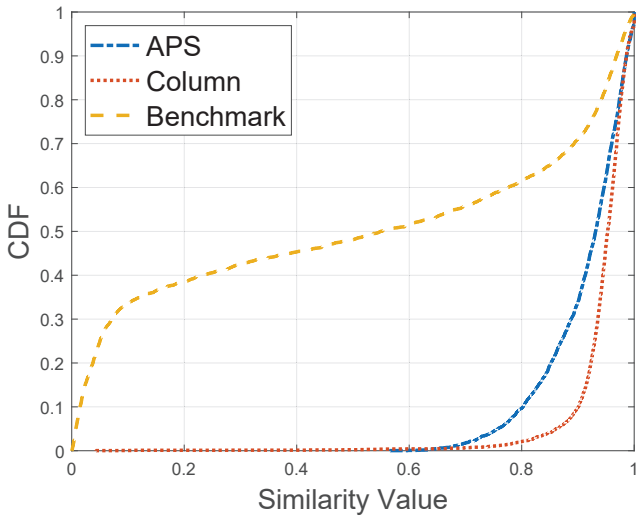


Figure 1.14. CDF of the similarity between the predicted communication APS and the true APS for the test dataset.

column prediction method outperforms the APS prediction method as high similarity occurs with higher frequency; specifically, the 10th-percentile similarity value reaches 0.9. The results are reasonable since for APS predictions, the DNN mainly captures the magnitude properties for the prediction, while the covariance column contains more complete inner features of the covariance matrix including the phase shift information.

1.4.3 mmWave Link Configuration

Using the parameters above, we evaluate the communication link in a Monte Carlo fashion to characterize the performance of the analog beamforming stage, and compare the three proposed network architectures for radar-to-communication mapping. 1000 independent environments were generated. For each environment, one active vehicle was selected at random to be in initial access, while the remaining three vehicles were selected to be in tracking. Channels were labeled as LOS if a direct path with zero reflections existed in the ray-tracing results. Otherwise, the channels were labeled as NLOS. Since the communication link operates in the band from 73 GHz to 74 GHz and the automotive radars operate in the band from 76 GHz to 77 GHz, we assume that spectral leakage is negligible.

As mentioned in Section 1.4.1, three beam training protocols were evaluated: exhaustive search, assisted narrow search, and assisted wide search. Each protocol defines a search space of beams which will be evaluated during the SS transmissions and requires a different number of SS blocks to be sent. Assume the RSU uses the same 2-bit DFT codebook for its $N_{\text{RSU}} = 64$ beams, and that the UE uses the same 2-bit DFT codebook for its $N_{\text{V}} = 16$ beams. Let $\mathcal{Q}(.)$ define the 2-bit quantization function. Then the RSU codebook $\mathbf{C}_{\text{RSU}}$ is defined as

$$[\mathbf{C}_{\text{RSU}}]_i = \mathcal{Q}\left(\frac{1}{N_{\text{RSU}}}\mathbf{a}_{\text{RSU}}\left(\arcsin\left(\frac{2i - N_{\text{RSU}} - 1}{N_{\text{RSU}}}\right)\right)\right) \quad \forall i \in [N_{\text{RSU}}]. \tag{1.29}$$

Similarly, the UE codebook $\mathbf{C}_{\text{V}}$ is defined as

$$[\mathbf{C}_{\text{V}}]_i = \mathcal{Q}\left(\frac{1}{N_{\text{V}}}\mathbf{a}_{\text{RSU}}\left(\arcsin\left(\frac{2i - N_{\text{V}} - 1}{N_{\text{V}}}\right)\right)\right) \quad \forall i \in [N_{\text{V}}]. \tag{1.30}$$

Let the beam search spaces for vehicle u be defined as a set $S_{\text{RSU},u}$ and $S_{\text{V},u}$. Then the best RF-precoder $\mathbf{f}_u$ and RF-combiner $\mathbf{w}_u$ are selected as

$$\{\mathbf{f}_u, \mathbf{w}_u\} = \arg \max_{\mathbf{f}_u \in S_{\text{RSU}}, \mathbf{w}_u \in S_{\text{V}}} \sum_{k=1}^{K} \log_2\left(1 + (\mathbf{w}_u^* \mathbf{H}_u[k]\mathbf{f}_u)^2\right). \tag{1.31}$$

We assume that the vehicles in tracking always select the best pair of RF-precoders and RF-combiners, so for those vehicles $S_{\text{RSU},u} = [N_{\text{RSU}}]$ and $S_{\text{V},u} = [N_{\text{V}}]$. Let the selected RF-precoders and RF-combiners be stacked into matrices such that $[\mathbf{F}_{\text{RF}}]_u = \mathbf{f}_u$ and $[\mathbf{W}_{\text{RF}}]_u = \mathbf{w}_u$. If we assume that $\mathbf{F}_{\text{BB}}[k]$ and $\mathbf{W}_{\text{BB}}[k]$ are diagonal matrices, then the signal power and interference power for the stream to UE u at subcarrier k can be defined as

$$P_{\text{sig},u}[k] = ([\mathbf{W}_{\text{RF}}]_u^* \mathbf{H}_u[k][\mathbf{F}_{\text{RF}}]_u)^2, \tag{1.32}$$

and

$$P_{\text{int},u}[k] = \sum_{l \neq u} ([\mathbf{W}_{\text{RF}}]_l^* \mathbf{H}_u[k][\mathbf{F}_{\text{RF}}]_l)^2. \tag{1.33}$$

Recall that we transmit each stream with a power of $P_t = 24$ dBm. With equal power allocation across all subcarriers, each subcarrier in each stream has a transmit power of $P_t = -9.1$ dBm. Assume a thermal noise of $N_0 = -174$ dBm/Hz, a system noise factor of 10dB, and a subcarrier spacing of $B_{\text{sc}} = 240\,\text{kHz}$. Then the noise power per subcarrier is $P_n = -110.2$ dBm. The signal-to-interference-noise ratio (SINR) can be defined as

$$\text{SINR}_u[k] = \frac{P_{\text{sig},u}[k] P_t}{P_{\text{int},u}[k] P_t + P_n}. \tag{1.34}$$

After computing the SINR, the spectral efficiency can be defined as

$$s_u = \sum_{k=1}^{K} \log_2 (1 + \text{SINR}_u[k]). \tag{1.35}$$

Let T_{train} be the effective time spent transmitting training data from the SS and CSI-RS blocks. "Effective" refers to an average over all subcarriers. Let N_{SS} be the number of SS blocks and $N_{\text{CSI-RS}}$ be the number of CSI-RS blocks sent in the transmitted frame. Also let ν denote the fraction of subcarriers used by CSI-RS blocks. The effective training time is then computed as

$$T_{\text{train}} = T_{\text{sym}} \frac{N_{\text{SS}} N_{\text{sym-per-SS}} + \nu N_{\text{CSI-RS}} N_{\text{sym-per-CSI-RS}}}{N_{\text{beams}}}. \tag{1.36}$$

Now define T_{coh} as the channel coherence time and assume that the communication system repeats its training protocol every T_{coh} seconds. We can then compute the effective rate R_u for each link as

$$R_u = (1 - \frac{T_{\text{train}}}{T_{\text{coh}}}) B_{\text{sc}} s_u. \tag{1.37}$$

The sum-rate is then defined as

$$R_{\Sigma} = \sum_{u} R_u. \tag{1.38}$$

Single-User Rate Results

We consider a single stream transmission ($N_s = 1$) with the transmit power of 24 dBm using 2048 subcarriers. The bandwidth for the transmission is 491.52 MHz. We adopt the codebook design based on 2-bit phase-shifters [19], while the region of the interest is a 180° sector spanning the angles $[-\frac{\pi}{2}, \frac{\pi}{2})$. Thus, the nth codeword in the TX is $\frac{1}{\sqrt{N_v}}\mathbf{a}_v(\arcsin(\frac{2n-N_v-1}{N_v}))$ and the codeword in the RX is defined similarly. We use the rate metric [7] for rate calculation considering a highly dynamic channel with the coherence time of $4N_{RSU}T_{sym}N_v$ (s), where $T_{sym} = 4.7667$ (μs) is the symbol period.

Exhaustive search and radar-only search are treated as the benchmarks to compare with the two methods exploiting DL. For exhaustive search, the training overhead is $N_{RSU} \times N_v$ OFDM blocks, and the optimal beam-pair is determined by which brings the largest absolute value of the received signals over all the sub-carriers. Radar-only search reduces the beam search size at RX from 64 to 12, which finds the best combiner from the top 12 beams whose angles are close to the reference angle corresponding to the largest entry of the radar APS. The beam search size of 12 is found heuristically to maximize the rate by minimizing the likelihood of missing the optimal beam and minimizing the overhead. Similarly, we use a search size of 12 for the APS prediction method, the only difference is the reference angle which corresponds to the largest component of the predicted communication APS instead of the original radar APS. As the covariance column prediction has the best similarity performance, the search size could be further reduced to 2 according to the estimated communication APS from the prediction.

The rate results are fully based on the test dataset. The average rate of the samples in the test dataset is used for performance comparison for each method. As shown in Fig. 1.15, exhaustive search achieves the lowest rate comparing with the other three methods, as it needs to allocate more resources to training blocks. The beam search based on R2C APS prediction

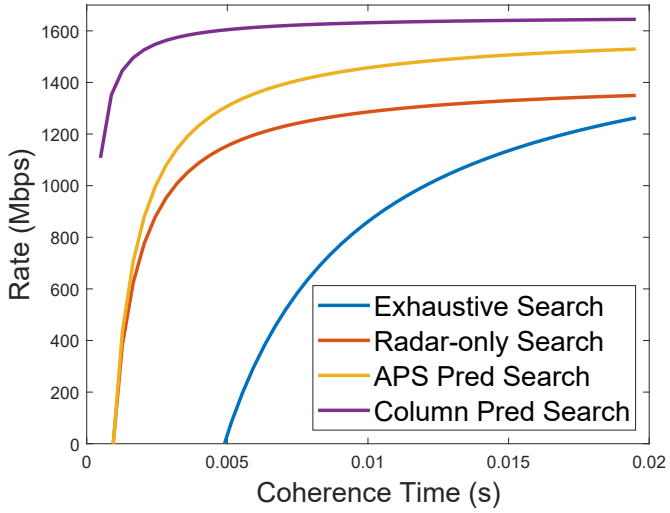


Figure 1.15. Rate results as a function of the coherence time. The beam search size is 12 for radar-only and APS prediction-based beam training and 2 for covariance prediction-based beamtraining.

requires similar training overhead as radar-only search due to the same beam search size, but later it achieves a rate of 1.450 Gbps, which is 200 Mbps (13.3%) higher than the rate achieved by radar-only search. The covariance column prediction method outperforms all other three methods by achieving the final rate of 1.63 Gbps (21.9% higher than radar-only search) with the lowest training overhead.

Linking the rate results to the similarity performance, higher similarity in the APS from the DNN predictions implies more accurate R2C translation, which directly contributes to the efficient beam search.

MU Case Rate Results

The sum-rate results are plotted in Fig. 1.16 for the assisted narrow search and Fig. 1.17 for the assisted wide search. Both plots include the exhaustive search results for comparison. Due to the large training overhead of the exhaustive search, the exhaustive strategy achieves poor sum-rates for short coherence times. Below 0.005 s, the exhaustive strategy requires more training symbols than what can be fit within one coherence interval, resulting in no data

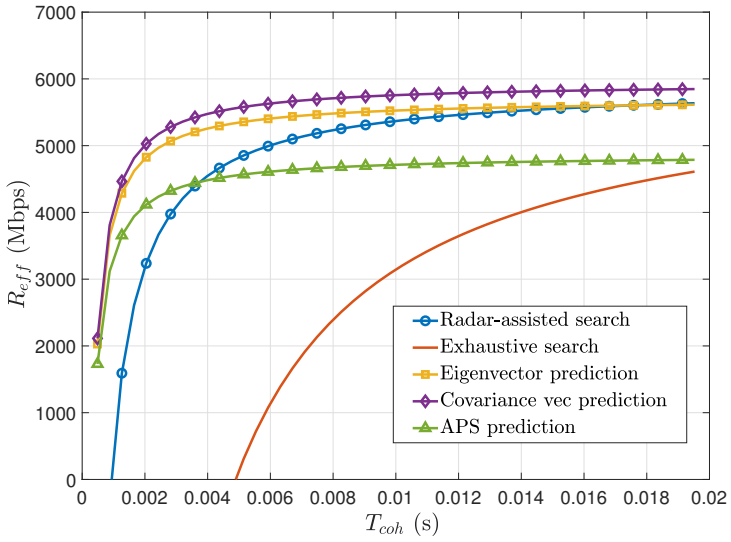


Figure 1.16. The Monte Carlo sum rate results over all 4 users using an exhaustive search and all versions of the assisted narrow search.

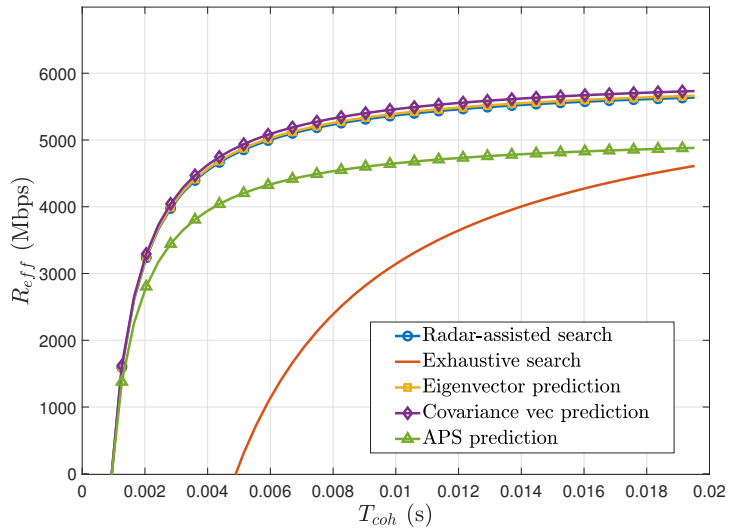


Figure 1.17. The Monte Carlo sum rate results over all 4 users using an exhaustive search and all versions of the assisted wide search.

transmission. Comparatively, the assisted strategies allow the links to be established for even short coherence intervals. In Fig. 1.16, the assisted search based on covariance vector prediction achieves the highest rate for the entire range of coherence times, with the assisted search based on the eigenvector prediction slightly below. The pure radar-assisted search without learning based radar-to-communication mapping has similar performance at long coherence times, but the

performance declines much faster at shorter times. The APS prediction neural network provides the lowest sum-rates of the assisted strategies at longer coherence times, yet outperforms the radar-assisted search at very short coherence times. All of the assisted strategies show significant rate improvements over the exhaustive search, which is infeasible for short coherence times due to the large training overhead required. In Fig. 1.17, all the assisted methods use slightly more training overhead to reduce the likelihood that the optimal beam is missed. As a result, the difference in sum-rates between assisted strategies based on covariance vector prediction, eigenvector prediction, and pure radar-assisted is reduced.

1.5 Conclusion and Future Work

In this chapter, we extended the application of radar-assisted beam training to multi-user communications. We designed a processing chain to estimate the individual spatial covariances of multiple interfering FMCW signals measured at a passive antenna array, and trained a neural network to predict selected features of the communication spatial covariances based on the estimated radar covariances. The proposed approaches were evaluated against a traditional exhaustive beam search strategy and an approach that used the APS of the estimated radar covariance without any further refinement based on a neural network. The features used for neural network prediction include the APS, the dominant eigenvector, and the covariance vector. Ray tracing software was used to generate mmWave radar and communication channels to create a training set for optimizing the neural network, and an evaluation dataset for comparing the beam training methods. Results indicate significant overhead reductions and performance improvements in single-user scenarios: using neural network-based APS and covariance translations yields rate increases of 13.3% and 21.9%, respectively, compared to directly leveraging radar APS without mapping. In MU environments, the proposed radar-assisted methods notably enhance sum-rates over exhaustive search methods. Specifically, the learned covariance vector approach consistently delivers the highest sum-rates and lowest outage probabilities across LOS and NLOS conditions.

These results demonstrate that out-of-band spatial information from passive radars can be leveraged for multi-user systems when special processing is implemented to estimate signal parameters and filter out interference. The intricate differences between radar and communication channel spatial characteristics can be learned through deep learning, yielding more accurate channel predictions even in NLOS channels. Predicting the full communication covariance, rather than a single dominant direction, preserves the rank information that determines how many spatial streams the link can support, making the proposed pipeline compatible with both single-stream beamforming and spatial multiplexing regimes. From a limited-feedback perspective, passive radar sensing replaces the uplink feedback path by providing the RSU with the minimum spatial statistic needed to self-configure the analog precoder.

Future work should explore the detection of automotive radars beyond FMCW, including OFDM and PMCW radar. On the communications side, addressing the problem of radar-aided beam reconfiguration, to track the variations of the vehicular channels after the link has been established, is also an interesting area of research. In particular, blockage prediction from radar tracks can preserve a covariance prior across blockage events, so that beam reselection starts from sensing-derived spatial information rather than from exhaustive search, which is especially relevant in the dense-deployment regime characteristic of mmWave. Finally, integrating the proposed processing chain with a software-defined mmWave RSU and commercial automotive radars in over-the-air experiments is a natural next step to validate the radar-to-communication mapping under hardware impairments not captured by ray tracing.

Acknowledgements

Chapter 1, in part, is a reprint, with permission, of the material as it appears in the papers: Y. Chen, A. Graff, N. González-Prelcic, T. Shimizu, “Radar Aided mmWave Vehicle-to-Infrastructure Link Configuration Using Deep Learning”, in The 2021 IEEE Global Communications Conference (IEEE GLOBECOM 2021), Madrid, Spain, Dec., 2021; A. Graff, Y. Chen, N. González-Prelcic, T. Shimizu, “Deep Learning-based Link Configuration for Radar-aided Multiuser mmWave Vehicle-to-Infrastructure Communication”, in IEEE Trans.

Veh. Technol., doi: 10.1109/TVT.2023.3239227, 2022; and N. González-Prelcic, A. Ali, Y. Chen, "Radar-aided Millimeter Wave Communication", in Signal Processing for Joint Radar Communications, Wiley-IEEE Press, 2024. The dissertation author was the primary investigator and author of these papers, which were supported in part by the NSF under Grant No. 2147955 and by a gift from Toyota Motor North America, Inc., USA.

Chapter 2

Learning to Localize with Attention: from Sparse mmWave Channel Estimates from a Single BS to High Accuracy 3D Location

2.1 Introduction

Wireless networks operating at mmWave bands exploit large arrays and bandwidths, which lead to a high angle and delay resolvability when performing basic functions in the receiver such as channel parameter estimation, either for communication or localization purposes. Unlike at lower frequency bands –where the multipath is dense and becomes an interference for localization– the sparsity of the mmWave channel makes it simpler to map the relevant channel paths to the geometry of the environment [54]. More specifically, the user location can be obtained from high resolution estimates of the multipath components of the channel between the user and a single BS, by exploiting the geometric relationships between the path parameters and the location of the scatterers, the BS (assumed to be known), and the user [54, 55]. This approach has the potential to become a cost-effective alternative for precise localization, required in many envisioned applications such as highly automated vehicles or robot automation in smart factories [56]. In the vehicular setting, a CSI based approach is robust to unfavorable weather or light conditions that may impact methods that exploit onboard automotive sensors, such as

radars, LiDAR, cameras, and IMUs [57–61]. Moreover, it does not suffer from the low accuracy of GNSS in urban scenarios. However, state-of-the-art solutions do not provide the required localization accuracy for some envisioned use cases –for example, an accuracy in the order of 0.1 m in vehicular settings [62]– when evaluated in a realistic propagation environment with a practical mmWave MIMO architecture.

2.1.1 Prior Work

Existing work on localization and channel estimation exploiting a snapshot from a single BS [55, 63–75] can be based on two kinds of methods: 1) Two-stage approaches [55, 63–73], where the first stage focuses on channel parameter estimation, while the second stage has to solve an optimization problem to determine the user location from the channel path parameters, usually exploiting the geometry of the propagation environment. 2) Joint statistical approaches [74, 75], which typically solve the optimization problems leveraging joint probability distributions of the parameters to be estimated. For the first category, the required accuracy of the channel estimation stage for precise localization is higher than for communication, since the estimated channel parameters are introduced into nonlinear geometric transformations very sensitive to estimation errors. Proposed channel estimation methods usually exploit the sparse nature of the mmWave channel, and include on-grid approaches based on variations of OMP such as simultaneous orthogonal matching pursuit (SOMP), distributed compressed sensing simultaneous orthogonal matching pursuit (DCS-SOMP), or others [55, 63–65], off-grid strategies including subspace-based algorithms like multidimensional estimation via rotational invariance techniques (MD-ESPRIT) [66–69] and atomic norm minimization [70], algorithms based on probability models such as the generalized turbo methodology [71] or maximum likelihood estimation (MLE) [72] to name a few. The user location can then be obtained by exploiting geometric relationships which involve path parameters and the known anchor node positions [55, 63, 67–71]. In particular, when the user location has to be determined from the parameters of the channel between the user and a single BS, measurements of the direction-of-arrival (DoA), direction-

of-departure (DoD), and delay are required. If the channel is LOS, the user can be localized if these angular and delay measurements are available for the LOS path and at least one first-order reflection. In the NLOS scenario, the measurements for at least three first-order reflections are required [76]. Methods of the second category consider a joint estimation of the channel and position parameters. The joint probability distributions of these parameters are exploited by methods such as MLE [74] or expectation maximization (EM) to reduce complexity [75]. The main limitation comes from the assumptions of specific joint distributions which may not hold for realistic channels. To understand the limitations of prior work on joint localization and channel estimation at mmWave we focus first on reviewing previous work on channel estimation. Greedy strategies for mmWave channel estimation exploit the sparsity of the channel to obtain the parameters for every multipath component using a dictionary-based approach [77–83]. For example, a greedy approach based on a low-complexity orthogonal matching pursuit (OMP) algorithm that operates with a reduced dictionary constructed by exploiting statistical information of the scatters is proposed in [77]. A parameter perturbed OMP algorithm is provided in [78]. Although a frequency-flat channel model is considered in [77, 78], other prior work [79–83] offers dictionary-based solutions for frequency-selective mmWave channels. The simultaneous OMP (SOMP) algorithm [84] is the core for the solutions developed in [79, 80]. In [79], a joint subcarrier-block-based scheme is proposed using continually distributed angles of arrival and departure, while in [80], simultaneous weighted OMP simultaneous weighted orthogonal matching pursuit (SWOMP) is presented to account for the correlated noise after combining. In [81], the authors focus on compressive channel estimation on the uplink to configure precoders/combiners for the downlink based on channel reciprocity, and develop two algorithms for both purely digital and hybrid architectures. In [82], a multi-layer sparse Bayesian learning (SBL) method for selectively increasing the angle resolution for channel estimation layer by layer helps to reduce the computational complexity and improves the performance. However, all these methods operate in the frequency domain, without estimating the path delays, which are required in any localization strategy that exploits the information about the channel between a user and a single

BS. A time domain channel estimation technique that accounts for the pulse shaping and filtering effect in the received signal, and can identify the path delays as well as angular parameters, was proposed in [83]. However, it suffers from an extremely high complexity when operating with planar arrays and high resolution dictionaries. To overcome this complexity limitation, an alternative approach called MOMP, which also operates in the time domain estimating directions and delays, was recently proposed in [85] and is applied in this chapter (see Sec. 2.4 for details). The main idea behind MOMP is to operate with a multidimensional dictionary and perform the matching operation by independent tensor multiplications along each dimension, to later introduce a refinement stage based on alternating minimization.

Off-grid sparse recovery strategies were also developed in previous work to solve the channel estimation problem at mmWave [86–88]. The method in [86] operates with a nonuniform grid, but it can be applied only to narrowband channels without filtering effects. The key idea in [87] is to approximate a continuous infinite dictionary, acquiring the channel parameters by solving a convex optimization problem, but again, it only applies to the unrealistic case of a narrowband channel without filtering effects. In [88], the authors assumed some specific distributions of the channel parameters, and a SBL-based block EM algorithm is proposed to perform Bayesian inference, assuming a MIMO-OTFS system, neglecting the filtering effects again, and focusing on time-varying channels. Other studies on off-grid mmWave channel estimation that focus on the narrowband case can be found in [89–92], but share the same limitation. An off-grid sparse recovery method is proposed in [93] for frequency selective mmWave channels including the filtering effect, but it operates in the frequency domain and cannot be used to estimate the delay parameters required for single snapshot localization from a single BS. Another category of off-grid methods exploits the ESPRIT algorithm [94, 95], but they also neglect the filtering effects in the channel model.

Methods based on deep learning (DL) have also been recently proposed to estimate the mmWave channel exploiting suitable datasets [96–98]. Different network architectures have been proposed, including a 3D convolutional neural network (CNN) that approximates the sparse

Bayesian learning process [96], a concatenated block architecture based on CNN for extracting the channel coefficients [97], and a FC network for beamspace channel amplitude estimation and channel reconstruction [98]. Strong limitations of all these DL methods are again that they operate in the frequency domain (i.e. the delays are not estimated), the discrete time channel model does not account for the pulse shaping and filtering stages at the receiver previous to analog-to-digital conversion, and the combining process is omitted in [97, 98], which results in the implicit assumption of using a fully digital architecture with high resolution converters, which is not feasible at mmWave.

Apart from using a DNN for channel estimation to subsequently derive the user locations, DNN can be directly applied to map channels to user locations based on channel fingerprinting, where channel characteristics such as reference signal received power (RSRP) [99], CSI [99–101], beamformed fingerprints [102, 103], angle-delay profiles [104], etc., are leveraged. Networks based on CNN architectures are proposed in [99, 101–104], where channel information can be structured as image-like inputs for convolutional layers to extract inherent features associated with user locations. While these methods require a stable environment with static channels, the work in [100] considers dynamic environments and enables robust feature learning via attention schemes. However, these methods assume the availability of perfect channel information, without running into issues related to the computation complexity and inaccurate channel representations. In addition, localization accuracy is compromised when avoiding overfitting. In addition to the previously discussed limitations of most of the previous work on the channel estimation strategy itself, specific work on model-based localization from a snapshot from a single BS exploiting the channel parameters suffers from additional drawbacks: 1) an oversimplified communication system model, which employs a limited number of antenna elements [55, 63, 65, 67, 69–75], or neglects the filtering effects at the receiver [55, 63–75]; 2) assumption of perfect TX-RX synchronization to exploit the time-of-arrival (ToA) [55, 65–68, 70, 71, 73–75]; 3) artificially controlled evaluation settings which lead to simplistic and impractical channels, resulting in the lack of strategies to extract the LOS and first-order NLOS paths [55, 63–75]; 4) high complexity

of the 3D high resolution channel estimation process [55, 63, 65]; 5) unsatisfactory localization accuracy–for example, $\geq$ 10 m [101, 102]–when evaluated with realistic channels.

2.1.2 Contributions

In this chapter, we propose a hybrid model/data-driven strategy for single shot joint localization and channel estimation. The data driven stage has been customized with data corresponding to vehicular channels, but the strategy could be applied to any scenario by using the appropriate datasets. Our approach begins by implementing a low complexity compressive channel estimation technique based on the MOMP algorithm [85, 105], which enables operation in realistic 3D environments. Then, a data driven method using *PathNet* is employed to solve the path classification problem, identifying the necessary LOS and first-order NLOS paths. A new position estimator, which can work in both LOS and NLOS channels with imperfect TX-RX synchronization, is then applied to convert the estimated parameters into the vehicle's 3D location. To further improve the localization accuracy, we introduce a novel strategy for position refinement – a DNN called *ChanFormer* inspired by the *Transformer* architecture [106]. It is used to analyze the estimated paths, evaluate the consistency between the estimated channel and the initial location estimate, and generate a probability distribution of the true position exploited to obtain a more precise location estimate. The main contributions of the paper are as follows:

- We propose a realistic 3D mmWave channel model that includes the effects of the filtering stages at the receiver and the unknown clock offset between the TX and the RX, which needs to be considered when the channel parameters are exploited for localization.

- We develop a two-stage MOMP channel estimation algorithm to reduce the complexity of the high resolution channel estimation process, turning computational burden from a product to a sum of terms. It operates by jointly estimating first, for every path, the DoD in azimuth and elevation, the delay, and a parameter that contains a combination of the DoA information and the complex gain. An additional estimation stage is defined to retrieve the

DoA information. We also include an off-grid channel estimation algorithm, ESPRIT-D which accounts for system filtering effects, as a benchmark for performance comparison.

- We build the lightweight yet effective *PathNet* architecture for classifying the estimated channel paths. The training loss function is formulated to minimize the misclassification of high-order reflections as an LOS or a first-order NLOS path. The network exhibits a strong generalization ability in new environments, achieving a classification accuracy of 99% (with perfect channel knowledge).

- We develop a model-driven location estimator that exploits the channel geometry and can operate in both LOS and NLOS situations, as long as a sufficient number of paths are estimated. It provides sub-meter accuracy for more than 85% of the users in LOS vehicular channels and for 35% of the users in purely NLOS channels.

- We design *ChanFormer*, a network that exploits the concept of "attention" to evaluate which estimated paths are more credible and assess the likelihood of a given location being accurate. A mathematical formulation that models the likelihood based on the straight-line distance to the true location is proposed. *ChanFormer* is intended to refine the location results obtained from the model-driven location estimator.

- We generate a dataset containing realistic vehicular channels together with their associated vehicle positions generated by ray-tracing in an urban environment. All the simulations and evaluations of our algorithms are based on these channels, which are mostly composed of high-order NLOS paths. The dataset is available at [107] and can be used by the research community to evaluate any new solution to the joint localization and channel estimation problem in vehicular channels. Simulation results show that 80% of the users in LOS conditions experience localization errors below 28 cm when exploiting our proposed strategy for localization, while sub-meter accuracy is achieved for 55% of users in NLOS conditions.

Our overall scheme has been built upon our initial design in [76], completing all the details and derivations of the channel estimation strategy, extending the initial datasets for path classification, modifying the model-based initial localization strategy, adding the Transformer-based location refinement stage and including additional numerical experiments and comparisons with prior work.

The rest of the chapter is structured as follows: Sec. 2.2 describes the general V2I communication setup, including the system model and the training strategy for joint channel estimation and localization. Sec. 2.4 develops the different stages of our hybrid model/data-driven approach to joint localization and channel estimation. Then, Sec. 2.5, shows the numerical results of the experiments designed to evaluate the proposed strategy and the comparisons with previous work. Finally, Sec. 2.6 concludes the chapter, summarizing the main results and outlining future research directions.

Notations: Non-bold Italic letters x, X are used for scalars; Bold lowercase $\mathbf{x}$ is used for column vectors, and bold uppercase $\mathbf{X}$ is used for matrices. $[\mathbf{x}]_i$ and $[\mathbf{X}]_{i,j}$, denote i-th entry of $\mathbf{x}$ and entry at the i-th row and j-th column of $\mathbf{X}$, respectively. $\mathbf{X}^*$, $\bar{\mathbf{X}}$ and $\mathbf{X}^\mathsf{T}$ are the conjugate transpose, conjugate and transpose of $\mathbf{X}$. $\|\mathbf{X}\|_\mathrm{F}$ denotes the Frobenius norm of $\mathbf{X}$. $[\mathbf{X},\mathbf{Y}]$ and $[\mathbf{X};\mathbf{Y}]$ are the horizontal and vertical concatenation of $\mathbf{X}$ and $\mathbf{Y}$. $\mathcal{N}(\mathbf{x},\mathbf{X})$ denotes a complex circularly symmetric Gaussian random vector with mean $\mathbf{x}$ and covariance $\mathbf{X}$. $\mathbf{I}_N$ denotes a N-by-N identity matrix. $\mathbb{N}$, $\mathbb{R}$, and $\mathbb{C}$ are the set of natural numbers, real numbers, and complex numbers, respectively. $\mathbb{E}[\cdot]$ denotes expectation. For mathematical calculations, $\mathbf{X}\otimes\mathbf{Y}$, $\mathbf{X}\odot\mathbf{Y}$, and $\mathbf{X}\circ\mathbf{Y}$ are the Kronecker product, Hadamard product, and Khatri-Rao product of $\mathbf{X}$ and $\mathbf{Y}$. $<\mathbf{x},\mathbf{y}>$ is the dot product of $\mathbf{x}$ and $\mathbf{y}$.

2.2 System Model

We consider a mmWave MIMO system where the users are active vehicles either communicating with the BS or in initial access. The BS is equipped with a single uniform

rectangular array (URA), and each vehicle is equipped with 4 URAs facing front, back, right, and left, as suggested by the 3GPP methodology to simulate vehicular channels [40]. The URA at the BS is equipped with $N_{\mathrm{t}} = N_{\mathrm{t}}^{\mathrm{x}} \times N_{\mathrm{t}}^{\mathrm{y}}$ antenna elements and is connected to $N_{\mathrm{t}}^{\mathrm{RF}}$ radio frequency (RF) chains, while each URA on the vehicle has $N_{\mathrm{r}} = N_{\mathrm{r}}^{\mathrm{x}} \times N_{\mathrm{r}}^{\mathrm{y}}$ antenna elements and is connected to $N_{\mathrm{r}}^{\mathrm{RF}}$ RF-chains. We focus on the downlink transmission during initial access, assuming hybrid analog-digital precoding and combining at both ends. We assume that N_s data streams are transmitted, with $N_s \leq \min\{N_{\mathrm{t}}^{\mathrm{RF}}, N_{\mathrm{r}}^{\mathrm{RF}}\}$. The hybrid precoder is defined as $\mathbf{F} = \mathbf{F}_{\mathrm{RF}}\mathbf{F}_{\mathrm{BB}} \in \mathbb{C}^{N_{\mathrm{t}} \times N_s}$, and the hybrid combiner is $\mathbf{W} = \mathbf{W}_{\mathrm{RF}}\mathbf{W}_{\mathrm{BB}} \in \mathbb{C}^{N_{\mathrm{r}} \times N_s}$, where the subscript RF stands for the analog counterpart of the precoder/combiner and BB for the digital one. We consider a fully connected phase shifting network [108].

To develop the 3D channel model we define θ^{x} and θ^{y} as the DoA in azimuth and elevation, while ϕ^{x} and ϕ^{y} represent the DoD also in both dimensions. Note that the azimuth and elevation angles are in the range of $[0, \pi)$ and $[-\frac{\pi}{2}, \frac{\pi}{2})$, respectively. The unitary vectors for the DoA and DoD are given by $\boldsymbol{\theta} = [\cos\theta^{\mathrm{y}} \cos\theta^{\mathrm{x}}, \cos\theta^{\mathrm{y}} \sin\theta^{\mathrm{x}}, \sin\theta^{\mathrm{y}}]^{\mathsf{T}}$, and $\boldsymbol{\phi} = [\cos\phi^{\mathrm{y}} \cos\phi^{\mathrm{x}}, \cos\phi^{\mathrm{y}} \sin\phi^{\mathrm{x}}, \sin\phi^{\mathrm{y}}]^{\mathsf{T}}$. Assuming the arrays are placed in the yz-plane with a half-wavelength element spacing, the array response at the receiver (RX) $\mathbf{a}(\boldsymbol{\theta})$ can be formulated where:

$$[\mathbf{a}(\boldsymbol{\theta})]_{(n_{\mathrm{r}}^{\mathrm{x}}-1)N_{\mathrm{r}}^{\mathrm{y}}+n_{\mathrm{r}}^{\mathrm{y}}} = e^{-j\pi\left((n_{\mathrm{r}}^{\mathrm{x}}-1)\cos\theta^{\mathrm{y}}\sin\theta^{\mathrm{x}}+(n_{\mathrm{r}}^{\mathrm{y}}-1)\sin\theta^{\mathrm{y}}\right)}, \tag{2.1}$$

which can be represented as the Kronecker product of two vectors as $\mathbf{a}(\boldsymbol{\theta}) = \mathbf{a}(\boldsymbol{\theta}^{\shortparallel}) \otimes \mathbf{a}(\boldsymbol{\theta}^{\perp})$ for later multidimensional operations, where $\left[\mathbf{a}(\boldsymbol{\theta}^{\shortparallel})\right]_n = e^{-j\pi(n-1)\cos\theta^{\mathrm{y}}\sin\theta^{\mathrm{x}}}$, and $\left[\mathbf{a}(\boldsymbol{\theta}^{\perp})\right]_n = e^{-j\pi(n-1)\sin\theta^{\mathrm{y}}}$. Similar definitions can be built for $\mathbf{a}(\boldsymbol{\phi})$ as $\mathbf{a}(\boldsymbol{\phi}) = \mathbf{a}(\boldsymbol{\phi}^{\shortparallel}) \otimes \mathbf{a}(\boldsymbol{\phi}^{\perp})$. The effective discrete time baseband channel is seen through the RF front end, so the effects of the filtering stages before analog-to-digital conversion should be included in the channel model. We represent the overall response of the filtering stages by the function f_{p}. The channel matrix for the n-th

delay tap is $\mathbf{H}_n \in \mathbb{C}^{N_\mathrm{r} \times N_\mathrm{t}},\ n = 0, ..., N_\mathrm{d} - 1$, which can be written as

$$\mathbf{H}_n = \sum_{\ell=1}^{L} \alpha_\ell f_\mathrm{p}\left(nT_s - (t_\ell - t_0)\right) \mathbf{a}_\mathrm{r}(\boldsymbol{\theta}_\ell)\mathbf{a}_\mathrm{t}(\boldsymbol{\phi}_\ell)^*, \tag{2.2}$$

where α_ℓ and t_ℓ are the complex gain and the ToA of the ℓ-th path, T_s is the sampling period, and t_0 is the unknown clock offset. Since the channel estimation algorithm in the mmWave receiver will provide an estimate of the relative delay $\tau_\ell = t_\ell - t_0$, $\ell = 1, \ldots, L$, it is convenient to define the channel model as a function of τ_ℓ instead of the absolute delays t_ℓ.

During initial access, training signals are transmitted/received through several pairs of training precoders and combiners to sound the channel and localize the vehicle. We focus on the initial access stage, seeking to realize sub-meter vehicle localization accuracy as a byproduct of the link establishment. Further refinement of the location is possible by exploiting subsequent channel tracking stages, but it is out of the scope of this chapter.

Next, we build the model for the received signal during training. The q-th instance of the training sequence is a vector denoted as $\mathbf{s}[q] \in \mathbb{C}^{N_s \times 1}, q = 1, ..., Q$, satisfying $\mathbb{E}[\mathbf{s}[q]\mathbf{s}[q]^*] = \frac{1}{N_s}\mathbf{I}_{N_s}$. We consider a frequency selective MIMO channel with N_d delay taps. Assuming that the transmitted power is denoted as P_t, the q-th instance of the received signal can be written as

$$\mathbf{y}[q] = \mathbf{W}^* \sum_{n=0}^{N_\mathrm{d}-1} \sqrt{P_\mathrm{t}}\mathbf{H}_n\mathbf{F}\mathbf{s}[q-n] + \mathbf{W}^*\mathbf{n}[q], \tag{2.3}$$

where $\mathbf{n}[q] \sim \mathcal{N}(\mathbf{0}, \sigma_\mathbf{n}^2\mathbf{I}_{N_\mathrm{r}})$ is additive white Gaussian noise. We compute the variance of the noise term as $\sigma_\mathbf{n}^2 = K_\mathrm{B}TB_c$, where K_B is the Boltzmann's constant, T is the absolute temperature of the receiver, and B_c is the system bandwidth. Note that the noise after combining is no longer white, i.e. $\mathbb{E}[\mathbf{W}^*\mathbf{n}[q]\mathbf{n}[q]^*\mathbf{W}] = \sigma_\mathbf{n}^2\mathbb{E}[\mathbf{W}^*\mathbf{W}] \neq \mathbf{I}$. To whiten the receive signal in (2.3), $\mathbf{y}[q]$ is multiplied by the inverse of a lower triangular matrix $\mathbf{L}$ as $\breve{\mathbf{y}}[q] = \mathbf{L}^{-1}\mathbf{y}[q]$, where $\mathbf{L}$ is obtained from the Cholesky decomposition $\mathbf{W}^*\mathbf{W} = \mathbf{L}\mathbf{L}^*$. Let $\breve{\mathbf{W}} = \mathbf{L}^{-1}\mathbf{W}$ and $\breve{\mathbf{n}}[q] = \mathbf{L}^{-1}\mathbf{W}^*\mathbf{n}[q]$, then

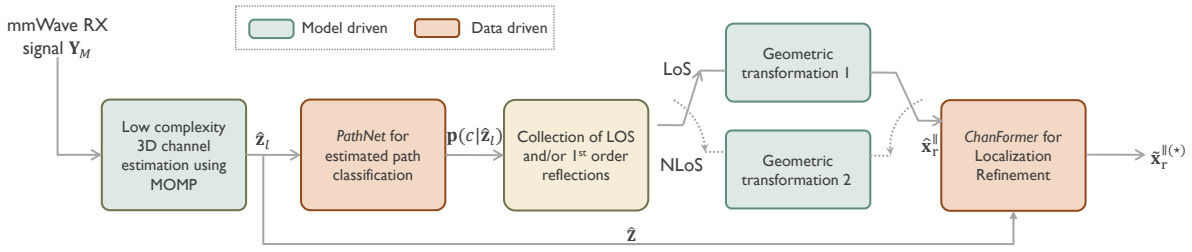


Figure 2.1. Diagram of the joint initial access and localization system model.

(2.3) can be rewritten as

$$\check{\mathbf{y}}[q] = \check{\mathbf{W}}^* \sum_{n=0}^{N_\mathrm{d}-1} \sqrt{P_\mathrm{t}} \mathbf{H}_n \mathbf{F} \mathbf{s}[q-n] + \check{\mathbf{n}}[q], \tag{2.4}$$

where $\mathbb{E}[\check{\mathbf{n}}[q]\check{\mathbf{n}}[q]^*] = \sigma_\mathbf{n}^2 \mathbf{I}$. Let $\check{\mathbf{Y}} = [\check{\mathbf{y}}[1], ..., \check{\mathbf{y}}[Q]] \in \mathbb{C}^{N_s \times Q}$ be the matrix collecting the received samples for the different training frames, and $\check{\mathbf{N}} = [\check{\mathbf{n}}[1], ..., \check{\mathbf{n}}[Q]]$ be the noise matrix. The whitened received signal matrix can be written as

$$\check{\mathbf{Y}} = \sqrt{P_\mathrm{t}} \check{\mathbf{W}}^* [\mathbf{H}_0, ..., \mathbf{H}_{N_\mathrm{d}-1}] \left((\mathbf{I}_{N_\mathrm{d}} \otimes \mathbf{F}) \mathbf{S} \right) + \check{\mathbf{N}}, \tag{2.5}$$

where

$$\mathbf{S} = \begin{bmatrix} \mathbf{s}[1] & \mathbf{s}[2] & \dots & \mathbf{s}[Q] \\ \mathbf{0} & \mathbf{s}[1] & \dots & \mathbf{s}[Q-1] \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & \mathbf{s}[Q-(N_\mathrm{d}-1)] \end{bmatrix}. \tag{2.6}$$

When using a set of M_t precoders $\{\mathbf{F}_{m_\mathrm{t}} | m_\mathrm{t} = 1, ..., M_\mathrm{t}\}$ and a set of M_r combiners $\{\mathbf{W}_{m_\mathrm{r}} | m_\mathrm{r} = 1, ..., M_\mathrm{r}\}$ for training, it is possible to write the expression of the received signal for a particular precoder/combiner pair $\check{\mathbf{Y}}_{m_\mathrm{r},m_\mathrm{t}}$ from (2.5) by substituting $\mathbf{F}$ by $\mathbf{F}_{m_\mathrm{t}}$, $\check{\mathbf{W}}$ by $\check{\mathbf{W}}_{m_\mathrm{r}}$ and $\check{\mathbf{N}}$ by $\check{\mathbf{N}}_{m_\mathrm{r},m_\mathrm{t}}$. In the next sections, we develop the stages that process this received signal for channel estimation and precise positioning.

The block diagram of our proposed joint initial access and 3D vehicle localization strategy

is shown in Fig. 2.1. First, we collect in $\mathbf{Y}_M$ the mmWave received signals for M different combinations of the training precoders and combiners. Then, we employ a low complexity channel estimation technique to acquire N_{est} estimated paths $\hat{\mathbf{Z}} = [\hat{\mathbf{z}}_1, ..., \hat{\mathbf{z}}_{N_{\text{est}}}]$, where each vector $\hat{\mathbf{z}}_\ell$ contains: the magnitude of the estimated channel gain $|\alpha_\ell|$, the relative delay $\tau_\ell = t_\ell - t_0$, the DoA $\boldsymbol{\theta}_\ell$, and the DoD $\boldsymbol{\phi}_\ell$. Then, every channel path is classified by *PathNet*, a lightweight network which predicts the probability of $\hat{\mathbf{z}}_\ell$ being a LOS component, a first-order reflection, or a high-order reflection, so that the LOS and first-order reflections are later exploited for localization using the geometric relationships between the path parameters and the vehicle's position. Note that these relationships depend on the channel state (LOS or NLOS). Since the location estimates provided by this stage cannot guarantee sub-meter accuracy for most of the users, an additional data driven stage realized with *ChanFormer*, a self-attention network, is thus proposed for location refinement. To this aim, a set of tiles with a given size is built around the initial position estimate, and the output from *ChanFormer* provides a probability map showing which tile contains the true location with the highest probability. *ChanFormer* analyzes the relationships among the estimated paths in $\hat{\mathbf{Z}}$ and measures the congruence between the channel features and the initial location estimate $\hat{\mathbf{x}}_{\text{r}}^{\text{II}}$. The input estimated channel features are extracted through self-attention in the encoder section of the network. These features are then decoded and matched to a more precise location estimate associated with the center of the highest probability tile. Though we use a square tile structure in this chapter, the number and the shape of the tiles could both be customized to suit the specific environment and required accuracy.

2.3 3D mmWave Channel Estimation via On-Grid and Off-Grid Methods for Initial Access

2.3.1 Two-Stage MOMP-Based Channel Estimation

Channel Estimation at mmWave Exploiting Sparsity and Conventional OMP

Prior work on compressive channel estimation at mmWave exploiting OMP (see for example [80, 81, 109]), leverages a representation of the channel in terms of a sparsifying dictionary $\mathbf{\Psi}$, defined as a Kronecker product of several matrices built from the steering vectors at the transmitter and at the receiver evaluated on a grid for the DoD and the DoA, and an additional component to represent the delay domain. By definition, the Kronecker structure creates a dictionary with a size related to the product of the sizes of the matrices which represent the angular and delay domains. These sizes are also related to the array dimension and the required resolution of the dictionary. When operating at mmWave with large planar arrays, the size of the dictionary becomes too large, posing challenges in terms of memory and the number of operations required to solve the sparse recovery problem. Mathematically, the first step to define the sparse recovery problem consists of vectorizing the received signal matrix in (5). By exploiting the properties of the vectorization operator, we can obtain that

$$\mathrm{vec}(\breve{\mathbf{Y}}) = \mathbf{\Upsilon}\mathbf{\Psi}\mathbf{c} + \mathrm{vec}(\breve{\mathbf{N}}), \tag{2.7}$$

where $\mathbf{\Upsilon} = \left((\mathbf{I}_{N_\mathrm{d}} \otimes \mathbf{F})\sqrt{P_\mathrm{t}}\mathbf{S}\right)^{\mathsf{T}} \otimes \breve{\mathbf{W}}^{*} \in \mathbb{C}^{N_s Q \times N_\mathrm{r} N_\mathrm{t} N_\mathrm{d}}$ and $\mathbf{\Psi} \in \mathbb{C}^{N_\mathrm{r} N_\mathrm{t} N_\mathrm{d} \times N_\mathrm{r}^\mathrm{a} N_\mathrm{t}^\mathrm{a} N_\mathrm{d}^\mathrm{a}}$ are the measurement matrix and the sparsifying dictionary, with N_r^a, N_t^a, and N_d^a depending on the required angular and delay resolutions, and $\mathbf{c} \in \mathbb{C}^{N_\mathrm{r}^\mathrm{a} N_\mathrm{t}^\mathrm{a} N_\mathrm{d}^\mathrm{a} \times 1}$ is the sparse vector representing the channel. The dictionary matrix $\mathbf{\Psi}$ is computed as [57]

$$\mathbf{\Psi} = \mathbf{A}_\mathrm{d} \otimes (\overline{\mathbf{A}}_\mathrm{t} \otimes \mathbf{A}_\mathrm{r}) \in \mathbb{C}^{N_\mathrm{r} N_\mathrm{t} N_\mathrm{d} \times N_\mathrm{r}^\mathrm{a} N_\mathrm{t}^\mathrm{a} N_\mathrm{d}^\mathrm{a}}, \tag{2.8}$$

where $\mathbf{A}_{\mathrm{d}} = \left[\mathbf{p}(\ddot{\imath}_1), ..., \mathbf{p}(\ddot{\imath}_{N_{\mathrm{d}}^{\mathrm{a}}})\right]$ is the dictionary for the delay, being $\mathbf{p}(t) = [f_{\mathrm{p}}(0 \cdot T_s - t), \ldots, f_{\mathrm{p}}((N_d - 1)T_s - t)]^{\mathsf{T}} \in \mathbb{R}^{N_d \times 1}$ a sampled version of the function that models the filtering effects in the discrete time equivalent channel model in (2), and $\{\ddot{\imath}_n | n = 1, ..., N_{\mathrm{d}}^{\mathrm{a}}\}$ the grid points in the delay domain; $\mathbf{A}_{\mathrm{t}} = \left[\mathbf{a}(\ddot{\boldsymbol{\phi}}_1), ..., \mathbf{a}(\ddot{\boldsymbol{\phi}}_{N_{\mathrm{t}}^{\mathrm{a}}})\right] \in \mathbb{C}^{N_{\mathrm{t}} \times N_{\mathrm{t}}^{\mathrm{a}}}$ is the dictionary for the DoD considering the grid points $\{\ddot{\boldsymbol{\phi}}_n | n = 1, ..., N_{\mathrm{t}}^{\mathrm{a}}\}$, and it can be decomposed as $\mathbf{A}_{\mathrm{t}} = \mathbf{A}_{\mathrm{t}}^{\mathrm{x}} \otimes \mathbf{A}_{\mathrm{t}}^{\mathrm{y}}$, where $[\mathbf{A}_{\mathrm{t}}^{\mathrm{x}}]_{:,n} = \mathbf{a}(\ddot{\boldsymbol{\phi}}_n^{\shortparallel})$ and $[\mathbf{A}_{\mathrm{t}}^{\mathrm{y}}]_{:,n} = \mathbf{a}(\ddot{\boldsymbol{\phi}}_n^{\perp})$, with $\ddot{\boldsymbol{\phi}}_n^{\shortparallel}$ and $\ddot{\boldsymbol{\phi}}_n^{\perp}$ the n-th selection of the grid values for the DoD in azimuth and elevation, respectively; finally, $\mathbf{A}_{\mathrm{r}} = \left[\mathbf{a}(\ddot{\boldsymbol{\theta}}_1), ..., \mathbf{a}(\ddot{\boldsymbol{\theta}}_{N_{\mathrm{r}}^{\mathrm{a}}})\right] = \mathbf{A}_{\mathrm{r}}^{\mathrm{x}} \otimes \mathbf{A}_{\mathrm{r}}^{\mathrm{y}} \in \mathbb{C}^{N_{\mathrm{r}} \times N_{\mathrm{r}}^{\mathrm{a}}}$ is the dictionary the DoA defined in a similar way as $\mathbf{A}_{\mathrm{t}}$. Given these definitions, prior work (see for example [57]) estimates the sparse representation of the channel $\mathbf{c}$ by exploiting OMP to solve the problem

$$\min_{\mathbf{c}} \left\| \breve{\mathbf{Y}} - \boldsymbol{\Upsilon}\boldsymbol{\Psi}\mathbf{c} \right\|^2, \tag{2.9}$$

which has a complexity $\mathcal{O}(N_{\mathrm{est}} N_s Q N_{\mathrm{r}} N_{\mathrm{t}} N_{\mathrm{d}} N_{\mathrm{r}}^{\mathrm{a}} N_{\mathrm{t}}^{\mathrm{a}} N_{\mathrm{d}}^{\mathrm{a}})$. Given the practical values of the parameters that impact this complexity when operating with large antenna arrays and fine dictionary resolutions, the OMP algorithm could not be executed in a conventional server or a high end personal computer. To address this limitation, the recently proposed MOMP algorithm [85, 105] solves the associated sparse recovery problem by exploiting independent dictionaries for every sparse dimension instead of a single, very large dictionary, built as a Kronecker product of these independent dictionaries, as described next.

MOMP based channel estimation

The fundamental idea of MOMP is to rearrange elements in $\boldsymbol{\Upsilon}$ and $\boldsymbol{\Psi}$ into N_{D} orthogonal dimensions and execute tensor multiplications independently along each dimension. Considering the received signals $\breve{\mathbf{Y}} \in \mathbb{C}^{N_s \times Q}$, the algorithm starts by constructing N_{D} independent sparsifying dictionaries $\{\boldsymbol{\Psi}_k \in \mathbb{C}^{N_k^{\mathrm{s}} \times N_k^{\mathrm{a}}} \mid k = 1, ..., N_{\mathrm{D}}\}$, where N_k^{a} is the number of atoms in the dictionary $\boldsymbol{\Psi}_k$, and N_k^{s} is the size of each atom. Then, the measurement tensor is defined as $\boldsymbol{\Phi} \in \mathbb{C}^{N_s \times \otimes_{k=1}^{N_{\mathrm{D}}} N_k^{\mathrm{s}}}$, where $\otimes_{k=1}^{N_{\mathrm{D}}} N_k^{\mathrm{s}}$ represents the tensor shape of $N_1^{\mathrm{s}} \times N_2^{\mathrm{s}} \times ... \times N_{\mathrm{D}}^{\mathrm{s}}$. The target of the algorithm is to

solve the multidimensional matching pursuit problem in (2.10) to extract the sparse coefficients from the tensor $\mathbf{C} \in \mathbb{C}^{\otimes_{k=1}^{N_\mathrm{D}} N_k^\mathrm{a} \times Q}$:

$$\min_{\mathbf{C}} \left(\left\| \check{\mathbf{Y}} - \sum_{\mathbf{i}\in\mathcal{I}} \sum_{\mathbf{j}\in\mathcal{J}} [\mathbf{\Phi}]_{:,\mathbf{i}} \left(\prod_{k=1}^{N_\mathrm{D}} [\mathbf{\Psi}_k]_{i_k,j_k} \right) [\mathbf{C}]_{\mathbf{j},:} \right\|_\mathrm{F}^2 \right), \tag{2.10}$$

where $\mathcal{I} = \{\mathbf{i} = (i_1, \ldots, i_{N_\mathrm{D}}) \in \mathbb{N}^{N_\mathrm{D}} | i_k \leq N_k^\mathrm{s},\ \forall k \leq N_\mathrm{D}\}$, and $\mathcal{J} = \{\mathbf{j} = (j_1, \ldots, j_{N_\mathrm{D}}) \in \mathbb{N}^{N_\mathrm{D}} | j_d \leq N_k^\mathrm{a},\ \forall k \leq N_\mathrm{D}\}$ represent the multidimensional indices.

The application of MOMP to joint localization and channel estimation in an indoor scenario was proposed in [85, 105], where all the additional details for the problem formulation and solution can be found, including a link to the algorithm implementation in GitHub (https://github.com/WiSeCom-Lab/MOMP-core.git). In this work, $N_\mathrm{D} = 5$ independent dictionaries are considered – for the DoD in azimuth, DoD in elevation, delay domain, DoA in azimuth, and DoA in elevation, namely, $\mathbf{\Psi}_1 = \overline{\mathbf{A}_\mathrm{t}^\mathrm{x}}$, $\mathbf{\Psi}_2 = \overline{\mathbf{A}_\mathrm{t}^\mathrm{y}}$, $\mathbf{\Psi}_3 = \mathbf{A}_\mathrm{d}$, $\mathbf{\Psi}_4 = \mathbf{A}_\mathrm{r}^\mathrm{x}$, and $\mathbf{\Psi}_5 = \mathbf{A}_\mathrm{r}^\mathrm{y}$. MOMP computes first the projections on these five different sparsifying dictionaries (which cover all possible angular domains and the delay component) independently, and exploits an alternating optimization strategy to converge to the same solution that conventional OMP would provide. Although MOMP requires additional steps for initialization and iterative refinement, the complexity of each step is much lower than those in OMP, resulting in a much lower overall complexity. In particular, the complexity is reduced to $\mathcal{O}\left(N_\mathrm{est} N_s Q N_\mathrm{iter} (\sum_{k=1}^{N_\mathrm{D}} N_k^\mathrm{a})(\prod_{k=1}^{N_\mathrm{D}} N_k^\mathrm{s})\right)$ from the previous $\mathcal{O}\left(N_\mathrm{est} N_s Q \prod_{k=1}^{N_\mathrm{D}} N_k^\mathrm{s} N_k^\mathrm{a}\right)$ when exploiting OMP, where N_iter is the number of iterations for refining the estimates associated with each dictionary. This advantage becomes particularly interesting when operating with high resolution dictionaries, because both computational complexity and memory requirements are significantly reduced, since $\sum_{k=1}^{N_\mathrm{D}} N_\mathrm{iter} N_k^\mathrm{a} \ll \prod_{k=1}^{N_\mathrm{D}} N_k^\mathrm{a}$.

Two-Stage MOMP Enabling Finer Resolutions

To further reduce memory requirements targeting the outdoor 3D localization problem, which considers large arrays and fine resolutions, we propose a modification of the MOMP-based

channel estimation strategy in [85, 105]. It comprises two stages: 1) estimating delays, DoDs, and a parameter that we define as equivalent gains, which include the combined effect of the path complex gains and the DoAs; this way we can apply MOMP for channel estimation with $N_{\mathrm{D}} = 3$ dictionaries instead of 5; and 2) estimating the DoAs from the equivalent gains. In the next paragraphs, we develop the estimators to implement this two-stage approach.

For stage 1), we start by constructing three independent sparsifying dictionaries $\mathbf{\Psi}_1$, $\mathbf{\Psi}_2$, and $\mathbf{\Psi}_3$ as defined before. Considering training with M_{r} combiners, the effect of combiners and the arrival angular information are embedded into what we define as the equivalent gain tensor $\mathbf{C} \in \mathbb{C}^{N_1^{\mathrm{a}} \times N_2^{\mathrm{a}} \times N_3^{\mathrm{a}} \times N_s M_{\mathrm{r}}}$, which is sparse, and can be written as

$$[\mathbf{C}]_{\mathbf{j},:} = \begin{cases} \boldsymbol{\beta}_\ell^{\mathsf{T}} & \text{if } \begin{array}{c} \phi_\ell^{\mathrm{x}} = \ddot{\phi}_{j_1}^{\shortparallel}, \phi_\ell^{\mathrm{y}} = \ddot{\phi}_{j_2}^{\perp}, \\ t_\ell - t_0 = \ddot{t}_{j_3} \end{array} ; \\ 0 & \text{o.w.} \end{cases}, \tag{2.11}$$

where $[\boldsymbol{\beta}_\ell]_{N_s(m_{\mathrm{r}}-1)+n_s} = \alpha_\ell [\breve{\mathbf{W}}_{m_{\mathrm{r}}}]^*_{:,n_s} \mathbf{a}_{\mathrm{r}}(\boldsymbol{\theta}_\ell)$. For a given training combiner $\breve{\mathbf{W}}_{m_{\mathrm{r}}}$, we define the part in (2.5) that contains the combiner and the channel matrices for different delays as the combined channel $\mathbf{H}^{(m_{\mathrm{r}})} = \breve{\mathbf{W}}^*_{m_{\mathrm{r}}} [\mathbf{H}_0, ..., \mathbf{H}_{N_{\mathrm{d}}-1}]$, with $\mathbf{H}^{(m_{\mathrm{r}})} \in \mathbb{C}^{N_s \times N_{\mathrm{d}} N_{\mathrm{t}}^{\mathrm{x}} N_{\mathrm{t}}^{\mathrm{y}}}$, which can also be represented by the multiplication of $\mathbf{\Psi}_k$ and $\mathbf{C}$ as

$$[\mathbf{H}^{(m_{\mathrm{r}})}]_{n_s, (i_3-1)N_{\mathrm{t}}^{\mathrm{x}} N_{\mathrm{t}}^{\mathrm{y}} + (i_1-1)N_{\mathrm{t}}^{\mathrm{y}} + i_2} = \sum_{\mathbf{j} \in \mathcal{J}} \left(\prod_{k=1}^{3} [\mathbf{\Psi}_k]_{i_k, j_k} \right) [\mathbf{C}]_{\mathbf{j}, (m_{\mathrm{r}}-1)N_s + n_s}. \tag{2.12}$$

Now (2.5) can be alternatively written as

$$\begin{aligned} [\breve{\mathbf{Y}}_{m_{\mathrm{r}}, m_{\mathrm{t}}}]_{n_s, q} = & \sum_{\mathbf{i} \in \mathcal{I}} [(\mathbf{I}_{N_{\mathrm{d}}} \otimes \mathbf{F}_{m_{\mathrm{t}}}) \sqrt{P_{\mathrm{t}}} \mathbf{S}]_{(i_3-1)N_{\mathrm{t}}^{\mathrm{x}} N_{\mathrm{t}}^{\mathrm{y}} + (i_1-1)N_{\mathrm{t}}^{\mathrm{y}} + i_2, q} \\ & \cdot \sum_{\mathbf{j} \in \mathcal{J}} \left(\prod_{k=1}^{3} [\mathbf{\Psi}_k]_{i_k, j_k} \right) [\mathbf{C}]_{\mathbf{j}, (m_{\mathrm{r}}-1)N_s + n_s} + [\breve{\mathbf{N}}]_{n_s, q}. \end{aligned} \tag{2.13}$$

We can now derive the measurement tensor, which is the remaining key component for solving the MOMP problem. In [85, 105], the measurement tensor $\mathbf{\Phi}$ includes the effect of both precoder and combiner, however, it currently only contains the information of the precoder in our solution, with the combiner effects factored into the equivalent gain $\mathbf{C}$ to be estimated. Hence, we define $\mathbf{\Phi}_{m_\mathrm{t}} \in \mathbb{C}^{Q\times N_\mathrm{t}^\mathrm{x}\times N_\mathrm{t}^\mathrm{y}\times N_\mathrm{d}}$ as the measurement tensor obtained with $\mathbf{F}_{m_\mathrm{t}}$, and the whole measurement tensor composed of $\mathbf{\Phi}_{m_\mathrm{t}}$, where $1 \le m_\mathrm{t} \le M_\mathrm{t}$, is $\mathbf{\Phi}_M \in \mathbb{C}^{QM_\mathrm{t}\times N_\mathrm{t}^\mathrm{x}\times N_\mathrm{t}^\mathrm{y}\times N_\mathrm{d}}$, where

$$[\mathbf{\Phi}_M]_{Q(m_\mathrm{t}-1)+q,\mathbf{i}} = [(\mathbf{I}_{N_\mathrm{d}} \otimes \mathbf{F}_{m_\mathrm{t}})\mathbf{S}]_{(i_3-1)N_\mathrm{t}^\mathrm{x}N_\mathrm{t}^\mathrm{y}+(i_1-1)N_\mathrm{t}^\mathrm{y}+i_2,q} \tag{2.14}$$

$$= [\mathbf{F}_{m_\mathrm{t}}\mathbf{s}[q-(i_3-1)]]_{(i_1-1)N_\mathrm{t}^\mathrm{y}+i_2}\,. \tag{2.15}$$

Now the components of (2.10) are ready, where $\breve{\mathbf{Y}}_M$ is formed by collecting multiple observations using different pairs of $\mathbf{F}_{m_\mathrm{t}}$ and $\mathbf{W}_{m_\mathrm{r}}$:

$$\breve{\mathbf{Y}}_M = \begin{bmatrix} \breve{\mathbf{Y}}_{1,1}^\mathsf{T} & \cdots & \breve{\mathbf{Y}}_{M_\mathrm{r},1}^\mathsf{T} \\ \vdots & \ddots & \vdots \\ \breve{\mathbf{Y}}_{1,M_\mathrm{t}}^\mathsf{T} & \cdots & \breve{\mathbf{Y}}_{M_\mathrm{r},M_\mathrm{t}}^\mathsf{T} \end{bmatrix} \in \mathbb{C}^{QM_\mathrm{t}\times N_s M_\mathrm{r}}. \tag{2.16}$$

We employ the MOMP algorithm in [85] to solve this problem and obtain the estimated values of $\boldsymbol{\phi}_\ell$, τ_ℓ, and $\boldsymbol{\beta}_\ell$, $\ell = 1,\ldots,L$.

For stage 2), to retrieve the DoA information from the non-zero coefficients of $\mathbf{C}$, i.e. $\boldsymbol{\beta}_\ell$, the main idea is to correlate the coefficients with angular dictionaries, so the DoA for the different paths can be obtained by finding the peaks of this correlation. Let $\mathbf{\Psi}_\mathrm{r} = \mathbf{\Psi}_4 \otimes \mathbf{\Psi}_5$ and $\breve{\mathbf{W}}_M = [\breve{\mathbf{W}}_1, \cdots, \breve{\mathbf{W}}_{M_\mathrm{r}}]$ (note that we remove the notation for measurement indices for simplicity), then $\boldsymbol{\beta}_\ell$ can be rewritten as $\boldsymbol{\beta}_\ell = \alpha_\ell \breve{\mathbf{W}}_M^* \mathbf{a}_\mathrm{r}(\boldsymbol{\theta}_\ell)$. Hence, assuming every entry of $\breve{\mathbf{W}}_M$ is orthonormal, by multiplying $\boldsymbol{\beta}_\ell^*$, $\breve{\mathbf{W}}_M^*$, and the angular dictionary leads to

$$\boldsymbol{\beta}_\ell^* \breve{\mathbf{W}}_M^* \mathbf{\Psi}_\mathrm{r} = \alpha_\ell \mathbf{a}_\mathrm{r}(\boldsymbol{\theta}_\ell)^* \breve{\mathbf{W}}_M \breve{\mathbf{W}}_M^* \mathbf{\Psi}_\mathrm{r} = \alpha_\ell \mathbf{a}(\boldsymbol{\theta}_\ell)^* \mathbf{\Psi}_\mathrm{r}, \tag{2.17}$$

and the DoAs can now be retrieved as

$$\hat{\boldsymbol{\theta}}_\ell = \arg\max_{\ddot{\boldsymbol{\theta}}} \hat{\boldsymbol{\beta}}_\ell^* \breve{\mathbf{W}}_M^* \boldsymbol{\Psi}_\mathrm{r}. \tag{2.18}$$

Note that (2.18) can also be solved using MOMP by independent tensor multiplications in the arrival angular domain in azimuth and elevation, especially when N_k^s and N_k^a are large. This way, the computational complexity of our two-stage channel estimation algorithm is $\mathcal{O}\left(N_\mathrm{est} N_s Q N_\mathrm{iter} (\sum_{k=1}^{3} N_k^\mathrm{a})(\prod_{k=1}^{3} N_k^\mathrm{s})\right)$, which is lower than using the single stage MOMP for simultaneous estimation across five dimensions [85]. Table 2.1 summarizes the computational complexity for the three channel estimation methods. The channel parameters required for localization – 3D DoDs/DoAs and TDoAs – are now available, except the path order which determines whether to discard an estimated path or not, as only LOS and first-order reflections will be used for localization. The path classification problem is addressed in Sec. 2.4.1.

Table 2.1. Complexity comparisons for various channel estimation algorithms.

Method	Complexity
Conventional OMP [57]	$\mathcal{O}\left(N_\mathrm{est} N_s Q \prod_{k=1}^{5} N_k^\mathrm{s} N_k^\mathrm{a}\right)$
MOMP [32,52]	$\mathcal{O}\left(N_\mathrm{est} N_s Q N_\mathrm{iter} (\sum_{k=1}^{5} N_k^\mathrm{a})(\prod_{k=1}^{5} N_k^\mathrm{s})\right)$
Two-stage MOMP (Proposed)	$\mathcal{O}\left(N_\mathrm{est} N_s Q N_\mathrm{iter} (\sum_{k=1}^{3} N_k^\mathrm{a})(\prod_{k=1}^{3} N_k^\mathrm{s})\right)$

2.3.2 Beamspace ESPRIT-D Channel Estimation

The 3D mmWave channel accounting for pulse shaping can be modelled as a tensor spanning five dimensions. Thereafter, to tackle the task of estimating directions and delays for multiple propagation paths, we introduce an innovative method termed ESPRIT-D. This method combines the strengths of beamspace ESPRIT for the simultaneous extraction of azimuth and elevation angles with a dictionary-based sparse recovery solution that targets delay estimation. Furthermore, the algorithm's performance is assessed through ray-tracing simulations, providing a robust evaluation in the context of realistic channel conditions.

Background of beamspace ESPRIT

The CP representation of a Zth-order tensor $\boldsymbol{\Upsilon} \in \mathbb{C}^{N_1 \times N_2 \times \ldots \times N_Z}$ of rank N_L is written as a linear combination of the tensor outer products of rank-1 tensors $\mathbf{u}_{z,\ell}$ [110] as

$$\boldsymbol{\Upsilon} = \sum_{\ell=1}^{N_L} \gamma_\ell \mathbf{u}_{1,\ell} \circ \mathbf{u}_{2,\ell} \circ \cdots \circ \mathbf{u}_{Z,\ell}, \tag{2.19}$$

where $\mathbf{u}_{z,\ell}$ is the rank-1 tensor spanning the z-th dimension of the ℓ-th tensor product, and γ_ℓ is the coefficient of the ℓ-th tensor product. To adapt the CP representation for our channel estimation problem, $\mathbf{u}_{z,\ell}$ is defined in the form of a steering vector as $\mathbf{u}_{z,\ell} = [e^{j0\omega_{z,\ell}}, e^{j1\omega_{z,\ell}}, \ldots, e^{j(N_z-1)\omega_{z,\ell}}]^\mathsf{T}$, with $\omega_{z,\ell}$ being the frequency component of the z-th dimension of the ℓ-th tensor product. A tensor decomposition method, e.g. canonical polyadic decomposition (CPD) in [110], can be adopted to derive the estimated rank-1 tensors, denoted as $\hat{\mathbf{u}}_{z,\ell}$. In this case, the optimization problem to be solved is

$$\min_{\hat{\mathbf{U}}_z} \left\| \boldsymbol{\Upsilon} - \sum_{\ell=1}^{N_L} \hat{\mathbf{u}}_{1,\ell} \circ \cdots \circ \hat{\mathbf{u}}_{Z,\ell} \right\|. \tag{2.20}$$

Let us define $\hat{\mathbf{U}}_z = [\hat{\mathbf{u}}_{z,1}, ..., \hat{\mathbf{u}}_{z,N_L}]$, and

$$\begin{cases} \mathbf{J}_z^{(1)} = \left[\mathbf{I}_{N_z-1} \ \mathbf{0}_{(N_z-1)\times 1}\right] \\ \mathbf{J}_z^{(2)} = \left[\mathbf{0}_{N_z-1\times 1} \ \mathbf{I}_{N_z-1}\right] \end{cases}, \tag{2.21}$$

then, the shift invariance property can be applied, obtaining

$$\mathbf{J}_z^{(1)} \hat{\mathbf{U}}_z = \mathbf{J}_z^{(2)} \hat{\mathbf{U}}_z \boldsymbol{\Phi}_z, \tag{2.22}$$

where $\boldsymbol{\Phi}_z$'s diagonal elements contain the estimated frequency components $\hat{\omega}_{z,\ell}$ for $1 \leq \ell \leq N_L$ for the z-th dimension, i.e., $\boldsymbol{\Phi}_z = \text{diag}[e^{-j\hat{\omega}_{z,1}}, ..., e^{-j\hat{\omega}_{z,N_L}}]$.

When using beamspace tensors, we define the beamspace rank-1 tensor $\mathbf{u}'_{z,\ell} = \mathbf{B}_z^* \mathbf{u}_{z,\ell} \in \mathbb{C}^{M_z \times 1}$, where $\mathbf{B}_z = \left[\mathbf{b}_{z,1}, ..., \mathbf{b}_{z,M_z}\right] \in \mathbb{C}^{N_z \times M_z}$ contains M_z "beams", each one of them defined

by a vector of length N_z. Using this identity, (2.19) becomes $\boldsymbol{\Upsilon} = \sum_{\ell=1}^{N_L} \gamma'_\ell \mathbf{u}'_{1,\ell} \circ \mathbf{u}'_{2,\ell} \circ \cdots \circ \mathbf{u}'_{Z,\ell}$. Replacing $\hat{\mathbf{u}}_{z,\ell}$ with the beamspace tensor $\hat{\mathbf{u}}'_{z,\ell}$ in (2.20), $\hat{\mathbf{u}}'_{z,\ell}$ can be still acquired via CPD as mentioned above. We denote the concatenation of the acquired beamspaces tensors of the z-th dimension as $\hat{\mathbf{U}}'_z = \left[\hat{\mathbf{u}}'_{z,1}, ..., \hat{\mathbf{u}}'_{z,N_L}\right]$, and it turns out (2.22) no longer holds, since $\mathbf{J}_z^{(1)}\hat{\mathbf{U}}'_z \neq \mathbf{J}_z^{(2)}\hat{\mathbf{U}}'_z\boldsymbol{\Phi}_z$. A way to restore the shift invariance structure is proposed in [67]. First, we find $\boldsymbol{\Phi}'_z$ s.t. $\mathbf{J}_z^{(1)}\mathbf{B}_z = \mathbf{J}_z^{(2)}\mathbf{B}_z\boldsymbol{\Phi}'_z$ through least squares (LS) estimation as

$$\boldsymbol{\Phi}'_z = (\mathbf{J}_r^{(2)}\mathbf{B}_z)^\dagger \mathbf{J}_z^{(1)}\mathbf{B}_z. \tag{2.23}$$

Then, there exists an adjustment matrix

$$\boldsymbol{\Psi}_z = \mathbf{I}_{M_z} - ([\mathbf{B}_z^*]_{:,N_z})([\mathbf{B}_z^*]_{:,N_z})^* - (\boldsymbol{\Phi}'_z)^*[\mathbf{B}_z^*]_{:,1}[\mathbf{B}_z]_{1,:}\boldsymbol{\Phi}'_z \in \mathbb{C}^{M_z \times M_z}, \tag{2.24}$$

s.t.

$$\begin{cases} \boldsymbol{\Psi}_z[\mathbf{B}_z^*]_{:,N_z} = \mathbf{0}_{M_z \times 1} \\ \boldsymbol{\Psi}_z(\boldsymbol{\Phi}'_z)^*[\mathbf{B}_z^*]_{:,1} = \mathbf{0}_{M_z \times 1} \end{cases}, \tag{2.25}$$

such that $\boldsymbol{\Psi}_z(\boldsymbol{\Phi}'_z)^*\mathbf{U}'_z = \boldsymbol{\Psi}_z\mathbf{U}'_z\boldsymbol{\Phi}_z^*$ holds. Let $\mathbf{U}'_z = \hat{\mathbf{U}}'_z\mathbf{D}_z$, where $\mathbf{D}_z \in \mathbb{C}^{N_L \times N_L}$ is an unknown non-singular matrix. We now have $\boldsymbol{\Psi}_z(\boldsymbol{\Phi}'_z)^*\hat{\mathbf{U}}'_z = \boldsymbol{\Psi}_z\hat{\mathbf{U}}'_z\mathbf{D}_z\boldsymbol{\Phi}_z^*(\mathbf{D}_z)^{-1}$. The value of $\boldsymbol{\Gamma}_z = \mathbf{D}_z\boldsymbol{\Phi}_z^*(\mathbf{D}_z)^{-1}$ can be determined by LS estimation as

$$\hat{\boldsymbol{\Gamma}}_z = (\boldsymbol{\Psi}_z\hat{\mathbf{U}}'_z)^\dagger \boldsymbol{\Psi}_z(\boldsymbol{\Phi}'_z)^*\hat{\mathbf{U}}'_z. \tag{2.26}$$

As $\boldsymbol{\Gamma}_z\mathbf{D}_z = \mathbf{D}_z\boldsymbol{\Phi}_z^*$, the eigenvalues of $\hat{\boldsymbol{\Gamma}}_z$, denoted as $\text{eig}(\hat{\boldsymbol{\Gamma}}_z) \in \mathbb{C}^{N_L \times 1}$, are the estimation of the diagonal elements of $\boldsymbol{\Phi}_z^* = \text{diag}[e^{j\hat{\omega}_{z,1}}, ..., e^{j\hat{\omega}_{z,N_L}}]$.

Beamspace ESPRIT-D for Joint 3D Angle and Delay Estimation

We first define the delay response vector as

$$\mathbf{p}(\tau_\ell) = [p(0 \cdot T_s - \tau_\ell), \ldots, p((N_d - 1)T_s - \tau_\ell)]^{\mathsf{T}} \in \mathbb{R}^{N_d \times 1}, \tag{2.27}$$

and rearrange the elements in (2.5) so that it can be reformatted as

$$\mathbf{y} = \sum_{\ell=1}^{L} \alpha_\ell \mathbf{w}^* \dot{\mathbf{a}}(\boldsymbol{\theta}_\ell) \dot{\mathbf{a}}(\boldsymbol{\phi}_\ell)^* \mathbf{f} \mathbf{p}(\tau_\ell)^{\mathsf{T}} \sqrt{P_{\mathrm{t}}} \mathbf{S} + \check{\mathbf{n}}. \tag{2.28}$$

Assuming that M_{r} combiners (the m_{r}-th of which is denoted as $\mathbf{w}_{m_{\mathrm{r}}}$) and M_{t} precoders (the m_{t}-th of which is denoted as $\mathbf{f}_{m_{\mathrm{t}}}$) are used to collect $M = M_{\mathrm{r}} M_{\mathrm{t}}$ measurements, a tensor containing all the measurements can be written in the outer product format as

$$\mathbf{Y} = \sum_{\ell=1}^{L} \mathbf{W}^* \dot{\mathbf{a}}(\boldsymbol{\theta}_\ell) \circ \mathbf{F}^{\mathsf{T}} \bar{\mathbf{a}}(\boldsymbol{\phi}_\ell) \circ \alpha_\ell \sqrt{P_{\mathrm{t}}} \mathbf{S}^{\mathsf{T}} \mathbf{p}(\tau_\ell) + \check{\mathbf{N}}, \tag{2.29}$$

where $\mathbf{W} = [\mathbf{w}_1, ..., \mathbf{w}_{M_{\mathrm{r}}}]$, $\mathbf{F} = [\mathbf{f}_1, ..., \mathbf{f}_{M_{\mathrm{t}}}]$, and $[\mathbf{Y}]_{m_{\mathrm{r}}, m_{\mathrm{t}}, q}$ and $[\check{\mathbf{N}}]_{m_{\mathrm{r}}, m_{\mathrm{t}}, q}$ are the q-th instance of the received signal and the noise acquired with $\mathbf{f}_{m_{\mathrm{t}}}$ and $\mathbf{w}_{m_{\mathrm{r}}}$. We introduce $\dot{\mathbf{H}}$ as the spatial representation of the channel including the precoder/combiner effects, whose ℓ-th component is

$$\dot{\mathbf{H}}_\ell = \mathbf{W}^* \dot{\mathbf{a}}(\boldsymbol{\theta}_\ell) \circ \mathbf{F}^{\mathsf{T}} \bar{\mathbf{a}}(\boldsymbol{\phi}_\ell), \tag{2.30}$$

which can be further extended to a 4D tensor as

$$\dot{\mathbf{H}}_\ell = \mathbf{W}_{\mathrm{x}}^* \mathbf{a}(\theta_\ell^{\mathrm{x}}) \circ \mathbf{W}_{\mathrm{y}}^* \mathbf{a}(\theta_\ell^{\mathrm{y}}) \circ \mathbf{F}_{\mathrm{x}}^{\mathsf{T}} \bar{\mathbf{a}}(\theta_\ell^{\mathrm{x}}) \circ \mathbf{F}_{\mathrm{y}}^{\mathsf{T}} \bar{\mathbf{a}}(\theta_\ell^{\mathrm{y}}), \tag{2.31}$$

where $\mathbf{W}_{\mathrm{x}} = \left[\mathbf{w}_{\mathrm{x}}^{(1)}, ..., \mathbf{w}_{\mathrm{x}}^{(M_{\mathrm{r}}^{\mathrm{x}})}\right] \in \mathbb{C}^{N_{\mathrm{r}}^{\mathrm{x}} \times M_{\mathrm{r}}^{\mathrm{x}}}$ and $\mathbf{W}_{\mathrm{y}} = \left[\mathbf{w}_{\mathrm{y}}^{(1)}, ..., \mathbf{w}_{\mathrm{y}}^{(M_{\mathrm{r}}^{\mathrm{y}})}\right] \in \mathbb{C}^{N_{\mathrm{r}}^{\mathrm{y}} \times M_{\mathrm{r}}^{\mathrm{y}}}$ are used to build $\mathbf{W}$ as $\mathbf{W}_{\mathrm{x}} \otimes \mathbf{W}_{\mathrm{y}}$, and $M_{\mathrm{r}} = M_{\mathrm{r}}^{\mathrm{x}} M_{\mathrm{r}}^{\mathrm{y}}$. Similarly, $\mathbf{F}_{\mathrm{x}} \otimes \mathbf{F}_{\mathrm{y}} = \mathbf{F}$, where $\mathbf{F}_{\mathrm{x}} = \left[\mathbf{f}_{\mathrm{x}}^{(1)}, ..., \mathbf{f}_{\mathrm{x}}^{(M_{\mathrm{t}}^{\mathrm{x}})}\right]$ and $\mathbf{F}_{\mathrm{y}} = \left[\mathbf{f}_{\mathrm{y}}^{(1)}, ..., \mathbf{f}_{\mathrm{y}}^{(M_{\mathrm{t}}^{\mathrm{y}})}\right]$. Now $\mathbf{Y}$ can be written as the outer product of tensors spreading five

dimensions as

$$\mathbf{Y}_{5\mathrm{D}} = \sum_{\ell=1}^{L} \mathbf{W}_{\mathrm{x}}^{*}\mathbf{a}(\theta_{\ell}^{\mathrm{x}}) \circ \mathbf{W}_{\mathrm{y}}^{*}\mathbf{a}(\theta_{\ell}^{\mathrm{y}}) \circ \mathbf{F}_{\mathrm{x}}^{\mathsf{T}}\bar{\mathbf{a}}(\theta_{\ell}^{\mathrm{x}}) \circ \mathbf{F}_{\mathrm{y}}^{\mathsf{T}}\bar{\mathbf{a}}(\theta_{\ell}^{\mathrm{y}}) \circ \alpha_{\ell}\sqrt{P_{\mathrm{t}}}\mathbf{S}^{\mathsf{T}}\mathbf{p}(\tau_{\ell}) + \check{\mathbf{N}}, \tag{2.32}$$

and the relationship between $\mathbf{Y}$ and $\mathbf{Y}_{5\mathrm{D}}$ can be written as

$$[\mathbf{Y}_{5\mathrm{D}}]_{m_{\mathrm{r}}^{\mathrm{x}},m_{\mathrm{r}}^{\mathrm{y}},m_{\mathrm{t}}^{\mathrm{x}},m_{\mathrm{t}}^{\mathrm{y}},q} = [\mathbf{Y}]_{(m_{\mathrm{r}}^{\mathrm{x}}-1)M_{\mathrm{r}}^{\mathrm{y}}+m_{\mathrm{r}}^{\mathrm{y}},(m_{\mathrm{t}}^{\mathrm{x}}-1)M_{\mathrm{t}}^{\mathrm{y}}+m_{\mathrm{t}}^{\mathrm{x}},q}.$$

Now we solve the tensor decomposition problem so that $\mathbf{Y}_{5\mathrm{D}}$ can be approximated by the summation of N_L vector outer products, i.e.,

$$\min_{\hat{\mathbf{U}}_1',\ldots,\hat{\mathbf{U}}_5'} \left\| \mathbf{Y}_{5\mathrm{D}} - \sum_{\ell=1}^{N_L} \gamma_{\ell} \left(\circ_{z=1}^{5} \hat{\mathbf{u}}_{z,\ell}' \right) \right\|, \tag{2.33}$$

where the first 4 dimensions will be associated with $\mathbf{W}_{\mathrm{x}}^{*}\mathbf{a}(\theta_{\ell}^{\mathrm{x}})$, $\mathbf{W}_{\mathrm{y}}^{*}\mathbf{a}(\theta_{\ell}^{\mathrm{y}})$, $\mathbf{F}_{\mathrm{x}}^{\mathsf{T}}\bar{\mathbf{a}}(\theta_{\ell}^{\mathrm{x}})$, and $\mathbf{F}_{\mathrm{y}}^{\mathsf{T}}\bar{\mathbf{a}}(\theta_{\ell}^{\mathrm{y}})$ respectively, and the last dimension corresponds to the delay domain. Let the tensor decomposition result be $\hat{\mathbf{U}}_z' = [\hat{\mathbf{u}}_{z,1}', \ldots, \hat{\mathbf{u}}_{z,N_L}']$ following that in Sec. 2.3.2, then the beamspace ESPRIT can be performed to extract the angular information, while the delay information needs to be acquired by solving the sparse recovery problem through a dictionary based method.

Angle estimation: To apply the beamspace model, we determine $\mathbf{B}_z$ for $1 \leq z \leq 4$ as $\mathbf{B}_1 = \mathbf{W}_{\mathrm{x}}$, $\mathbf{B}_2 = \mathbf{W}_{\mathrm{y}}$, $\mathbf{B}_3 = \bar{\mathbf{F}}_{\mathrm{x}}$, and $\mathbf{B}_4 = \bar{\mathbf{F}}_{\mathrm{y}}$. The frequency component of the steering vector in each dimension is defined as:

$$\begin{cases} \omega_{1,\ell} = -\pi \cos\theta_{\ell}^{\mathrm{y}} \sin\theta_{\ell}^{\mathrm{x}} \\ \omega_{2,\ell} = -\pi \sin\theta_{\ell}^{\mathrm{y}} \\ \omega_{3,\ell} = \pi \cos\phi_{\ell}^{\mathrm{y}} \sin\phi_{\ell}^{\mathrm{x}} \\ \omega_{4,\ell} = \pi \sin\phi_{\ell}^{\mathrm{y}} \end{cases}, \tag{2.34}$$

Algorithm 2. Beamspace ESPRIT-D.

Inputs: The received signal in the form of a 5D tensor $\mathbf{Y}_{5\mathrm{D}}$ as in (2.32); Transmitted signal matrix $\mathbf{S}$; The combiner matrices $\mathbf{W}_{\mathrm{x}}$ and $\mathbf{W}_{\mathrm{y}}$ that can construct $\mathbf{W}$; The precoder matrices $\mathbf{F}_{\mathrm{x}}$ and $\mathbf{F}_{\mathrm{y}}$ that can construct $\mathbf{F}$; The number of taps N_d; The delay dictionary resolution G_t; The number of channel paths N_L to be estimated.

Initialize system parameters:

The rotation matrices $\mathbf{J}_z^{(1)}$ and $\mathbf{J}_z^{(2)}$ as defined in (2.21);

The matrices: $\mathbf{B}_1 = \mathbf{W}_{\mathrm{x}}$, $\mathbf{B}_2 = \mathbf{W}_{\mathrm{y}}$, $\mathbf{B}_3 = \bar{\mathbf{F}}_{\mathrm{x}}$, $\mathbf{B}_4 = \bar{\mathbf{F}}_{\mathrm{y}}$, and $\mathbf{B}_5 = \bar{\mathbf{S}}$;

The number of atoms N_z in $\mathbf{u}'_z$ for each dimension: $N_1 = M_{\mathrm{r}}^{\mathrm{x}}$, $N_2 = M_{\mathrm{r}}^{\mathrm{y}}$, $N_3 = M_{\mathrm{t}}^{\mathrm{x}}$, $N_4 = M_{\mathrm{t}}^{\mathrm{y}}$, and $N_5 = Q$;

The delay dictionary $\mathbf{P}$ as in (2.37);

The tensor decomposition results by solving (2.33): $\hat{\mathbf{U}}'_z = [\hat{\mathbf{u}}'_{z,1}, ..., \hat{\mathbf{u}}'_{z,\ell}, ..., \hat{\mathbf{u}}'_{z,N_L}]$.

% *Angle estimation*

for $z = [1:4]$ **do**

  Calculate the diagonal matrix $\mathbf{\Phi}'_z$ as in (2.23);

  Calculate the adjustment matrix $\mathbf{\Psi}_z$ as in (2.24);

  Acquire $\hat{\mathbf{\Gamma}}_z$ as in (2.26), and its eigenvalues $\mathrm{eig}(\hat{\mathbf{\Gamma}}_z)$;

  Estimated frequency components: $\hat{\boldsymbol{\omega}}_z = -j \ln\left(\mathrm{eig}\left(\hat{\mathbf{\Gamma}}_z\right)\right)$;

  if $z \in \{1,2\}$ **then**

    Estimated AoAs: $\begin{cases} \hat{\theta}_\ell^{\mathrm{y}} = \left[\sin^{-1}\left(\frac{\hat{\boldsymbol{\omega}}_2}{\pi}\right)\right]_\ell; \\ \hat{\theta}_\ell^{\mathrm{x}} = \sin^{-1}\left(\frac{[\hat{\boldsymbol{\omega}}_1]_\ell}{-\pi\cos\hat{\theta}_\ell^{\mathrm{y}}}\right). \end{cases}$

  else

    Estimated AoDs: $\begin{cases} \hat{\phi}_\ell^{\mathrm{y}} = \left[\sin^{-1}\left(\frac{\hat{\boldsymbol{\omega}}_4}{\pi}\right)\right]_\ell; \\ \hat{\phi}_\ell^{\mathrm{x}} = \sin^{-1}\left(\frac{[\hat{\boldsymbol{\omega}}_3]_\ell}{\pi\cos\hat{\phi}_\ell^{\mathrm{y}}}\right). \end{cases}$

  end if

end for

% *Delay estimation*

for $\ell = [1:N_L]$ **do**

  Acquire the estimation of $\eta_\ell \mathbf{p}(\tau_\ell)$ as in (2.39);

  Calculate the distributed correlation $\mathbf{c}$ as in (2.40);

  The index of the estimated delay in $\mathbf{P}$: $i_\ell = \arg\max_{i_\ell} \mathbf{c}$;

  The estimated delay: $\hat{\tau}_\ell = \ddot{\imath}_{i_\ell}$;

end for

Calculate gain estimations $\boldsymbol{\gamma}$ as in (2.42);

so that $\mathbf{u}_{1,\ell} = \mathbf{a}(\theta_\ell^{\mathrm{x}})$, $\mathbf{u}_{2,\ell} = \mathbf{a}(\theta_\ell^{\mathrm{y}})$, $\mathbf{u}_{3,\ell} = \bar{\mathbf{a}}(\phi_\ell^{\mathrm{x}})$ and $\mathbf{u}_{4,\ell} = \bar{\mathbf{a}}(\phi_\ell^{\mathrm{y}})$. Hereby $\dot{\mathbf{H}}_\ell$ defined in (2.31) becomes

$$\dot{\mathbf{H}}_\ell = \circ_{z=1}^{4} \mathbf{B}_z^* \mathbf{u}_{z,\ell} \approx \circ_{z=1}^{4} \hat{\mathbf{u}}'_{z,\ell}. \tag{2.35}$$

Based on beamspace ESPRIT mentioned in (2.21)-(2.26), the frequency components of each dimension are estimated referring to $\text{eig}(\hat{\mathbf{\Gamma}}_z)$

$$\hat{\omega}_{z,\ell} = -j \ln\left([\text{eig}(\hat{\mathbf{\Gamma}}_z)]_\ell\right), \tag{2.36}$$

and the angle estimates are retrieved using (2.34).

Delay estimation: Previous work [67,95,111,112] neglecting the pulse shaping and filtering effects express the channel at the k-the subcarrier as $\mathbf{H}[k] = \sum_{\ell=1}^{L} \alpha_\ell \dot{\mathbf{a}}(\boldsymbol{\theta}_\ell)\dot{\mathbf{a}}(\boldsymbol{\phi}_\ell)^* e^{-j2\pi k\Delta f \tau_\ell}$, where Δf denotes the subcarrier spacing. In such cases, beamspace ESPRIT remains applicable considering $\mathbf{B}_5 = \mathbf{I}_K$ with K being the number of subcarriers, and $\omega_{5,l} = -2\pi\Delta f \tau_\ell$. In our model, however, the delay response $\mathbf{p}$ does not follow the formulation of a steering vector, and the decomposed vectors of the 5th dimension should be converted into delay estimates. Consequently, a dictionary based method emerges as the most viable solution to estimate the delay through sparse recovery. We first construct the delay dictionary $\mathbf{P}$ as:

$$\mathbf{P} = [\mathbf{p}(\ddot{\imath}_1), \ldots, \mathbf{p}(\ddot{\imath}_{G_t})], \tag{2.37}$$

where G_t determines the dictionary resolution, and the g-th column is the delay response w.r.t time $\ddot{\imath}_g = (g-1)\cdot\frac{N_d T_s}{G_t}$. From (2.32) and (2.33), we know that $\alpha_\ell\sqrt{P_\text{t}}\mathbf{S}^\mathsf{T}\mathbf{p}(\tau_\ell)$ corresponds to $\gamma_\ell\hat{\mathbf{u}}'_5$, and $\|\gamma_\ell\| \propto \|\alpha_\ell\sqrt{P_\text{t}}\mathbf{S}^\mathsf{T}\mathbf{p}(\tau_\ell)\|$. As we now focus on delay estimation, we treat $\frac{\alpha_\ell\sqrt{P_\text{t}}\mathbf{S}^\mathsf{T}\mathbf{p}(\tau_\ell)}{\gamma_\ell}$ as $\eta_\ell\mathbf{S}^\mathsf{T}\mathbf{p}(\tau_\ell)$, which can be approximated by $\hat{\mathbf{u}}'_5$. The delay information for the ℓ-th path can be estimated by solving the sparse recovery problem:

$$\min_{\mathbf{x}_\ell} \|\mathbf{P}\mathbf{x}_\ell - \eta_\ell\mathbf{p}(\tau_\ell)\|, \tag{2.38}$$

where $\eta_\ell\mathbf{p}(\tau_\ell)$ is obtained by LS estimation:

$$\widehat{\eta_\ell\mathbf{p}(\tau_\ell)} = (\mathbf{S}\mathbf{S}^\mathsf{T})^\dagger\mathbf{S}\hat{\mathbf{u}}_{5,\ell}. \tag{2.39}$$

We calculate the distributed correlation of $\mathbf{P}$ and $\eta_\ell \mathbf{p}(\tau_\ell)$ as

$$\mathbf{c} = \mathbf{P}^* \widehat{\eta_\ell \mathbf{p}(\tau_\ell)}, \tag{2.40}$$

then the maximum projection gives the index of the support of the sparse vector $\mathbf{x}_\ell$ as

$$i_\ell = \arg\max_{i_\ell} \mathbf{c}, \tag{2.41}$$

so the delay estimation is given by $\hat{\tau}_\ell = \ddot{t}_{i_\ell}$.

Finally, let $\boldsymbol{\gamma} = [\gamma_1, ..., \gamma_{N_L}]^\mathsf{T}$, then γ_ℓ can be retrieved by solving (2.42) via LS estimation:

$$\min_{\boldsymbol{\gamma}} \left\| \text{vec}(\mathbf{Y}_{5\text{D}}) - \left[\text{vec}(\circ_{z=1}^{5} \hat{\mathbf{u}}_{z,1}), ..., \text{vec}(\circ_{z=1}^{5} \hat{\mathbf{u}}_{z,N_L}) \right] \boldsymbol{\gamma} \right\|. \tag{2.42}$$

The ESPRIT-D algorithm exhibits spatial complexity of $\mathcal{O}\left(\sum_{z=1}^{4} N_z M_z + N_5 G\right)$, as outlined in Algorithm 2, where each N_z corresponds to the response vector length, M_z relates to number of beams, and G depends on the resolution of the dictionary for the delay. In comparison to the time-domain dictionary based channel estimation methods [109], which suffer from a spatial complexity of $\mathcal{O}\left(\prod_{z=1}^{5} N_z G_z\right)$ with each G_z linked to dictionary resolution along the dimension, resulting in exponential complexity growth with finer resolutions, our proposed method offers a much more practical and manageable computational approach. Our implementation of ESPRIT-D in MATLAB can be found in [113].

2.4 Hybrid Model/Data Driven Precise Localization

2.4.1 *PathNet* for LOS and First-Order Reflection Identifications

To obtain the user position given the channel path parameters we can exploit different geometric relationships for the LOS and NLOS cases. In general, only the parameters of the LOS and first-order paths are leveraged for localization. By exploiting the laws of physics, it

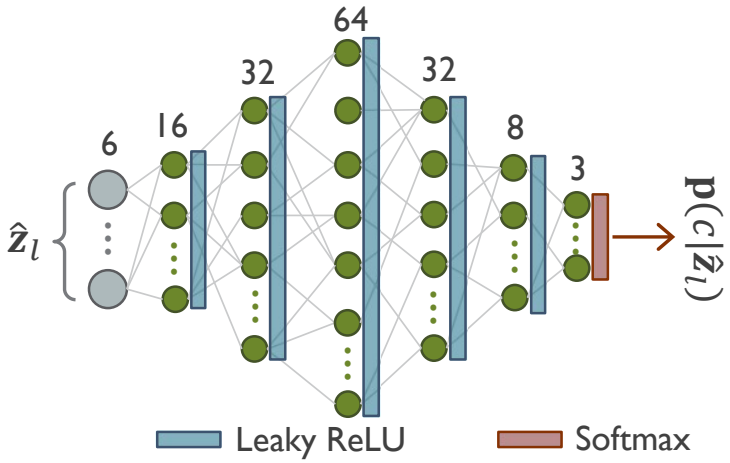


Figure 2.2. Architecture of *PathNet*.

is possible to define a mathematical model to decide if a channel path is a LOS or a first-order reflection given their parameters (see [105] for example). In practice, when applying the model to an estimated path, there will be numerous misclassifications due to the channel estimation error. The introduction of parameters of misclassified paths into the geometric equations that exploit the first-order or LOS nature of the path for localization will lead to a high positioning error. To overcome this limitation, we design in this section a path classification network, named *PathNet*, very robust to channel estimation errors.

To build a suitable network, we study the correlations between the channel parameters and the path order, and as expected, $|\alpha|$, TDoA τ, and azimuth and elevation DoAs/DoDs θ^{x}, θ^{y}, ϕ^{x}, ϕ^{y} are related to the path order, while the phase of the path is uncorrelated with its order. Therefore, we define the input of the network to be the normalized version of all the path parameters but the phase, denoted as $\mathbf{z} = [|\alpha|^2, \tau, \theta^{\mathrm{x}}, \theta^{\mathrm{y}}, \phi^{\mathrm{x}}, \phi^{\mathrm{y}}]$. For localization purposes, every estimated path $\hat{\mathbf{z}}_\ell$ must be classified into one of the three following categories: LOS ($c = 1$), first-order reflections ($c = 2$), or others ($c = 3$). This classification is performed given the probability vector output from *PathNet*, which is defined as

$$\mathbf{p}(c|\hat{\mathbf{z}}) = \mathcal{F}(|\alpha|^2, \tau, \theta^{\mathrm{x}}, \theta^{\mathrm{y}}, \phi^{\mathrm{x}}, \phi^{\mathrm{y}}; \boldsymbol{\mu}), \tag{2.43}$$

where $[\mathbf{p}(c|\hat{\mathbf{z}})]_i = p(c=i|\hat{\mathbf{z}}, i \in \{1,2,3\})$, $\mathcal{F}(\cdot)$ represents the operations performed by *PathNet*, and $\boldsymbol{\mu}$ represents the network parameters to be trained. Hence, among N_{est} estimated paths of a channel, the LOS and first-order reflections can be identified according to

$$\hat{c}(\hat{\mathbf{z}}) = \arg \max_{i \in \{1,2,3\}} p(c=i|\hat{\mathbf{z}}). \tag{2.44}$$

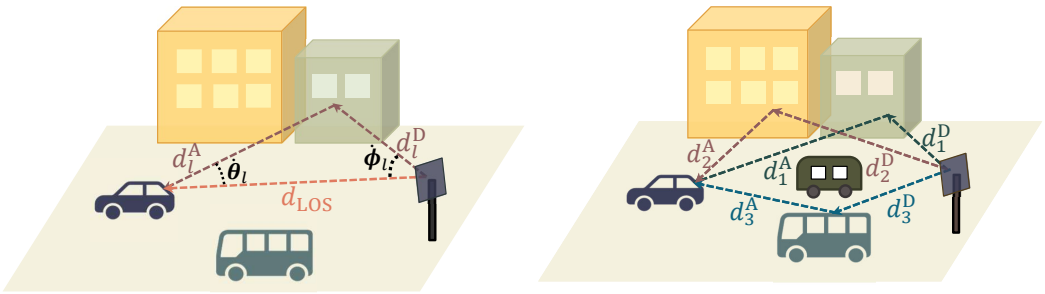
Unlike images, the input parameters to PathNet do not exhibit visible local features, and unlike in natural language processing (NLP) problems, they do not contain context-sensitive or historical information. Therefore, we simply adopt FC layers as the major components of the network, which provide a low complexity solution to learning non-linear combinations of features embedded in the input. The proposed architecture is shown in Fig. 2.2. To select an effective loss function to train the network, we notice first that a higher penalization needs to be applied when classifying a second or high-order path as LOS/first-order reflection, since this would significantly deteriorate the localization performance. Regarding the misclassification of LOS/first-order reflections as high-order paths, a lower penalization should be applied, since they usually indicate an inaccurate channel parameter estimation, so that it is beneficial to discard those paths for localization. With this in mind, we propose a weighted cross-entropy loss instead of a regular cross-entropy loss to adjust the penalties, i.e.

$$\mathcal{L}(\boldsymbol{\mu}) = -e^{-\eta(c(\mathbf{z})-\hat{c}(\mathbf{z}))} \cdot [\mathbf{p}(c|\mathbf{z})]_{c(\mathbf{z})}, \tag{2.45}$$

where η is the customized weight coefficient.

2.4.2 Geometric Localization w.r.t LOS/NLOS Scenarios

In this section, we propose two different geometric localization strategies for the LOS and NLOS scenarios, both accounting for the clock offset between the TX and RX.



(a) **(b)**

Figure 2.3. Illustration of the geometric relationships to be exploited for localization in (a) LOS+NLOS and (b) NLOS scenarios.

LOS+NLOS Scenario

We consider the geometry that can be exploited in the LOS+NLOS scenario as illustrated in Fig. 2.3a. We define the angle between the DoAs of the LOS path and the l-th multipath component as $\dot{\theta} = \arccos(\boldsymbol{\theta}_{\mathrm{LOS}}^{\mathsf{T}}\boldsymbol{\theta}_\ell)$. Similarly, $\dot{\phi}_\ell = \arccos(\boldsymbol{\phi}_{\mathrm{LOS}}^{\mathsf{T}}\boldsymbol{\phi}_\ell)$ represents the angle between the DoDs. For any pair of rays composed of the LOS and any first-order reflection we can apply the Law of Sines as

$$\frac{d_{\mathrm{LOS}}}{\sin(\dot{\theta}_\ell + \dot{\phi}_\ell)} = \frac{d_\ell^{\mathrm{D}}}{\sin(\dot{\theta}_\ell)} = \frac{d_\ell^{\mathrm{A}}}{\sin(\dot{\phi}_\ell)}, \tag{2.46}$$

the where $d_{\mathrm{LOS}} = ||\mathbf{x}_{\mathrm{t}} - \mathbf{x}_{\mathrm{r}}||$ is the distance between the positions of the TX, $\mathbf{x}_{\mathrm{t}}$, and the RX, $\mathbf{x}_{\mathrm{r}}$, which can be computed as $d_{\mathrm{LOS}} = v_{\mathrm{c}} t_{\mathrm{LOS}}$, where v_{c} is the speed of light and t_{LOS} is the time of flight; d_ℓ^{D} (d_ℓ^{A}) is the distance between the TX (RX) and the interaction point on any surface, so that $d_\ell^{\mathrm{D}} + d_\ell^{\mathrm{A}} = v_{\mathrm{c}} t_\ell$. Considering these definitions we have

$$d_\ell^{\mathrm{D}} + d_\ell^{\mathrm{A}} - d_{\mathrm{LOS}} = v_{\mathrm{c}}(t_\ell - t_{\mathrm{LOS}}) = v_{\mathrm{c}}\tau_\ell, \tag{2.47}$$

where τ_ℓ is the TDoA between the l-th first-order reflection and the LOS. Combining (2.46) and (2.47), d_{LOS} can be written as

$$\hat{d}_{\text{LOS}} = \frac{v_{\text{c}}\tau_\ell \sin(\dot{\theta}_\ell + \dot{\phi}_\ell)}{\sin(\dot{\theta}_\ell) + \sin(\dot{\phi}_\ell) - \sin(\dot{\theta}_\ell + \dot{\phi}_\ell)}. \tag{2.48}$$

Let us now define the vectors $\boldsymbol{\tau} = [\tau_1, ..., \tau_{L_{c=2}}]^{\mathsf{T}}$, $\dot{\boldsymbol{\theta}} = [\dot{\theta}_1, ..., \dot{\theta}_{L_{c=2}}]^{\mathsf{T}}$, and $\dot{\boldsymbol{\phi}} = [\dot{\phi}_1, ..., \dot{\phi}_{L_{c=2}}]^{\mathsf{T}}$. When the number of estimated first-order reflections $L_{c=2} \geq 1$, then $\hat{d}_{\text{LOS}}$ can be obtained by solving a LS problem with solution

$$\hat{d}_{\text{LOS}} = \frac{< v_{\text{c}} \cdot \boldsymbol{\tau} \odot \sin(\dot{\boldsymbol{\theta}} + \dot{\boldsymbol{\phi}}), \sin(\dot{\boldsymbol{\theta}}) + \sin(\dot{\boldsymbol{\phi}}) - \sin(\dot{\boldsymbol{\theta}} + \dot{\boldsymbol{\phi}}) >}{\| \sin(\dot{\boldsymbol{\theta}}) + \sin(\dot{\boldsymbol{\phi}}) - \sin(\dot{\boldsymbol{\theta}} + \dot{\boldsymbol{\phi}})\|^2}. \tag{2.49}$$

Finally, the vehicle location could be determined as

$$\hat{\mathbf{x}}_{\text{r}} = \mathbf{x}_{\text{t}} + \hat{d}_{\text{LOS}} \cdot \boldsymbol{\phi}_{\text{LOS}}. \tag{2.50}$$

NLOS Scenario

In this case, illustrated in Fig. 2.3b, the geometric equations for path l could be created with an extension of (2.47) as

$$\begin{cases} \mathbf{x}_{\text{r}} + \boldsymbol{\theta}_\ell d_\ell^{\text{A}} = \mathbf{x}_{\text{t}} + \boldsymbol{\phi}_\ell d_\ell^{\text{D}} \\ d_\ell^{\text{A}} + d_\ell^{\text{D}} = \Delta d_\ell + d_0 \end{cases}, \tag{2.51}$$

where $\Delta d_\ell = v_{\text{c}}(t_\ell - t_0)$, and $d_0 = v_{\text{c}} t_0$. The vehicle location can be now expressed as

$$\mathbf{x}_{\text{r}} = \mathbf{x}_{\text{t}} + (\boldsymbol{\phi}_\ell + \boldsymbol{\theta}_\ell) d_\ell^{\text{D}} - \boldsymbol{\theta}_\ell(\Delta d_\ell + d_0), \tag{2.52}$$

with d_ℓ^{D} is estimated as

$$\hat{d}_\ell^{\text{D}} = \frac{< \boldsymbol{\phi}_\ell + \boldsymbol{\theta}_\ell, \mathbf{x}_{\text{r}} - \mathbf{x}_{\text{t}} + \boldsymbol{\theta}_\ell(\Delta d_\ell + d_0) >}{||\boldsymbol{\phi}_\ell + \boldsymbol{\theta}_\ell||^2}. \tag{2.53}$$

Now we substitute d_ℓ^{D} in (2.52) with the expression in (2.53), and define $\Theta_\ell = \frac{(\boldsymbol{\theta}_\ell+\boldsymbol{\phi}_\ell)(\boldsymbol{\theta}_\ell+\boldsymbol{\phi}_\ell)^\mathsf{T}}{||\boldsymbol{\theta}_\ell+\boldsymbol{\phi}_\ell||^2}$. Considering these definitions, the vehicle position can be expressed now as

$$\mathbf{x}_\mathrm{r} = (\mathbf{I}-\Theta_\ell)\mathbf{x}_\mathrm{t} + \Theta_\ell\mathbf{x}_\mathrm{r} - (\mathbf{I}-\Theta_\ell)\boldsymbol{\theta}_\ell(\Delta d_\ell + d_0), \tag{2.54}$$

or alternatively

$$(\mathbf{I}-\Theta_\ell)(\mathbf{x}_\mathrm{r}+\boldsymbol{\theta}_\ell d_0) = (\mathbf{I}-\Theta_\ell)[\mathbf{I},\boldsymbol{\theta}_\ell][\mathbf{x}_\mathrm{r};d_0] \tag{2.55}$$

$$= (\mathbf{I}-\Theta_\ell)(\mathbf{x}_\mathrm{t}-\boldsymbol{\theta}_\ell\Delta d_\ell). \tag{2.56}$$

A least square estimation problem can be formulated, i.e., $[\hat{\mathbf{x}}_\mathrm{r};\hat{d_0}] = \mathbf{A}^{-1}\mathbf{b}$, where

$$\begin{cases} \mathbf{A} = \sum_{l=1}^{L_{c=2}}[\mathbf{I},\boldsymbol{\theta}_\ell]^\mathsf{T}(\mathbf{I}-\Theta_\ell)[\mathbf{I},\boldsymbol{\theta}_\ell] \\ \mathbf{b} = \sum_{l=1}^{L_{c=2}}[\mathbf{I},\boldsymbol{\theta}_\ell]^\mathsf{T}(\mathbf{I}-\Theta_\ell)(\mathbf{x}_\mathrm{t}-\boldsymbol{\theta}_\ell\Delta d_\ell) \end{cases},$$

to obtain the 3D vehicle position $\mathbf{x}_\mathrm{r}$ and the clock offset. Because the rank of Θ_ℓ is 1 which leads to matrix $(\mathbf{I}-\Theta_\ell)$ being rank 2, and the matrix $[\mathbf{I},\boldsymbol{\theta}_\ell]$ is of rank 3, the rank of $(\mathbf{I}-\Theta)[\mathbf{I},\boldsymbol{\theta}_\ell]$ is $\min\{2,3\} = 2$. Accordingly, the solution of the least square estimation problem is unique when at least 3 estimated first-order reflections are present. In addition, after computing $\hat{\mathbf{x}}_\mathrm{r}$, the location of the reflection points can be determined by introducing $\hat{\mathbf{x}}_\mathrm{r}$ into (2.51).

For both the LOS+NLOS and NLOS scenarios, multiple combinations of paths could exist, and will yield to different location estimates. In such a case, iterating over various combinations of paths and discarding implausible localization results, e.g., those with unrealistic height estimations $\hat{x}_\mathrm{r}^\perp = [\hat{\mathbf{x}}_\mathrm{r}]_3$, can lead to better localization results.

2.4.3 *ChanFormer* for Refining 2D Location Estimates

Instead of designing a network to solve the challenging regression problem of estimating the user position given the received signal, or even the channel parameters, we consider the design

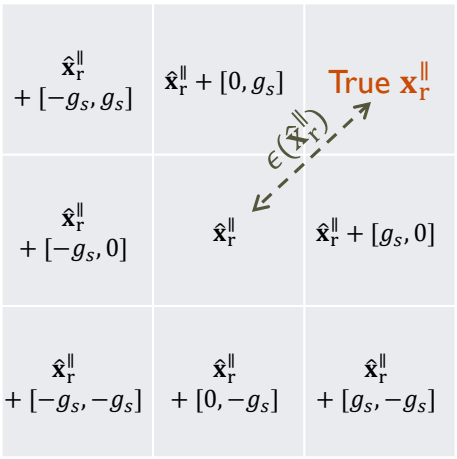


Figure 2.4. Illustration of a $N_g \times N_g = 3 \times 3$ tile structure for position refinement.

of a network for position refinement after obtaining an initial estimate by geometric localization. With this approach in mind, we will formulate the position refinement problem as a classification task, which is usually less challenging for an ML-based approach. To this aim, we consider a $N_g \times N_g$ tile structure with a specified grid size g_s around the initial 2D location estimate $\hat{\mathbf{x}}_{\mathrm{r}}^{\shortparallel}$ (obtained from the 3D location estimate $\hat{\mathbf{x}}_{\mathrm{r}}$ provided by geometric localization), as shown in the example in Fig. 2.4. Our goal in this section is to create a network called *ChanFormer* that obtains the probability of a tile containing the true location $\mathbf{x}_{\mathrm{r}}^{\shortparallel}$. Mathematically,

$$p\left(\tilde{\mathbf{x}}_{\mathrm{r}}^{\shortparallel}|\hat{\mathbf{Z}}\right) = p\left(\mathbf{x}_{\mathrm{r}}^{\shortparallel} = \tilde{\mathbf{x}}_{\mathrm{r}}^{\shortparallel}|\hat{\mathbf{Z}}, \tilde{\mathbf{x}}_{\mathrm{r}}^{\shortparallel} = \hat{\mathbf{x}}_{\mathrm{r}}^{\shortparallel} + [n_{\mathrm{x}} g_s, n_{\mathrm{y}} g_s]^{\mathsf{T}}\right), \tag{2.57}$$

where $\hat{\mathbf{Z}} = [\hat{\mathbf{z}}_1; ...; \hat{\mathbf{z}}_{N_{\mathrm{est}}}]$ is the estimated channel containing N_{est} estimated paths, and $|n_{\mathrm{x}}|, |n_{\mathrm{y}}| \in \left\{0, 1, ..., \frac{N_g - 1}{2}\right\}$. $p(\tilde{\mathbf{x}}_{\mathrm{r}}^{\shortparallel}|\hat{\mathbf{Z}})$ should be negatively related to the distance between $\tilde{\mathbf{x}}_{\mathrm{r}}^{\shortparallel}$ and $\mathbf{x}_{\mathrm{r}}^{\shortparallel}$, which is formulated as :

$$p\left(\tilde{\mathbf{x}}_{\mathrm{r}}^{\shortparallel}|\hat{\mathbf{Z}}\right) = \frac{1}{1 + e^{-\gamma\left(1 - \frac{||\tilde{\mathbf{x}}_{\mathrm{r}}^{\shortparallel} - \mathbf{x}_{\mathrm{r}}^{\shortparallel}||}{\delta}\right)}}, \tag{2.58}$$

where γ is the belief factor, and δ is the scale factor for the distance $\varepsilon(\tilde{\mathbf{x}}_{\mathrm{r}}^{\shortparallel}) = ||\tilde{\mathbf{x}}_{\mathrm{r}}^{\shortparallel} - \mathbf{x}_{\mathrm{r}}^{\shortparallel}||$. *ChanFormer* is meant to analyze $\hat{\mathbf{x}}_{\mathrm{r}}^{\shortparallel}$ and its surroundings $\tilde{\mathbf{x}}_{\mathrm{r}}^{\shortparallel}$ to find the one that most likely meets the current estimated channel condition, i.e., $\max_{\tilde{\mathbf{x}}_{\mathrm{r}}^{\shortparallel}} p(\hat{\mathbf{Z}}|\tilde{\mathbf{x}}_{\mathrm{r}}^{\shortparallel})$. The entire network can be

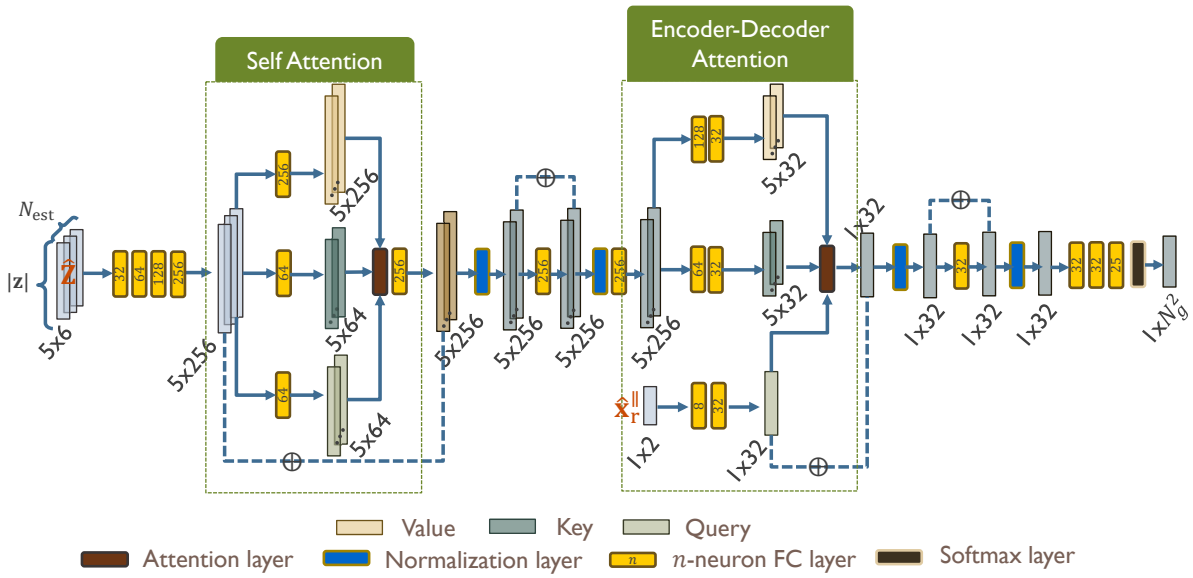


Figure 2.5. Architecture of *ChanFormer*. The self-attention block extracts the intra and crossover features of the input estimated paths. The encoder-decoder block analyzes the relationship between the initial location estimate and the extracted features of the estimated paths.

formulated as

$$\hat{\mathbf{P}}\left(\hat{\mathbf{Z}}, \hat{\mathbf{x}}_{\mathrm{r}}^{\mathrm{II}}\right) = \mathcal{T}\left(\hat{\mathbf{Z}}, \hat{\mathbf{x}}_{\mathrm{r}}^{\mathrm{II}}; \boldsymbol{\omega}\right) \in \mathbb{R}^{1 \times N_g^2}, \tag{2.59}$$

where $\boldsymbol{\omega}$ is the network parameters to be trained. Inspired by the idea of the original *Transformer* [106], the core concept of *ChanFormer* is an encoder for *Self-Attention* to extract features of the input estimated channel $\hat{\mathbf{Z}}$, and a decoder to analyze the relationships between the estimated channel features and the initial location estimate $\hat{\mathbf{x}}_{\mathrm{r}}^{\mathrm{II}}$ using *Encoder-Decoder Attention*. The proposed architecture is shown in Fig. 2.5. **Encoder:** The workflow starts with FC layers embedding the input estimated paths to vectors with a length of 256. Then, the self-attention process begins by creating three abstraction matrices – *query*, *key*, and *value* matrices, denoted as $\mathbf{Q}$, $\mathbf{K}$, and $\mathbf{V}$ respectively, where each row of the matrices corresponds to an estimated path. Conceptually, each $\hat{\mathbf{z}}_\ell$ now has a high-dimensional interpretation of its features in its *value* $[\mathbf{V}]_{l,:}$, which can be indexed by its *key* $[\mathbf{K}]_{l,:}$. The following attention layer then evaluates the

relationships among the paths by

$$\text{Attention}(\mathbf{Q},\mathbf{K},\mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^{\mathsf{T}}}{\sqrt{d_{\text{k}}}}\right)\mathbf{V}, \tag{2.60}$$

where d_{k} is the dimension of $[\mathbf{K}]_{l,:}$, and softmax is for atoms along axis $= 2$. The softmax score determines how much true channel information is represented by the l-th estimated path by examining the correlation between $\hat{\mathbf{z}}_\ell$ and all the paths in $\hat{\mathbf{Z}}$. It is expected that $\hat{\mathbf{z}}_l$ will have the highest softmax score with itself, but other paths that have quite accurate estimations will also be assigned a relatively high score. Therefore, the Attention output is the expression of each path that integrates information from all other paths. The less reliable estimated paths will have a smaller influence on the expressions. This means that the attention mechanism emphasizes more accurate paths in this step, so that the true channel is better represented, regardless of the presence of the noises from the channel estimation process and/or the misclassified paths. In addition, the attention layer is capable of analyzing the input without being constrained by its chronicle orders, which exceeds the capabilities of the convolutional layer that considers path relationships within a fixed window, or the FC layer that relies on connections of all the input parameters.

Decoder: The input $\hat{\mathbf{x}}_{\text{r}}^{\text{II}}$ serves as a *query*, referring to which the network generates the probability map $\mathbf{P}(\hat{\mathbf{Z}})^\star \in \mathbb{R}^{N_g \times N_g}$ of the tiles with $\hat{\mathbf{x}}_{\text{r}}^{\text{II}}$ at the center. Note that, though $\hat{\mathbf{x}}_{\text{r}}^{\text{II}}$ is the only input at the decoder, the network actually evaluates all candidate locations within the tiles given the grid size. In this part, the attention layer improves the initial location estimate accuracy by assigning a higher probability to a candidate location $\widetilde{\mathbf{x}}_{\text{r}}^{\text{II}}$ that aligns better with the channel representation obtained from the encoder. Note that the output $\hat{\mathbf{P}}(\hat{\mathbf{Z}},\hat{\mathbf{x}}_{\text{r}}^{\text{II}})$ is reshaped to $\mathbf{P}(\hat{\mathbf{Z}},\hat{\mathbf{x}}_{\text{r}}^{\text{II}})^\star \in \mathbb{R}^{N_g \times N_g}$ to acquire the probability map to simplify the process of accessing $\widetilde{\mathbf{x}}_{\text{r}}^{\text{II}}$ associated with $p(\widetilde{\mathbf{x}}_{\text{r}}^{\text{II}}|\hat{\mathbf{Z}})$. By referring to the tile with the highest probability:

$$[j^\star, i^\star] = \arg\max_{j,i} \left[\hat{\mathbf{P}}(\hat{\mathbf{Z}},\hat{\mathbf{x}}_{\text{r}}^{\text{II}})^\star\right]_{j,i} \Rightarrow [n_{\text{x}}^\star, n_{\text{y}}^\star] = \left[i^\star - \frac{N_g+1}{2}, \frac{N_g+1}{2} - j^\star\right], \tag{2.61}$$

the refined location is given by:

$$\widetilde{\mathbf{x}}_{\mathrm{r}}^{\shortparallel(\star)} = \hat{\mathbf{x}}^{\shortparallel} + [n_{\mathrm{x}}^{\star} g_s, n_{\mathrm{y}}^{\star} g_s]^{\mathsf{T}}. \tag{2.62}$$

To train the network, we evaluate both MSE loss and Kullback-Leibler (KL) divergence loss for learning the probability map distribution. Denoting the target probability vector as $\mathbf{p}^{\star} \in \mathbb{R}^{N_g^2}$, whose k-th element is $p_k^{\star} = p(\widetilde{\mathbf{x}}_{\mathrm{r},k}^{\shortparallel} | \hat{\mathbf{Z}})$ computed from (2.58) for the center coordinate $\widetilde{\mathbf{x}}_{\mathrm{r},k}^{\shortparallel}$ of the k-th tile, the MSE loss is

$$\mathcal{L}_{\mathrm{MSE}}(\boldsymbol{\omega}) = \frac{1}{N_g^2} \left\| \hat{\mathbf{P}}\left(\hat{\mathbf{Z}}, \hat{\mathbf{x}}_{\mathrm{r}}^{\shortparallel}; \boldsymbol{\omega}\right) - \mathbf{p}^{\star} \right\|^2, \tag{2.63}$$

which penalizes the ℓ_2 deviation of the predicted probability from the target uniformly across all tiles. The KL divergence loss is

$$\mathcal{L}_{\mathrm{KL}}(\boldsymbol{\omega}) = \sum_{k=1}^{N_g^2} p_k^{\star} \log \frac{p_k^{\star}}{\left[\hat{\mathbf{P}}\left(\hat{\mathbf{Z}}, \hat{\mathbf{x}}_{\mathrm{r}}^{\shortparallel}; \boldsymbol{\omega}\right)\right]_k + \varepsilon}, \tag{2.64}$$

where ε is a small constant for numerical stability. While $\mathcal{L}_{\mathrm{MSE}}$ treats all tiles equally, $\mathcal{L}_{\mathrm{KL}}$ places greater emphasis on tiles with high target probability, making it more sensitive to the distributional shape around the true position. The network parameters $\boldsymbol{\omega}$ are optimized by minimizing the chosen loss function using the Adam optimizer with an initial learning rate of 2×10^{-4} and a decay rate of 0.95 per 200 epochs, with early stopping based on the validation loss to prevent overfitting.

2.5 Simulation Results

In this section, we present simulation results to evaluate the performance of the different modules designed in this chapter. We begin by detailing the simulation setup in Sec. 2.5.1. Then, in Sec. 2.5.2, we assess the accuracy of the MOMP-based channel estimation stage. In Sec. 2.5.4,

we demonstrate the results of path classification using *PathNet*. We finally discuss the localization results, including the initial estimation only exploiting geometric relationships and the enhanced performance after applying *ChanFormer*. Note that, *PathNet* is exclusively trained on perfect channels, while the two-stage MOMP based channel estimates are used in the testing stage for *PathNet*, and both training and testing for *ChanFormer*. As channel estimation errors propagate through the system, we analyze their impact on the different stages of our solution, including path order classification, geometric localization, and position refinement using *ChanFormer*.

2.5.1 Simulation Setup

Ray-Tracing Simulation for Realistic Channels: We run 2500 electromagnetic simulations of a vehicular environment in Rosslyn City, Virginia, on a 240×120 m^2 plane, using *Wireless Insite* software [50]. In each simulation, around 30 vehicles are randomly distributed across the four lanes for initial access, with 80% being cars and 20% being trucks according to the 3GPP methodology for simulation of vehicular communication systems [40]. The BS is located at $\mathbf{x}_{\mathrm{t}} = [120, -21, 5]$ m, down facing the road. Parameters for materials of the building/territorial surfaces, the vehicle sizes, placements of antennas on the vehicle and BS, etc., follow the deployments in [7]. 4 active cars are randomly selected to communicate with the BS in the 73 GHz band. With each car equipped with 4 communication arrays, the simulations provide $4 \times 4 \times 2500 = 40$k channels as the dataset $\mathcal{S}$, and every channel has a maximum of $L = 25$ multipath components. The first 24k channels denoted as $\mathcal{S}_{\mathrm{tr}}$ are split into 3 : 1 for training and validations, and the remaining 16k channels serve as the testing set $\mathcal{S}_{\mathrm{te}}$ for all the performance evaluations.

Communication System: In this chapter, we use an antenna setting of $N_{\mathrm{t}} = N_{\mathrm{t}}^{\mathrm{x}} \times N_{\mathrm{t}}^{\mathrm{y}} = 16 \times 16$, $N_{\mathrm{r}} = N_{\mathrm{r}}^{\mathrm{x}} \times N_{\mathrm{r}}^{\mathrm{y}} = 8 \times 8$, and a transmitted power of $P_{\mathrm{t}} = 40$ dBm. The number of RF chains at TX and RX are set to be $N_{\mathrm{t}}^{\mathrm{RF}} = 8$ and $N_{\mathrm{r}}^{\mathrm{RF}} = 4$. The communication system operates at a carrier frequency $f_c = 73$ GHz with a bandwidth $B_c = 1$ GHz. The noise power is computed as $\sigma_{\mathbf{n}}^2 = -84$ dBm, assuming an environmental temperature of $T = 288$ °F. Given the RMS

delay-spread of the simulated channels and the bandwidth, the number of delay taps is fixed to $N_\mathrm{d} = 64$. $N_s = \min\{N_\mathrm{t}^\mathrm{RF}, N_\mathrm{r}^\mathrm{RF}\} = 4$ training data streams with a length of $Q = 64$ are transmitted. We use the raised-cosine filter with a roll-off factor of 0.4 to simulate pulse shaping and other filtering effects in the discrete equivalent channel.

2.5.2 MOMP Based Low Complexity 3D Channel Estimation

The training matrix $\mathbf{F}$ is constructed by the Khatri-Rao product of the precoders along the azimuth and elevation planes, i.e., $\mathbf{F} = \mathbf{F}^\mathrm{x} \circ \mathbf{F}^\mathrm{y}$. Each column of $\mathbf{F}^\mathrm{x}$ ($\mathbf{F}^\mathrm{y}$) is extracted from the DFT codebook of size N_t^x (N_t^y), e.g., $\forall [\mathbf{F}^\mathrm{x}]_{:,i} \in \left\{ \mathbf{a}'(\varphi) | \varphi = 0, \frac{2\pi\cdot 1}{N_\mathrm{t}^\mathrm{x}}, \ldots, \frac{2\pi\cdot(N_\mathrm{t}^\mathrm{x}-1)}{N_\mathrm{t}^\mathrm{x}} \right\}$, where $\mathbf{a}'(\varphi) = \frac{1}{\sqrt{N_\mathrm{t}^\mathrm{x}}} \left[0, e^{j\cdot 1\cdot\varphi}, \ldots, e^{j\cdot(N_\mathrm{t}^\mathrm{x}-1)\cdot\varphi}\right]^\mathsf{T}$. The same procedure applies to the combiners $\mathbf{W}$. We use a setting of $M_\mathrm{t} = 16$, $M_\mathrm{r} = 64$, and form $\mathbf{Y}_M$ by collecting a total of $M = M_\mathrm{r} \times M_\mathrm{t} = 1024$ frames. The size –or the resolution– of the dictionaries is based on the number of their atoms along the dimension, with a specific constant K_res determining the proportion, i.e., $N_k^\mathrm{a} = K_\mathrm{res} \cdot N_k^\mathrm{s}$. In our case, $N_1^\mathrm{s} = N_\mathrm{t}^\mathrm{x}$, $N_2^\mathrm{s} = N_\mathrm{t}^\mathrm{y}$, and $N_3^\mathrm{s} = N_\mathrm{d}$. The impact of K_res is studied in [85]. Here we set $K_\mathrm{res} = 128$ as it brings a comparable performance to using a higher resolution setting, such as $K_\mathrm{res} = 1024$, while being computationally more efficient.

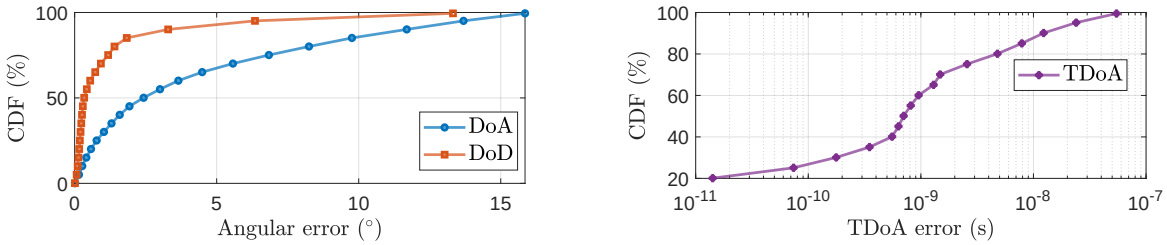


Figure 2.6. MOMP-based channel estimation performance using a setting with a 16×16 array at the TX and a 8×8 array at the RX. Plots acquired based on the whole dataset $\mathcal{S}$.

The angle and delay estimation performance is in Fig. 2.6. The estimation errors are calculated by matching an estimated path to its closest true path in the channel. We obtain DoD estimaties with an average error of 0.5°, and DoA estimates with an average error of 2.5°. This is

reasonable as the TX is equipped with a 16×16 antenna array, while the RX array size is 8×8. In addition, to reduce complexity, the DoA is extracted after DoD estimation, which further reduces the ability of the algorithm to provide high accuracy. An average delay error of $1e-9$ s is also observed. Note that these results include all channel paths, not only first-order reflections, and are deteriorated by the larger estimation error on second or high-order paths. However, this does not reflect the localization accuracy, as *PathNet* identifies the LOS and first-order reflections for localization. We do not introduce any comparison to prior work since there is no existing algorithm that considers the filtering effects, performs time domain estimation so the delays can also be extracted, and can run with the realistic array sizes used in our setup. This is because the memory and computational complexity requirements of the other approaches exceed what a current personal computer or server can provide.

2.5.3 Beamspace ESPRIT-D Channel Estimation

We first show an example of channel estimation results for the 5 strongest paths using ESPRIT-D, as depicted in Fig. 2.7, where the square boxes overlap with "+" markers, and the diamond boxes overlap with "$*$" markers, representing accurate estimations for both AoAs and

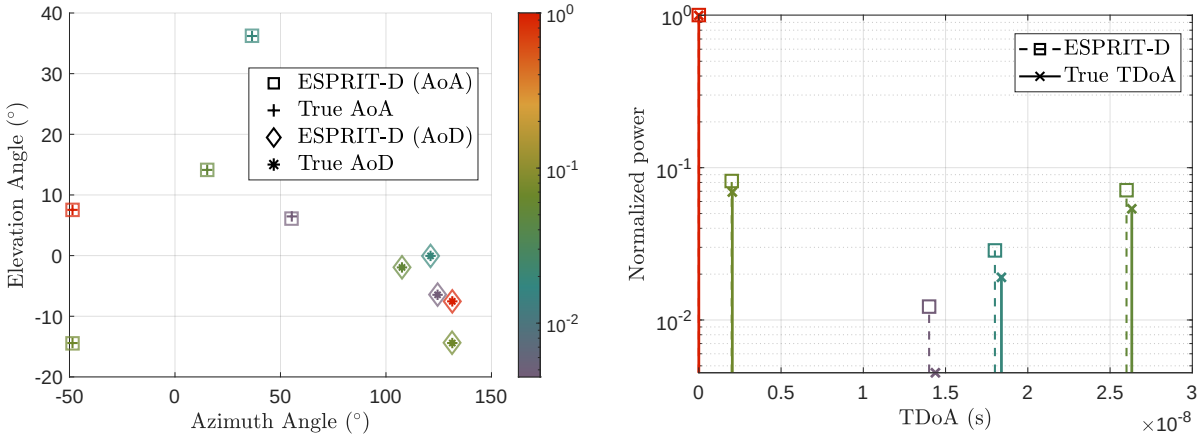
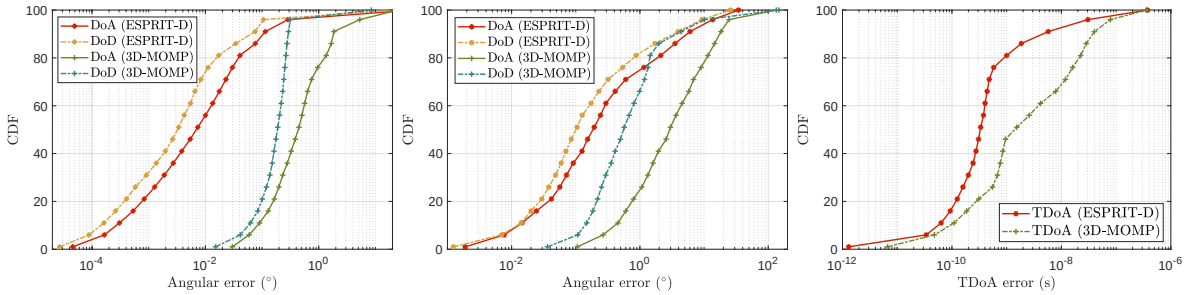


Figure 2.7. An example of channel estimation results of the 5 strongest paths using ESPRIT-D, with the colorbar indicating the normalized power strength. DoA estimation error $\leq 0.38°$; DoD estimation error $\leq 0.11°$; TDoA estimation error $\leq 3.8\text{e}^{-10}$ s.



(a) LoS paths **(b)** NLoS paths **(c)** TDoA

Figure 2.8. CDF of estimation errors based on a total of 200 channels: (a) angular error for LoS paths; (b) angular error for NLoS paths; (c) error in the TDoA.

AoDs. The average estimation errors for AoA and AoD are 0.06° and 0.02°, respectively. The 3D angular estimation error is calculated by $\varepsilon(\hat{\boldsymbol{\varphi}}) = \cos^{-1}(<\hat{\boldsymbol{\varphi}}, \boldsymbol{\varphi}>)$, which are 0.10° and 0.04° in average for DoA and DoD estimations in the instance. In addition, as we assume an unknown clock offset, the delay estimation results are represented by TDoA (s) w.r.t the shortest path in the channel, and the estimation error is 2.84e^{-10} s on average.

The CDF plots in Fig. 2.8 illustrate the estimation errors of angular and TDoA estimations based on the 200 channels, where the LOS paths and NLOS paths are analyzed separately. ESPRIT-D achieves an angular accuracy of 0.04° for 90% of the LOS cases, and an accuracy of 2.35° (AoAs) / 1.20° (AoDs) for 90% of the NLOS cases. AoD estimations exhibit a higher accuracy compared to AoA estimations, primarily due to the larger array employed at the RSU. TDoA estimation with ESPRIT-D keeps the errors mostly $\leq 4.57\text{e}^{-9}$ s. This level of precision is particularly valuable for applications such as vehicle positioning. Compared with results using MOMP [107], the proposed method brings an improvement of more than $10\times$ on average for both the angular and delay estimations. Note that outliers with significant deviations from the ground truth persist, possibly attributed to the intricacies of real-world outdoor channels, such as large power attenuation ($\geq$ 30 dB) for NLOS paths within a LOS channel [107].

2.5.4 *PathNet* for LOS and First-Order Reflection Identifications

PathNet is trained based on $\mathcal{S}_{\text{tr}}$, which lasts for 1000 epochs with an early stopping depending on the convergence of the validation loss. The customized weight for tweaking the penalty in $\mathcal{L}(\boldsymbol{\mu})$ is set to $\eta = 0.2$. We adopt Adam optimizer [114], and set the learning rate $1e-3$ with a decay rate of 0.95 every 200 training epochs. The path classification performance represented by confusion matrices in Fig. 2.9 is evaluated with channels in $\mathcal{S}_{\text{te}}$, where Fig. 2.9a and Fig. 2.9b show the path order classification results for true and MOMP estimated channels, respectively. With the perfect channel parameters, the classification accuracy reaches $\sim 99\%$, highlighting the generalization capability of the simple yet effective network. When using MOMP estimated channels, the classification accuracy reduces to 94.7% for LOS, 90.0% for first order reflections, and 80.5% for other paths. Nevertheless, paths are more likely to be misclassified as higher-order paths rather than as first-order reflections or LOS, which are subsequently discarded from localization. In the rare instances where high-order paths are misclassified as LOS or first-order reflections, mitigation strategies are incorporated, including disregarding paths that yield anomalous height estimates and using the LOS path with the highest power when multiple LOS paths are identified. This degradation from ideal to estimated channels underscores the importance of robust channel estimation in the pipeline.

Table 2.2. Localization error percentiles (m) before and after applying *ChanFormer*. Red percentages in brackets indicate error reduction with *ChanFormer*.

Method	5th	50th	80th	95th	$p(\varepsilon < 1$ m$)$
PathNet+Geo-LOS	0.05	0.44	0.90	1.42	87%
PathNet+Geo-NLOS	0.10	1.73	3.90	5.73	36%
PathNet+Geo-LOS +*ChanFormer*	**0.04**	**0.18** (59% ↓)	**0.28** (69% ↓)	**0.58** (59% ↓)	**98%**
PathNet+Geo-NLOS +*ChanFormer*	**0.07**	**0.77** (55% ↓)	**3.09** (21% ↓)	**5.60**	**55%**

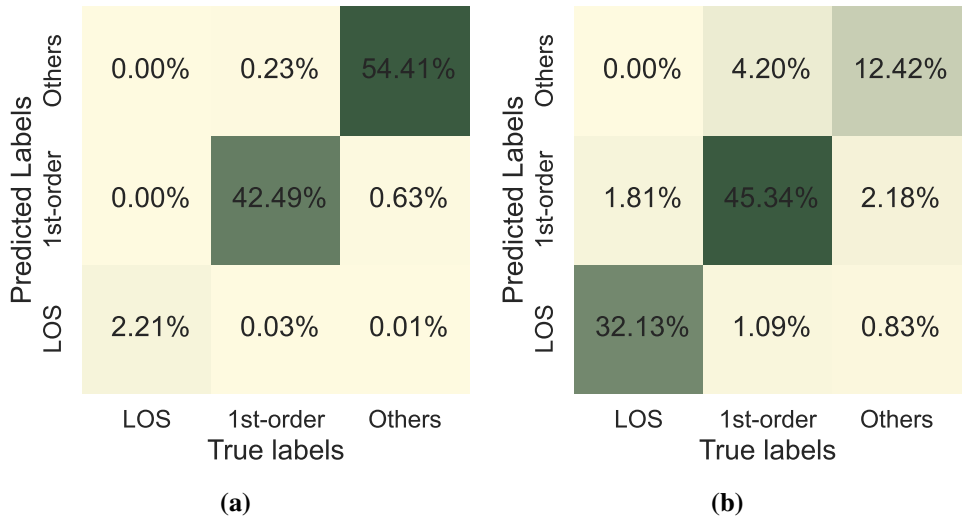


Figure 2.9. Path order classification performance with *PathNet*. (a) Classification with perfect channel parameters (20 paths per channel); (b) Classification with MOMP estimated channels (5 estimated paths per channel).

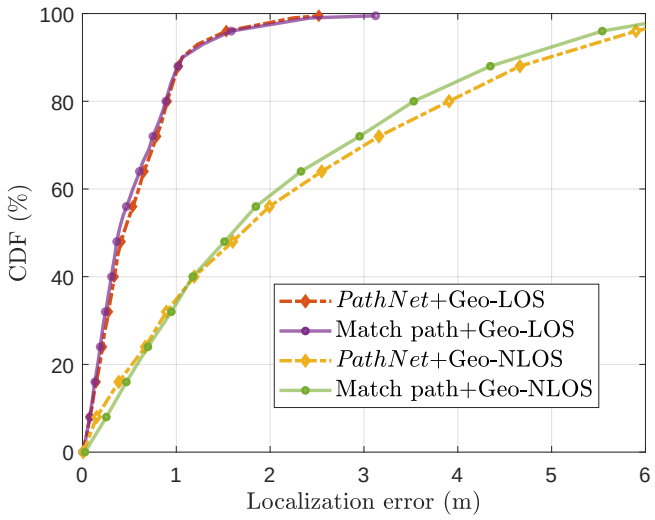


Figure 2.10. Geometric localization performance for MOMP estimated channels in $\mathcal{S}_{\text{te}}^{+}$. "Geo-LOS" and "Geo-NLOS" refer to LOS+NLOS and NLOS-only methods. Results with path orders matched to their closest true-channel counterparts ("Match path") are included for comparison.

2.5.5 Geometric Localization in LOS/NLOS Scenarios

Dataset Preparation

Considering practical localization applications, the vehicle location is calculated based on the array with the strongest received power. Given 4 arrays deployed per vehicle, it reduces the number of channels by $4\times$, resulting in a total of 10k channels. We further examine the database for valid LOS channels ($L_{c=2} \geq 1$, meaning at least one first-order reflection exists alongside the LOS path as required by the geometric localization algorithm) and valid NLOS channels ($L_{c=2} \geq 3$), and establish criteria to exclude channels where the vehicle cannot be located exploiting measurements from a single BS. For LOS channels, we measure the power gap $\Delta|\alpha|^2$ between the LOS and the strongest first-order path. We find that 40% of the channels obtained with Wireless Insite include very weak first-order reflections, with $\Delta|\alpha|^2 > 30$ dB. We assume that if $\Delta|\alpha|^2 > 30$ dB for all the first-order paths, the channel is purely LOS, and the vehicle cannot be located with a single BS due to the lack of strong first-order reflections. Analogously, for valid NLOS channels, we check the received power levels to determine a threshold to ensure a sufficient number of first-order paths (≥ 3) so the vehicle can be located. In particular, we require paths received with an attenuation < 40 dB, and exclude from the database any NLOS channels containing less than 3 qualifying paths. Therefore, the new sets $\mathcal{S}_{\text{tr}}^{+} \in \mathcal{S}_{\text{tr}}$ containing 4085 LOS and 1085 NLOS channels, and $\mathcal{S}_{\text{te}}^{+} \in \mathcal{S}_{\text{te}}$ containing 1385 LOS and 375 NLOS channels are formed. The following evaluations of the localization performance are based on $\mathcal{S}_{\text{te}}^{+}$.

Geometric (Initial) Localization Performance

Fig. 2.10 shows the CDF of localization error (m), and the 5, 50, 80, and 95th-percentile accuracies are presented in Table 2.2. We observe that sub-meter accuracy localization is realized for 87% of the users in LOS channels and 36% of the users in NLOS channels. The compromised performance for NLOS channels is due to the small pool of qualified estimated paths and the decreased accuracy of MOMP channel estimations. Nevertheless, the achieved performance should be considered the worst-case scenario, as the real-world channels are likely to include more

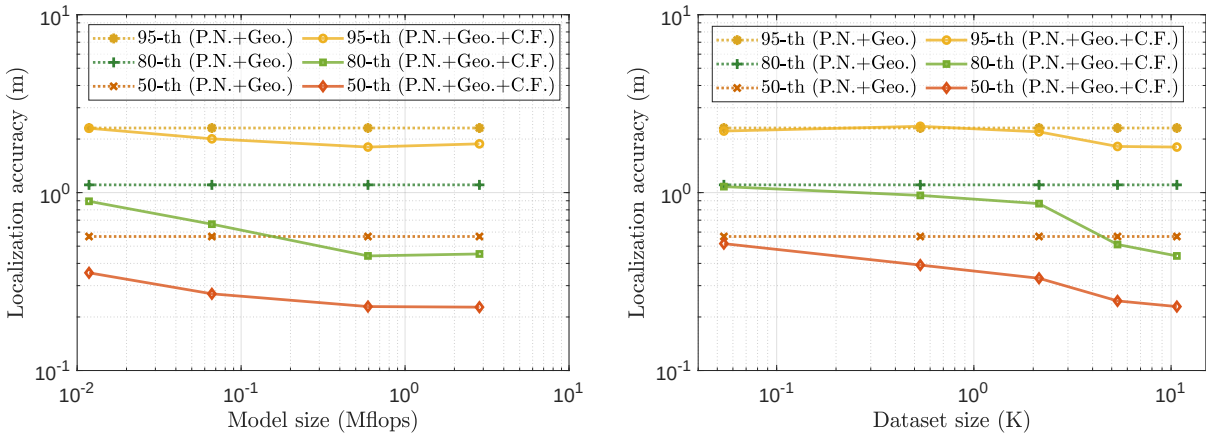


Figure 2.11. Localization refinement using *ChanFormer* with various model and dataset size configurations. "P.N.+Geo." denotes *PathNet*+geometric localization, and "C.F." denotes refinement with *ChanFormer*. (a) Bottlenecked by model size (large dataset used); (b) bottlenecked by dataset size (optimal model size used).

usable reflections with higher power from traffic lights, building windows, and other details of the vehicular scenario which are not present in our electromagnetic simulation of the environment. To study the combined impact of channel estimation errors and *PathNet* classification errors, we also include in Fig. 8 the localization results where the path orders are determined by matching the estimated paths to their closest counterparts in the true channel (denoted as "Match path") instead of using *PathNet* predictions. The performance is comparable for LOS channels, due to the very high path classification accuracy in this case. For NLOS cases, we do observe the impact of the misclassifications and channel estimation errors in performance, since NLOS paths are weaker, their estimation accuracy is lower, and this also increases the likelihood of being misclassified.

2.5.6 2D Location Estimate Refinement with *ChanFormer*

We employ a 5×5 tile structure with a grid size $g_s = 0.4$ m as the output for *ChanFormer*, which is found to be the best option among the test settings. The labels for the 5×5 grids are calculated by setting $\gamma = 5$ and $\delta = 1$ in (2.58), where $p(\bar{\mathbf{x}}_{\mathbf{R}}^{\text{II}})|\mathbf{Z})$ drops to $\leq 0.6\%$ for the ranging

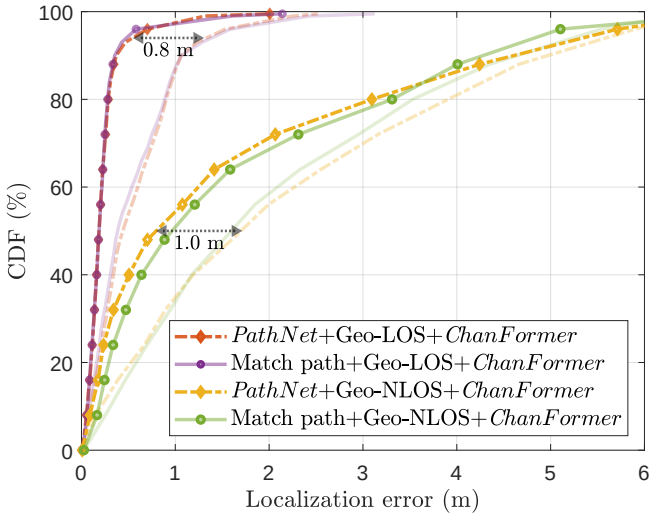


Figure 2.12. Position estimates refined with *ChanFormer*. Transparent lines show the geometric-localization reference (Fig. 2.10). The 95th-percentile 2D error drops by 0.8 m (59% ↓) for LOS users and 1 m (55% ↓) at the median for NLOS users, achieving sub-meter localization.

error $\varepsilon(\bar{\mathbf{x}}^{\text{II}}_{\mathbf{R}}) \geq 2$ m. This network is trained with a batch size of 64 using the Adam optimizer with the learning rate of $2e-4$. To determine the optimal model architecture and dataset sizes, we train the network with various model and dataset configurations. The complexity of the model is varied by adjusting the number of layers and neurons per layer, allowing us to investigate the impact of model size on performance. Additionally, we assess the impact of the dataset size by adjusting the training length. Fig. 2.11 presents the localization accuracy percentiles on $\mathcal{S}^{+}_{\text{te}}$ for various configurations, including initial estimates to highlight *ChanFormer*'s capability to improve position estimation accuracy. The resulting accuracy initially improves as the model size increases. However, a saturation point can also be observed, indicating that the model has become overly complex. On the other hand, expanding the dataset enhances the accuracy, as it provides the network with more comprehensive features to learn

Fig. 2.11a shows the effect of *ChanFormer* model complexity on localization accuracy when a full-scale dataset is used. As model size grows, the refinement benefit increases, with the median error dropping to ∼0.18–0.2 m and the 80th-percentile error to ∼0.3–0.4 m, achieving ∼60–70% improvement over the geometric initial estimate; beyond the optimal size, performance

saturates, indicating that data availability rather than model expressiveness becomes the binding constraint. The 95th-percentile tail improves more modestly from $\sim$1.8–2.0 m to $\sim$1.2–1.4 m ($\sim$30–40%), reflecting the irreducible difficulty of severe NLOS cases where path estimates remain ambiguous regardless of network capacity. Fig. 2.11b reveals the data-scaling behavior using the optimal model: with only $\sim$500 training samples the median error is $\sim$0.4 m, and scaling to $\sim$10K samples reduces it to $\sim$0.2 m ($\sim$50% reduction), with $\sim$60% gains at the 80th percentile and a more modest $\sim$20% at the 95th percentile, consistent with the tail-case difficulty observed in Fig. 2.11a. Beyond $\sim$10K samples the curves flatten, consistent with the saturation behavior seen in Fig. 2.11a. These results confirm that *ChanFormer* is both data-efficient and model-scalable: the dominant geometry-correction behavior is captured with a few thousand samples, and a moderate network size is sufficient to realize the bulk of the refinement benefit. With the optimal model and dataset size, we assess the localization refinement performance based on set $\mathcal{S}_{\text{te}}^{+}$, and the CDF of localization errors is in Fig. 2.12. The performance per percentiles can be found in Table 2.2, highlighting the accuracy improvement when applying *ChanFormer* to geometry based localization results. The refinement reduces the 95th-percentile error to 0.58 m from 1.42 m, i.e., 59% accuracy improvement, for users in the LOS. An error reduction to 0.77 m from 1.73 m, i.e., 55% accuracy improvement, is achieved for half of the users in the NLOS scenario. In conclusion, 98% of the users in LOS and 55% of the users in the NLOS case achieve sub-meter accuracy. The marginally better performance with *PathNet* predicted orders over using matched paths can be attributed to grid resolution constraints. Localization with matching paths has slightly lower errors and undergoes less refinement compared to using *PathNet* determined paths, resulting in a smaller accuracy improvement.

Comparison to Prior Work: Most of the previous studies mentioned in Sec. 2.1.1 assume unrealistic channels and simplistic communication system settings (for example operating with the true channel parameters instead of the estimated ones, or neglecting the clock offset), limiting their performance when evaluated with our data set and system model. To perform comparisons to prior work, we use our realistic ray-tracing simulated channels and signal model

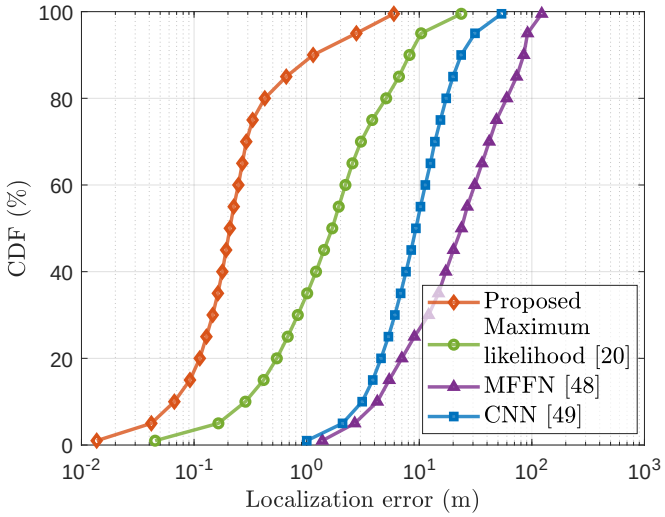


Figure 2.13. Comparison of state-of-the art and proposed scheme. CDF obtained using $\mathcal{S}_{\text{te}}^{+}$.

that accounts for filtering effects. We implemented three different localization strategies in prior work as baselines, one exploiting geometric localization [72] and two other exploiting deep learning architectures [101, 102]. To guarantee a fair comparison, all the approaches exploit the channel parameters estimated with two-stage MOMP as described in Section III.A.3. The localization results with this experimental setting can be found in Fig. 2.13. Our hybrid model/data driven localization method significantly outperforms all the solutions in prior work, achieving sub-meter accuracy for 90% of the users and errors below 30 cm for 50% of the users. Note that the performance obtained with [72] degrades compared to that shown in the original paper due to the use of realistic channels in our simulations instead of ideal ones – with only LOS and first-order reflections– and the introduction of filtering effects. Similarly, the performance degradation of the approach in [101] comes from exploiting the true (not estimated) channel parameters in the original work.

2.6 Conclusion and Future Work

We developed a hybrid model/data-driven approach to obtain 3D localization in a vehicular network operating at mmWave. We considered a realistic channel model accounting for filtering

effects and an unknown TX-RX clock offset. We generated realistic channel datasets for evaluation using ray-tracing. We proposed two low-complexity channel estimation methods suitable for large antenna arrays and wide bandwidths, namely two-stage MOMP and ESPRIT-D, enabling high-resolution 3D channel estimation for precise localization. Then, we designed *PathNet*, a data-driven path classification strategy to select the LOS and first-order paths from the estimated channel, achieving a classification accuracy of 99% (with perfect channel knowledge). We also developed a model-driven 3D positioning strategy which exploits the geometric relationships between the channel parameters and the positions of the BS and the user. This geometric localization strategy can operate in LOS and NLOS channels, providing sub-meter accuracy for 85% of users in LOS channels and for 35% of users in NLOS channels. We developed *ChanFormer*, a location refinement network that enhances channel representation and identifies the most probable vehicle location. After position refinement with *ChanFormer*, 95% of users in the LOS channels achieve a localization accuracy of 0.58 m, and 50% of users in the NLOS channels achieve an accuracy of 0.77 m. Overall, the performance has been improved by $59\% \sim 69\%$ depending on the scenario. The results demonstrate that attention mechanisms are well-suited to the joint localization and communication problem, and highlight the potential of migrating advanced neural network architectures to the field of wireless communications.

The angular and delay errors achieved by two-stage MOMP and ESPRIT-D are within an order of magnitude of the Cramér-Rao bound for mmWave channel-parameter estimation under our system model [115], with the residual gap dominated by dictionary granularity in MOMP and subspace-noise effects in ESPRIT-D. The end-to-end position error is reported as empirical CDFs because the effective likelihood induced by *PathNet* and *ChanFormer* does not admit a closed-form bound; *ChanFormer*'s cross-attention can be interpreted as a learned message-passing step on a multipath factor graph, where each estimated path is a node and the geometric position estimate is the variable being refined.

Several extensions are natural. Multi-RSU channel-parameter fusion is the most direct way to push beyond the single-snapshot positioning limit, as it converts higher-order reflections

that are currently discarded by *PathNet* into additional geometric constraints anchored at distinct base stations. A principled factor-graph / belief-propagation formulation of multipath-based localization, in the spirit of multipath-based SLAM [116], would also provide an interpretable convergence point for the learned refinement and a clean handle on path-association uncertainty. On the evaluation side, set-based metrics such as GOSPA between estimated and ray-traced path parameters would more directly characterize the channel estimator's behavior under births, deaths, and mismatches. Finally, validation with a software-defined mmWave testbed would confirm that the filter-aware estimators and the attention-based refinement remain accurate under hardware impairments not modeled by ray tracing.

Acknowledgements

Chapter 2, in part, is a reprint, with permission, of the material as it appears in the papers: Y. Chen, J. Palacios, N. González-Prelcic, T. Shimizu, H. Lu, "Joint Initial Access and Localization in Millimeter Wave Vehicular Networks: a Hybrid Model/Data Driven Approach", in 2022 IEEE 12th Sensor Array and Multichannel Signal Processing Workshop (SAM 2022), Trondheim, Norway, Jun., 2022; Y. Chen, N. González-Prelcic, T. Shimizu, and H. Lu, "Learning to localize with attention: From sparse mmWave channel estimates from a single BS to high accuracy 3D location," IEEE Trans. Wireless Commun., March 2024 (second round of reviews); and Y. Chen, N. González-Prelcic, T. Shimizu, H. Lu, C. Mahabal, "Beamspace ESPRIT-D for Joint 3D Angle and Delay Estimation for Joint Localization and Communication at MmWave", IEEE International Conference on Communications (IEEE ICC 2024), Denver, USA, Jun., 2024. The dissertation author was the primary investigator and author of these papers, which were partially supported by the NSF under grant no. 2433782, no. 2147955, and in part by funds from the federal agency and industry partners as specified in the Resilient & Intelligent NextG Systems (RINGS) program, and by a gift from Toyota Motor North America, Inc., USA.

Chapter 3

A Hybrid Model/Data-Driven Solution to Channel, Position and Orientation Tracking in mmWave Vehicular Systems

3.1 Introduction

The advancement of mmWave MIMO communication systems employing wide bandwidth and large antenna arrays enables high-resolution channel estimation, including accurate delay and angle acquisitions [109, 117]. Unlike lower frequency bands with dense multipath components (MPCs) [118, 119], mmWave channels exhibit sparsity and facilitate geometric localization [54]. For example, in an outdoor vehicular scenario, the vehicle's location can be derived from high-resolution estimates of the channel between the vehicle and a single BS by exploiting geometric relationships between the path parameters and the locations of the scatterers, the BS, and the vehicle [54]. Therefore, joint channel estimation and localization is a promising technology for real-world deployments as a cost-effective method for precise positioning while meeting the accuracy requirements of automated vehicles in various environments [5]. While accurate single-shot joint channel and position estimation for the *initial access* phase (without including orientation) has been studied in our previous work [120, 121], this chapter focuses on reliable and high-accuracy vehicle position/orientation (PO) *tracking*. Sensor-based PO tracking methods in vehicular settings utilizing IMU [122], cameras [123, 124], LiDAR [125], radar [126],

or sensor fusion [127–129], are well-studied, but often suffer from compromised localization accuracy, e.g., with the GNSS in urban canyons, or reduced reliability under adverse weather or lighting conditions. While solutions relying on mmWave communication signals are a promising alternative, SOTA tracking solutions exploiting the link with a single BS suffer from some limitations, as discussed in Sec. 3.1.1. These methods either fail to model the system realistically or do not achieve the desired localization accuracy for certain use cases when evaluated with practical mmWave communication channels and architectures.

3.1.1 Prior Work

Representative channel-parameter-enabled position tracking methods are presented in [63, 73, 116, 130–143]. Two-stage approaches, in which channel parameters are acquired and subsequently employed for tracking, are studied in [63, 73, 130–135]. In [130], antenna-level carrier phase measurements are used to acquire channel parameters including delays and angels that relate the PO between the BS and the extended reality (XR) devices, and an extended Kalman filter (EKF) is adopted to enable six-degrees-of-freedom (6DoF) tracking. However, the channel parameters are simulated using an error distribution function rather than estimated directly from received signals, which simplifies the complexity of real-world signal acquisition and processing. In contrast, channel parameter acquisition is included in [131] and [73]. In [131], optimal beam selection based on the Fisher information matrix (FIM) is used to maximize the accuracy of delay and angle estimation, after which the channel parameters are tracked with an EKF to obtain the user position, which is then tracked by another EKF. Meanwhile, [73] considers a high-speed outdoor vehicular scenario, addressing the complexity of estimating AoA, ToA, and Doppler shift using a sequential approach, and subsequently realizes localization by solving a weighted least squares (WLS) optimization problem via Newton's method. However, these methods require LOS components and assume perfect synchronization between the transmitter (TX) and RX for the ranging purpose. Methods for tracking user PO in both LOS and NLOS scenarios are proposed in [63, 132]. A tensor decomposition algorithm is proposed in [132] to

extract geometrical channel parameters and track the moving target by referencing the Doppler frequency shift (DFS) through intersecting virtual lines of estimated angles. The solution in [63] introduces a compressed sensing (CS)-based high-resolution multipath parameter estimation method, relating the known BS PO to the user PO by solving a LS estimation problem and nonlinear equations. However, both approaches neglect the filtering effects in the communication system and fail to address higher-order reflections, which negatively affect localization accuracy.

Apart from the model-based solutions, approaches incorporating DL are discussed in [133–136]. In [133], a variational autoencoder architecture is proposed to extract position-related parameters such as TDoA and AoA from channel impulse response (CIR) waveforms, mitigating errors in ranging and angles of NLOS components, and the parameters are fused using a federated filter for user position tracking. Without explicit channel parameter extraction approaches, [134] assumes the availability of ideally simulated channel parameters, and in [135], channel parameters are generated with uncertainties. In such cases, [134] introduces an ensemble-learning way to identify LOS and single-bounce NLOS components for geometrical localization, and adopt an unscented Kalman filter (UKF) together with supplemental odometer data to refine location estimates. A long short-term memory (LSTM) DNN is employed in [135] to extract CSI features in frequency and time domains and aggregates the information of ToA, AoA, and pair-wise received powers, for PO tracking. Furthermore, a fingerprinting solution is presented in [136], where the beamformed fingerprint data is input into a Transformer network to predict the user trajectories.

In addition to the two-stage strategies, joint channel and position tracking approaches are explored in [116, 137–143], leveraging the joint probability distribution of user PO and channel multipath parameters, and employing various filtering methods for user state (PO) tracking. In [137], a factor graph is formulated with a sum-product algorithm (SPA) to calculate marginal posterior distributions of state variables including user PO and channel parameters, enhancing NLOS delay and amplitude estimates, followed by a particle-based implementation for state predictions. The factor graph with belief propagation (BP) methods are commonly applied

to channel simultaneous localization and mapping (SLAM) [116, 138], modeling the state of users (e.g., PO, velocity), physical anchors/BS, virtual anchors (VAs), and the geometry of MPCs, with probability distribution functions (PDFs). In [116], where a super-resolution channel estimation algorithm is used to extract MPCs and higher-order reflections are addressed by incorporating a ray-tracing module, a factor graph representation for the user state, anchor state, and channel measurements is established, a SPA is employed for belief calculation, and finally the user's PO are updated through a minimum mean squared error (MMSE) estimator. In [138], where the factor graph construction remains similar, a continuous measurement correction method incorporating time-sequential measurements is integrated into the BP process, enabling efficient message passing. Besides, in [139, 140], angle-based SLAM are provided where the multipath angle estimates are acquired through beam sweeping. The user PO state is obtained with IMU inputs through particle filtering [139], or a BP framework to calculate PDFs of user states, the anchors, and channel measurements, followed by a MMSE estimator to update user PO [140]. While [139, 140] rely on LOS and first-order reflections, [141–143] address NLOS situations, providing more comprehensive approaches considering multipath birth and disappearance. In [141], a Poisson multi-Bernoulli mixture density is used to represent the joint distribution of environmental landmarks including the anchors, and an EKF is adopted to jointly update motion sensor and landmark states for user PO inference. Alternatively, as in [142], landmark changes are predicted considering a birth probability hypothesis density (PHD) added to the previous landmark PHD, and particle filtering is adopted for updating user POs. In [143], a snapshot SLAM method based on multipath geometry–excluding higher-order reflections–is incorporated to get the initial estimates for a multi-hypothesis linear filter, which contains a nearest neighbor filter for user state tracking and a PHD filter for landmark tracking.

The above methods face several limitations, summarized in Table 3.1. Many studies assume idealized channel multipath parameters without incorporating real-world channel estimation or tracking techniques [130, 134, 136, 138]. For approaches that include estimation or tracking of channel MPCs [63, 73, 116, 131–133, 135, 139–143], simplified communication systems are often

considered by using ULA instead of URA at one or both ends to reduce processing complexity [63, 116,131,132,135,139–143], and filtering effects for time-domain channel processing are neglected [63, 73, 135]. Furthermore, some methods do not address orientation tracking [73, 131, 133, 136]. PO tracking algorithms relying on the presence of LOS paths for ranging [73, 116, 130, 131, 134] assume perfect synchronization between the TX and RX [73, 116, 131, 132, 137, 142], or require round trip time (RTT) measurements to cancel clock biases [133–135]. In addition, algorithms that depend on LOS and/or first-order reflections [132, 133, 138–142] fail to address higher-order reflections negatively affecting the tracking performance. Most methods are evaluated in indoor environments [116, 130, 132, 137–140, 142, 143], while the accuracy can degrade for outdoor complex scenarios. In addition, the complexity of real environments introduces additional challenges for these approaches relying on idealized assumptions and simplified system models, particularly in modeling real-world channel effects and achieving the required localization accuracy. DL can be employed to address aspects that conventional models cannot handle effectively, such as localization errors arising from system noise and accumulated errors that are difficult to represent with closed-form formulations, and some pure DL-based fingerprinting solutions have achieved reasonable accuracy, e.g., root mean squared error (RMSE) of $1 \sim 2$ m [135, 136], but they remain inadequate for achieving submeter-level tracking.

We intend to propose a hybrid approach that combines model-based and data-driven solutions. When closed-form analytical solutions are feasible, we employ low-complexity and effective model-based methods. For tasks that involve capturing complex, nonlinear relationships in high-dimensional data, such as those arising from driver behaviors, environmental noise, raw channel impulse responses, and the spatial configuration of transmitters and receivers, we leverage DL. In these cases, DL is used to extract robust features related to POs from processed signals, ensuring reliable performance even in the presence of noise, multipath propagation, NLOS conditions, and hardware imperfections.

Table 3.1. Capabilities of representative prior channel-parameter-enabled PO tracking methods. ✓/× = property satisfied/not. Columns denote: realistic channel estimation, full URA at both ends, time-domain filtering, orientation tracking, LOS-free ranging, sync-free, RTT-free, high-order reflections, and outdoor evaluation.

Reference	Realistic channel	URA	Filtering	Orientation	LOS-free	Sync-free	RTT-free	High-order refl.	Outdoor
Talvitie *et al.* (2023) [130]	×	✓	✓	✓	×	✓	✓	✓	×
Koivisto *et al.* (2022) [131]	✓	×	✓	×	×	×	✓	✓	✓
Gong *et al.* (2023) [73]	✓	✓	×	×	×	×	✓	✓	✓
Zhao *et al.* (2023) [132]	✓	×	✓	✓	✓	×	✓	×	×
Gómez-Cuba *et al.* (2023) [63]	✓	×	×	✓	✓	✓	✓	✓	✓
Wang *et al.* (2024) [133]	✓	✓	✓	×	✓	✓	×	×	✓
Bader *et al.* (2024) [134]	×	✓	✓	✓	×	✓	×	✓	✓
Klus *et al.* (2024) [135]	✓	×	×	✓	✓	✓	×	✓	✓
Shamsesalehi *et al.* (2024) [136]	×	✓	✓	×	✓	✓	✓	✓	✓
Venus *et al.* (2024) [137]	✓	✓	✓	✓	✓	×	✓	✓	×
Leitinger *et al.* (2023) [116]	✓	×	✓	✓	×	×	✓	✓	×
Gao *et al.* (2024) [138]	×	✓	✓	✓	✓	✓	✓	×	×
Que *et al.* (2023) [139]	✓	×	✓	✓	✓	✓	✓	×	×
Yang *et al.* (2023) [140]	✓	×	✓	✓	✓	✓	✓	×	×
Ge *et al.* (2022) [141]	✓	×	✓	✓	✓	✓	✓	×	✓
Du *et al.* (2024) [142]	✓	×	✓	✓	✓	×	✓	×	×
Kaltiokallio *et al.* (2024) [143]	✓	×	✓	✓	✓	✓	✓	✓	×

3.1.2 Contributions

In this chapter, we propose a novel hybrid model/data-driven framework for precise vehicle PO tracking as a byproduct of mmWave channel tracking. Considering a vehicle under tracking communicating with a roadside BS, the approach starts with a low-complexity channel tracking algorithm, F-MOMP, to enable high-resolution channel estimates. To address the challenge of unknown vehicle orientations in realistic driving scenarios, which affect localization accuracy, we propose an attention-based network, VO-ChAT, to predict the current vehicle orientation with the input sequence of channel estimates. Following orientation compensation, the estimated channel paths are weighted for single-shot localization through a geometric transformation. Subsequently, we leverage historical channel and position information to enhance position estimation accuracy using a Transformer-inspired network, VP-ChAT, which shares a partial architecture with VO-ChAT in its encoder for processing the channel estimate sequence, while incorporating position estimates through its decoder to output the correction for the current single-shot position estimate. Our contributions are as follows:

- We consider a mmWave MIMO communication system employing URAs between a

single BS and a vehicle with the driver following realistic driving behavior models. The system model accounts for unknown clock offset drifts between the TX and RX and the system filtering effects. While conventional channel estimation algorithms fail due to the computational complexity associated with high-resolution channel estimation required for localization, we develop the F-MOMP algorithm (available at [144]) for low-complexity and accurate channel tracking –with the delay accuracy of 0.1 ns and angular accuracy of 2° (at the 80th percentile)– to enable vehicle localization through the estimated MPCs.

- To address the unknown vehicle orientation incorporated in the estimated channel angular parameters that will affect the localization process, we design an attention-based network, VO-ChAT, to track the vehicle orientations. The network processes the input sequence of channel estimates to acquire the channel spatial and temporal evolution features, and concurrently integrates the historical orientation information to predict the current orientation. It achieves the orientation prediction error of $\leq 0.5^\circ$ for 80% of the situations.

- After orientation compensation based on VO-ChAT predictions, we identify LOS and first-order reflections referring to the vehicle's height obtained during the initial access phase, using which we implement the single-shot localization through channel path geometric transformations using a WLS algorithm. Subsequently, we propose a Transformer-inspired network, VP-ChAT, to process channel and position estimate sequences. Specifically, a module structurally similar to VO-ChAT serves as the encoder for VP-ChAT to process the channel estimate sequence and extract channel spatial and temporal evolution features, while the decoder of VP-ChAT processes the current single-shot position estimate plus the position history as a query to determine the correction of the current position estimate. This approach achieves position tracking accuracy of 0.15 m at the 50th percentile and 0.36 m at the 95th percentile, outperforming SOTA methods which report 95th-percentile accuracies of 3.1 m (standard EKF baseline), 1.2 m (learning-incorporated UKF [134]), and 2.1 m (NLOS with LSTM [135]).

- Our methods are evaluated using realistic ray-tracing simulated channels, generated based on snapshots captured along the vehicle's trajectory. The simulated channel database is open sourced [145] to provide the research community with a resource for evaluating new solutions to the joint channel and PO estimation/tracking problem using vehicular communication channels.

- We further extend the single-vehicle framework to a multi-vehicle collaborative positioning scenario (Sec. 3.5), where vehicles in NLOS conditions are localized as *targets* via V2V sidelinks, using vehicles with a LOS to the BS as *anchors*. A model-based geometric framework, building on ESPRIT-D channel estimation, explicitly estimates per-vehicle orientation offsets and clock offsets from V2V link parameters. Combined with a KF for temporal smoothing, this approach achieves an average 2D localization error of 0.3 m for NLOS targets, with sub-meter accuracy for 90% of cases.

The framework described in this chapter is built upon the foundational design presented in our prior work [146] with several key improvements. Vehicle trajectories are generated based on realistic driving behaviors, accounting for dynamic orientation changes which are tracked through the newly added VO-ChAT. The original channel tracking strategy based on MOMP is extended to the F-MOMP-based solution which incorporates the factoring operation to reduce computational complexity. The original single-shot localization through geometric transformations is refined as solving a WLS estimation problem. The original V-ChAT network is tuned into VP-ChAT to accommodate updated vehicle trajectories for corrections of the single-shot position estimates. A larger and more comprehensive dataset is formed, and additional numerical experiments and comparisons with SOTA studies are included.

The rest of this chapter is organized as follows: Sec. 3.2 outlines the general V2I communication setup, the driving behavior model, and the communication system model. Sec. 3.3 details the stages of the proposed hybrid model/data-driven approach, including channel tracking, orientation prediction and compensation, single-shot localization, and position

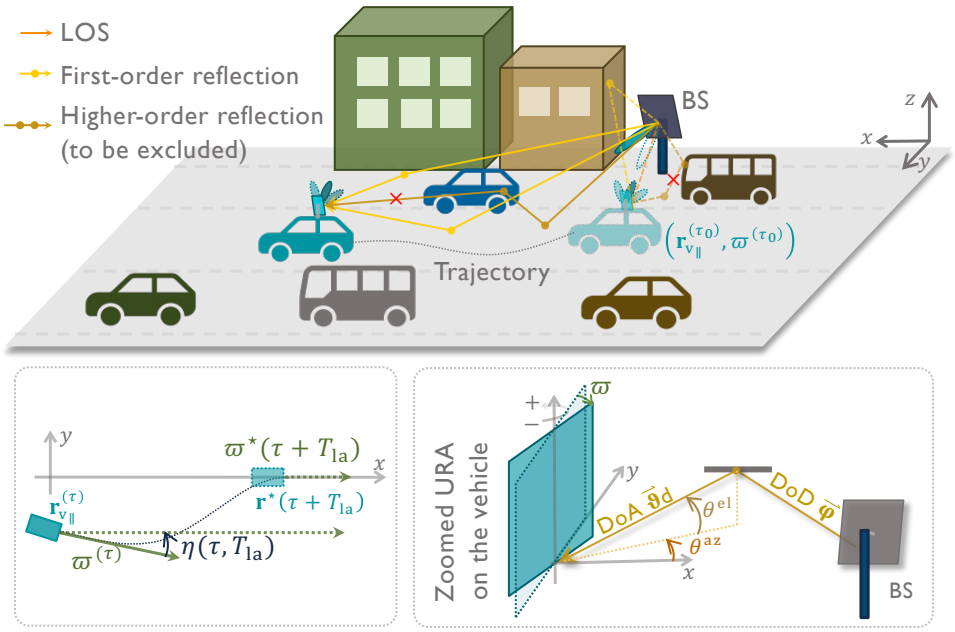


Figure 3.1. System model for tracking a vehicle in the urban canyon environment.

corrections leveraging historical channel and position estimates. Sec. 3.4 presents numerical results evaluating the proposed strategy and comparisons with prior work. Sec. 3.5 extends the framework to multi-vehicle collaborative positioning via V2V sidelinks for localization of NLOS vehicles. Finally, Sec. 3.6 concludes the chapter by summarizing the key findings.

Notations: $[\mathbf{x}]_i$ and $[\mathbf{X}]_{i,j}$ denote the i-th entry of a vector $\mathbf{x}$ and the entry at i-th row and j-th column of a matrix $\mathbf{X}$ (the same rule applies for a tensor). $\mathbf{X}^{\mathsf{T}}$, $\bar{\mathbf{X}}$, $\mathbf{X}^*$, and $\mathbf{X}^\dagger$ are the transpose, conjugate, conjugate transpose, and pseudo inverse of $\mathbf{X}$. $[\mathbf{X},\mathbf{Y}]$ and $[\mathbf{X};\mathbf{Y}]$ are the horizontal and vertical concatenation of $\mathbf{X}$ and $\mathbf{Y}$. $\mathbf{X}\otimes\mathbf{Y}$ and $\mathbf{X}\odot\mathbf{Y}$ are the Kronecker product and Khatri-Rao product of $\mathbf{X}$ and $\mathbf{Y}$. $\mathfrak{X}\cup\mathfrak{Y}$ is the union set of set $\mathfrak{X}$ and $\mathfrak{Y}$. $x\sim\mathcal{N}(\dot{x},\sigma_x^2)$ denotes the variable x follows the Gaussian distribution with mean $\dot{x}$ and variance σ_x^2. All variables marked with a circumflex $\hat{x}$, $\hat{\mathbf{x}}$, and $\hat{\mathbf{X}}$ denote estimates of the corresponding quantities. For clarity, we provide a summary of notations in Table 3.2.

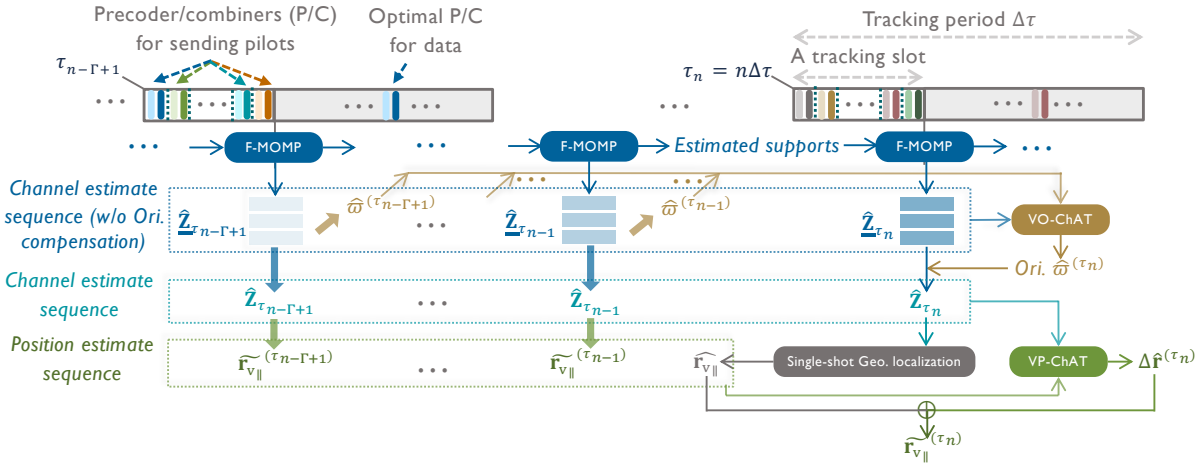


Figure 3.2. System diagram consisting of F-MOMP for channel tracking, VO-ChAT for vehicle orientation tracking, single-shot geometric (Geo.) localization using angles after orientation compensation, and VP-ChAT for vehicle position tracking.

3.2 System Model

We consider a mmWave vehicular communication system where an active car under tracking moves at the fast lane with the heading (orientation) changing according to driver behavior models [147]. The active vehicle starts from position $\mathbf{r}_{\mathrm{v}_{\mathrm{II}}}^{(\tau_0)}$, a two-dimensional vector comprising x and y coordinates, with the initial orientation $\varpi^{(\tau_0)}$. At time τ when the vehicle is at position $\mathbf{r}_{\mathrm{v}_{\mathrm{II}}}^{(\tau)}$ with the orientation $\varpi^{(\tau)}$ and the moving speed $v_{\mathrm{v}}^{(\tau)}$, the driver checks the aiming point $\mathbf{r}^{\star}(\tau+T_{\mathrm{la}})$ lying on the lane center line, where T_{la} is the looking ahead time depending on the environmental visibility, and steers the wheel continuously during the period from τ to $\tau+T_{\mathrm{la}}$ expecting to arrive at $\mathbf{r}^{\star}(\tau+T_{\mathrm{la}})$ with the orientation $\varpi^{\star}(\tau+T_{\mathrm{la}})$. Hence, the bearing angle from τ to $\tau+T_{\mathrm{la}}$, denoted as $\eta(\tau,T_{\mathrm{la}})$, is calculated by $\eta(\tau,T_{\mathrm{la}})=\varpi^{\star}(\tau+T_{\mathrm{la}})-\varpi^{(\tau)}$. We define the driver compensate control as $\delta(\tau)$, which needs to be increased meaning the driver controls the steering wheel with higher strength when $\eta(\tau,T_{\mathrm{la}})$ is large. Considering a continuous operation model, the following equations hold:

$$\eta(\tau,T_{\mathrm{la}})=\int_{\tau}^{\tau+T_{\mathrm{la}}}\omega(t)\delta(t)\partial t; \tag{3.1}$$

$$\mathbf{r}^{\star}(\tau+T_{\mathrm{la}})-\mathbf{r}_{\mathrm{v}_{\mathrm{II}}}^{(\tau)}=\int_{\tau}^{\tau+T_{\mathrm{la}}}[v_{\mathrm{v}}^{(\tau)}\cos(\varpi^{(t)}),v_{\mathrm{v}}^{(\tau)}\sin(\varpi^{(t)})]^{\mathsf{T}}\partial t, \tag{3.2}$$

where $\omega(t)\sim\mathcal{N}(\dot{\omega},\sigma_{\omega}^{2})$ is the wheel steering rate at t, (3.1) represents the cumulated orientation changes, and (3.2) represents the vehicle position changes. When implemented in the discrete domain, let $\Delta\tau=\tau_{n+1}-\tau_n$ be the sampling interval, vehicle's orientation and position can be

Table 3.2. List of notations.

Symbol	Description	Symbol	Description	Symbol	Description
VEHICLE DYNAMICS AND DRIVER MODEL					
$\mathbf{r}_{\mathrm{v}_{\shortparallel}}^{(\tau)}$	Vehicle position at time (x, y coordinates) τ	$\varpi^{(\tau)}$	Vehicle orientation at time τ	$v_{\mathrm{v}}^{(\tau)}$	Vehicle speed at time τ
$\mathbf{r}^{\star}(\tau+T_{\mathrm{la}})$	Aiming point on lane center	$\varpi^{\star}(\tau+T_{\mathrm{la}})$	Target orientation at aiming point	T_{la}	Look-ahead time for driver model
$\eta(\tau,T_{\mathrm{la}})$	Bearing angle over look-ahead interval	$\delta(\tau)$	Driver compensation control	$\omega(t)$	Wheel steering rate at time t
$\dot{\omega}$	Mean wheel steering rate	σ_{ω}^2	Wheel steering rate variance	K_{e}	Driver gain constant
T_{e}	Driver leading time constant	$n_{\delta}(t)$	Driver control noise	σ_{δ}^2	Driver control noise variance
COMMUNICATION SYSTEM					
τ	Time index	τ_n	Discrete time index	τ_{n+1}	Next discrete time index ($\tau_n+\Delta\tau$)
N_{t}	BS transmit antenna array size	$N_{\mathrm{t}}^{\mathrm{x}}$	Number of TX antennas (horizontal)	N_{t}^{y}	Number of TX antennas (vertical)
N_{r}	Vehicle receive antenna array size	$N_{\mathrm{r}}^{\mathrm{x}}$	Number of RX antennas (horizontal)	N_{r}^{y}	Number of RX antennas (vertical)
$N_{\mathrm{t}}^{\mathrm{rf}}$	Number of BS RF chains	$N_{\mathrm{r}}^{\mathrm{rf}}$	Number of vehicle RF chains	L	Number of multipath components
d	Channel tap index	T_{s}	Sampling interval		
$\mathbf{H}_{d}^{(\tau_n)}$	Channel matrix at tap d	$\alpha_{\ell}^{(\tau_n)}$	Complex gain of path ℓ	$t_{\ell}^{(\tau_n)}$	Time of arrival of path ℓ
$t_{\mathrm{off}}^{(\tau_n)}$	Clock offset between TX and RX	$f_{\mathrm{p}}(\cdot)$	Filtering function	$\theta_{\ell}^{\mathrm{az}(\tau_n)}$	Azimuth AoA
$\theta_{\ell}^{\mathrm{el}(\tau_n)}$	Elevation AoA	$\phi_{\ell}^{\mathrm{az}(\tau_n)}$	Azimuth AoD	$\phi_{\ell}^{\mathrm{el}(\tau_n)}$	Elevation AoD
$\mathbf{a}_{\mathrm{r}}(\cdot)$	Receive array response vector	$\mathbf{a}_{\mathrm{t}}(\cdot)$	Transmit array response vector	$\mathbf{a}(\cdot)$	Steering vector
$\theta^{\shortparallel}$	Parallel component of AoA	$\theta^{\perp}$	Perpendicular component of AoA	$\phi^{\shortparallel}$	Parallel component of AoD
$\phi^{\perp}$	Perpendicular component of AoD	$[\mathbf{a}(\vartheta)]_n$	n-th element of steering vector		
N_{s}	Number of data streams	Q	Pilot sequence length	$\mathbf{s}[q]$	q-th transmitted pilot instance
$\mathbf{F}$	Hybrid precoder matrix	$\mathbf{F}^{\mathrm{rf}}$	Analog precoder	$\mathbf{F}^{\mathrm{bb}}$	Digital precoder
$\mathbf{W}$	Hybrid combiner matrix	$\mathbf{W}^{\mathrm{rf}}$	Analog combiner	$\mathbf{W}^{\mathrm{bb}}$	Digital combiner
M	Number of precoder/combiner pairs	$\mathbf{F}_m$	m-th precoder matrix	$\mathbf{W}_m$	m-th combiner matrix
$\mathbf{y}_m[q]$	Received signal at m-th pair	P_{t}	Transmitted power	N_{d}	Number of channel taps
$\mathbf{n}_m[q]$	Additive white Gaussian noise	$\sigma_{\mathbf{n}}^2$	Noise variance	K_{B}	Boltzmann constant
T_{F}	Environmental temperature (°F)	B_{c}	Communication system bandwidth	$\check{\mathbf{y}}_m[q]$	Whitened received signal
$\mathbf{L}_m$	Cholesky decomposition matrix	$\check{\mathbf{W}}_m^{*}$	Whitened combiner	$\check{\mathbf{n}}_m[q]$	Whitened noise
$\check{\mathbf{Y}}_m$	Whitened measurement matrix	$\check{\mathbf{N}}_m$	Whitened noise matrix	$\mathbf{S}$	Pilot signal matrix
F-MOMP CHANNEL TRACKING ALGORITHM					
$\boldsymbol{\Upsilon}_m$	Measurement matrix	$\boldsymbol{\Psi}$	Dictionary matrix	$\mathbf{x}$	Sparse vector
$\mathbf{A}_{\mathrm{d}}$	Delay dictionary	$\mathbf{A}_{\mathrm{t}}$	Transmit array dictionary	$\mathbf{A}_{\mathrm{r}}$	Receive array dictionary
$\mathbf{A}_{\mathrm{t}}^{\shortparallel}$, $\mathbf{A}_{\mathrm{t}}^{\perp}$	TX azimuth and elevation dictionary	$\mathbf{A}_{\mathrm{r}}^{\shortparallel}$, $\mathbf{A}_{\mathrm{r}}^{\perp}$	RX azimuth and elevation dictionary	$\mathbf{p}(t)$	Sampled filtering function
N_k^{a}	k-th dictionary grid size	N_k^{s}	k-th system dimension size	$\ddot{\imath}_{j_1}$	Delay grid values
$\ddot{\phi}_{j_2}^{\shortparallel}$, $\ddot{\phi}_{j_3}^{\perp}$	Azimuth and elevation AoD grids	$\ddot{\theta}_{j_4}^{\shortparallel}$, $\ddot{\theta}_{j_5}^{\perp}$	Azimuth and elevation AoA grids		
N_{est}	Number of estimated components	$\mathfrak{x}$	Support set of sparse vector	ℓ_{s}	Support index
$\mathbf{i}$	Multidimensional measurement tensor index	$\mathbf{j}$	Multidimensional dictionary tensor index	$\mathfrak{I}$	Measurement tensor index set
$\mathfrak{J}$	Dictionary tensor index set	$\boldsymbol{\Phi}_m$	Measurement tensor	$\mathbf{X}$	Sparse tensor
$\boldsymbol{\Psi}_k$	k-th dimension full dictionary	N_{iter}	Number of iterations	$f_{\mathbf{i}}$, $f_{\mathbf{j}}$	Index mapping function
$\zeta_q^{\mathrm{S}}(j_1)$	Signal factor	$\zeta_m^{\mathrm{F}}(j_2,j_3)$	Precoder factor	$\zeta_m^{\mathrm{W}}(j_4,j_5)$	Combiner factor
$\hat{\mathfrak{J}}_{\mathrm{sup}}^{(\tau_{n-1})}$	Previous estimated support set	n_{est}	Estimated path index	$\boldsymbol{\Psi}_{k,n_{\mathrm{est}}}^{(\tau_n)}$	Reduced dictionary with grid spanning g_k
$\mathfrak{J}_{k,n_{\mathrm{est}}}$	Local grid search space	$\check{\mathbf{Y}}_m^{\mathrm{res}}$	Residual matrix	$\hat{\jmath}_{k,n_{\mathrm{est}}}^{(\tau_n)}$	Estimated k-th dictionary grid index
$\underline{\hat{\boldsymbol{Z}}}_{\tau_n}$	Estimated channel containing N_{est} paths	$\hat{\boldsymbol{\alpha}}_{\tau_n}$	Estimated gain vector	$\underline{\hat{\mathbf{t}}}_{\tau_n}$	Estimated delay vector (w/o $t_{\mathrm{off}}^{(\tau_n)}$ compensation)
$\underline{\hat{\boldsymbol{\theta}}}_{\tau_n}^{\mathrm{az}}$	Estimated azimuth AoA vector (w/o $\varpi^{(\tau_n)}$ compensation)	$\hat{\boldsymbol{\theta}}_{\tau_n}^{\mathrm{el}}$	Estimated elevation AoA vector	$\hat{\boldsymbol{\phi}}_{\tau_n}^{\mathrm{az}}$	Estimated azimuth AoD vector
$\hat{\boldsymbol{\phi}}_{\tau_n}^{\mathrm{el}}$	Estimated elevation AoD vector				
GEOMETRIC LOCALIZATION					
$\hat{\boldsymbol{Z}}_{\tau_n}$	Compensated channel estimate	$\hat{\boldsymbol{\theta}}_{\tau_n}^{\mathrm{az}}$	Compensated azimuth AoA	$\hat{\mathbf{t}}_{\tau_n}$	Compensated ToA estimates
$\mathbf{r}_{\mathrm{v}}^{(\tau_n)}$	Vehicle 3D position	$\mathbf{r}_{\mathrm{B}}$	Known BS array 3D position	$\overrightarrow{\varphi}_{\ell}$	Channel component DoD
$\overrightarrow{\vartheta}_{\ell}$	Channel component DoA	$d_{\ell}^{\vartheta(\tau_n)}$	Vehicle to scatterer distance	$d_{\ell}^{\varphi(\tau_n)}$	Scatterer to BS distance
v_{c}	Speed of light	h_{v}	Vehicle array height	$\lvert\alpha_{\mathrm{th}}\rvert$	Path gain threshold for path selections
$\mathfrak{L}$	Selected path set	w_{ℓ}	Path weight for WLS algorithms	$\mathbf{B}_{\ell}$	WLS coefficient matrix
$\mathbf{b}_{\ell}$	WLS observation vector	$\hat{\mathbf{r}}_{\mathrm{v}_{\shortparallel}}^{(\tau_n)}$	Single-shot 2D position estimate		
VO-ChAT/VP-ChAT TRACKING SYSTEM AND ATTENTION MECHANISM					
Γ	History length for tracking	$\hat{\boldsymbol{\varpi}}_{\tau_n\lvert\Gamma}$	Orientation sequence acquired at τ_n with length Γ	$\hat{\mathbf{r}}_{\tau_n\lvert\Gamma}$	Position sequence acquired at τ_n with length Γ
$\mathbf{W}^{\mathrm{o}}$	VO-ChAT learnable parameters	$\mathbf{W}^{\mathrm{p}}$	VP-ChAT learnable parameters	N_{h}	Number of attention heads
$\mathbf{V}_{n_{\mathrm{h}}}$	Value representation of head n_{h}	$\mathbf{K}_{n_{\mathrm{h}}}$	Key representation of head n_{h}	$\mathbf{Q}_{n_{\mathrm{h}}}$	Query representation of head n_{h}
V_{dim}	Value embedding dimension	K_{dim}	Key embedding dimension	Q_{dim}	Query embedding dimension
$\mathbf{R}_{n_{\mathrm{h}}}$	Attention output	$\mathbf{R}$	Spatial feature output	$\mathbf{V}'_{n_{\mathrm{h}}}$	Temporal value representation
$\mathbf{K}'_{n_{\mathrm{h}}}$	Temporal key representation	$\mathbf{Q}'_{n_{\mathrm{h}}}$	Temporal query representation	$\mathbf{R}'_{n_{\mathrm{h}}}$	Temporal attention output
$\Delta\hat{\mathbf{r}}^{(\tau_n)}$	2D position correction vector	$\widetilde{\mathbf{r}}_{\mathrm{v}_{\shortparallel}}^{(\tau_n)}$	Corrected vehicle position		
$\mathcal{S}_{\mathrm{tr}}$	Training trajectory set	$\mathcal{S}_{\mathrm{te}}$	Testing trajectory set		

updated as:

$$\delta(\tau_{n+1}) = K_{\mathrm{e}}\left(\delta(\tau_n) + T_{\mathrm{e}}\frac{\eta(\tau_{n+1}, T_{\mathrm{la}}{-}\Delta\tau) - \eta(\tau_n, T_{\mathrm{la}})}{\Delta\tau}\right) + n_\delta(\tau_{n+1}); \tag{3.3}$$

$$\varpi^{(\tau_{n+1})} = \varpi^{(\tau_n)} + \omega(\tau_n)\delta(\tau_n)\Delta\tau; \tag{3.4}$$

$$\mathbf{r}_{\mathrm{V}_{\mathrm{II}}}^{(\tau_{n+1})} = \mathbf{r}_{\mathrm{V}_{\mathrm{II}}}^{(\tau_n)} + \Delta\tau[v_{\mathrm{v}}^{(\tau_n)}\cos(\varpi^{(\tau_n)}), v_{\mathrm{v}}^{(\tau_n)}\sin(\varpi^{(\tau_n)})]^{\mathsf{T}}, \tag{3.5}$$

where K_{e} and T_{e} are the driver gain and leading time constants, $n_\delta(t) \sim \mathcal{N}(0, \sigma_\delta^2)$ is the driver control noise being σ_δ^2 the distribution variance.

Downlink communication is performed between the active car and a single BS at the roadside for tracking the channel and POs. The BS is equipped with a URA of size $N_{\mathrm{t}} = N_{\mathrm{t}}^{\mathrm{x}} \times N_{\mathrm{t}}^{\mathrm{y}}$ facing the road, and the vehicle has 4 smaller URAs placed vertically on the hardtop as in [120,121], each of which has a size of $N_{\mathrm{r}} = N_{\mathrm{r}}^{\mathrm{x}} \times N_{\mathrm{r}}^{\mathrm{y}}$. A hybrid MIMO communication architecture is adopted, with $N_{\mathrm{t}}^{\mathrm{rf}}$ and $N_{\mathrm{r}}^{\mathrm{rf}}$ RF chains deployed at the TX and RX. Hereby, the frequency selective mmWave channel containing L MPCs at a given time τ_n can be defined as

$$\mathbf{H}_d^{(\tau_n)} = \sum_{\ell=1}^{L}\left(\alpha_\ell^{(\tau_n)} f_{\mathrm{p}}\left(dT_{\mathrm{s}} - \left(t_\ell^{(\tau_n)} - t_{\mathrm{off}}^{(\tau_n)}\right)\right) \cdot \mathbf{a}_{\mathrm{r}}\left(\theta_\ell^{\mathrm{az}(\tau_n)} - \varpi^{(\tau_n)}, \theta_\ell^{\mathrm{el}(\tau_n)}\right)\mathbf{a}_{\mathrm{t}}\left(\phi_\ell^{\mathrm{az}(\tau_n)}, \phi_\ell^{\mathrm{el}(\tau_n)}\right)^*\right), \tag{3.6}$$

where d is the channel tap index, T_{s} is the sampling interval, $t_{\mathrm{off}}^{(\tau_n)}$ is the unknown clock offset between the TX and RX, $f_{\mathrm{p}}(\cdot)$ is the filtering function that factors in filtering effects in the system, $\alpha_\ell^{(\tau_n)}$ and $t_\ell^{(\tau_n)}$ are the complex gain and the ToA of the ℓ-th path, $\mathbf{a}_{\mathrm{r}}\left(\theta_\ell^{\mathrm{az}(\tau_n)} - \varpi^{(\tau_n)}, \theta_\ell^{\mathrm{el}(\tau_n)}\right)$ represents the RX array response evaluated at the azimuth and elevation AoA, denoted as $\theta_\ell^{\mathrm{az}(\tau_n)} - \varpi^{(\tau_n)}$ and $\theta_\ell^{\mathrm{el}(\tau_n)}$, and $\mathbf{a}_{\mathrm{t}}\left(\phi_\ell^{\mathrm{az}(\tau_n)}, \phi_\ell^{\mathrm{el}(\tau_n)}\right)$ is the TX array response evaluated at the azimuth and elevation AoD, denoted as $\phi_\ell^{\mathrm{az}(\tau_n)}$ and $\phi_\ell^{\mathrm{el}(\tau_n)}$. Note that, $\theta_\ell^{\mathrm{az}(\tau_n)}$ and $\theta_\ell^{\mathrm{el}(\tau_n)}$ are azimuth and elevation AoAs in the global coordinate system, and the same applies to azimuth

and elevation AoDs $\phi_\ell^{\mathrm{az}(\tau_n)}$ and $\phi_\ell^{\mathrm{el}(\tau_n)}$. The array responses can be formulated in the Kronecker product form as

$$\begin{cases} \mathbf{a}_{\mathrm{r}}(\theta^{\mathrm{az}}-\varpi,\theta^{\mathrm{el}}) = \mathbf{a}(\theta^{\shortparallel},\theta^{\perp}) = \mathbf{a}(\theta^{\shortparallel})\otimes\mathbf{a}(\theta^{\perp}) \\ \mathbf{a}_{\mathrm{t}}(\phi^{\mathrm{az}},\phi^{\mathrm{el}}) = \mathbf{a}(\phi^{\shortparallel},\phi^{\perp}) = \mathbf{a}(\phi^{\shortparallel})\otimes\mathbf{a}(\phi^{\perp}) \end{cases}, \tag{3.7}$$

where $\theta^{\shortparallel} = \cos(\theta^{\mathrm{el}})\sin(\theta^{\mathrm{az}}-\varpi)$, $\theta^{\perp} = \sin(\theta^{\mathrm{el}})$, $\phi^{\shortparallel} = \cos(\phi^{\mathrm{el}})\sin(\phi^{\mathrm{az}})$, $\phi^{\perp} = \sin(\phi^{\mathrm{el}})$, and $\mathbf{a}(\cdot)$ is the steering vector where $[\mathbf{a}(\vartheta)]_n = e^{-j\pi(n-1)\vartheta}$ considering a half-wavelength element spacing for the planar arrays.

Pilots in the form of $N_{\mathrm{s}} \leq \min\{N_{\mathrm{t}}^{\mathrm{rf}}, N_{\mathrm{r}}^{\mathrm{rf}}\}$ data streams of length Q are transmitted for channel tracking at each τ_n (we omit the upper right "(τ_n)" for simplicity for the following notation definition and notations), where the q-th instance is denoted as $\mathbf{s}[q] \in \mathbb{C}^{N_{\mathrm{s}}\times 1}$ with $\mathbb{E}[\mathbf{s}[q]\mathbf{s}[q]^*] = \frac{1}{N_{\mathrm{s}}}\mathbf{I}_{N_{\mathrm{s}}}$. Hybrid precoder and combiner are employed, which are denoted as $\mathbf{F} = \mathbf{F}^{\mathrm{rf}}\mathbf{F}^{\mathrm{bb}} \in \mathbb{C}^{N_{\mathrm{t}}\times N_{\mathrm{s}}}$ and $\mathbf{W} = \mathbf{W}^{\mathrm{rf}}\mathbf{W}^{\mathrm{bb}} \in \mathbb{C}^{N_{\mathrm{r}}\times N_{\mathrm{s}}}$, where $\mathbf{F}^{\mathrm{rf}}$ and $\mathbf{F}^{\mathrm{bb}}$ are the analog and digital precoders, and $\mathbf{W}^{\mathrm{rf}}$ and $\mathbf{W}^{\mathrm{bb}}$ are the analog and digital combiners. Within a channel tracking interval whose duration is less than the channel coherence time, M precoder and combiner pairs are employed, denoted as $\mathbf{F}_m$ and $\mathbf{W}_m$, $m = 1,2,...,M$, for the m-th pair. Accordingly, the q-th instance of the received signal using $\mathbf{F}_m$ and $\mathbf{W}_m$ is given as

$$\mathbf{y}_m[q] = \mathbf{W}_m^* \sum_{d=0}^{N_{\mathrm{d}}-1} \sqrt{P_{\mathrm{t}}}\mathbf{H}_d\mathbf{F}_m\mathbf{s}[q-d] + \mathbf{W}_m^*\mathbf{n}_m[q], \tag{3.8}$$

where P_{t} is the transmitted power, N_{d} is the number of channel taps, and $\mathbf{n}_m[q] \sim \mathcal{N}(\mathbf{0}, \frac{\sigma_{\mathbf{n}}^2}{N_{\mathrm{r}}}\mathbf{I}_{N_{\mathrm{r}}})$ is modeled as additive white Gaussian noise (AWGN) where $\sigma_{\mathbf{n}}^2 = K_{\mathrm{B}}T_{\mathrm{F}}B_{\mathrm{c}}$, being K_{B} the Boltzmann constant and T_{F} the environmental temperature in Fahrenheit. Due to the noise being combined with $\mathbf{W}_m^*$, it is no longer white. Therefore, we whiten the received signal as $\breve{\mathbf{y}}_m[q] = \mathbf{L}_m^{-1}\mathbf{y}_m[q]$, where $\mathbf{L}_m$ is computed via Cholesky decomposition of $\mathbf{W}_m^*\mathbf{W}_m = \mathbf{L}_m\mathbf{L}_m^*$, so that $\mathbb{E}\left[\mathbf{L}_m^{-1}\mathbf{W}_m^*\mathbf{n}_m[q](\mathbf{L}_m^{-1}\mathbf{W}_m^*\mathbf{n}_m[q])^*\right] = \sigma_{\mathbf{n}}^2\mathbf{I}_{N_{\mathrm{s}}}$. Let $\breve{\mathbf{W}}_m^* = \mathbf{L}_m^{-1}\mathbf{W}_m^*$ and $\breve{\mathbf{n}}_m[q] = \mathbf{L}_m^{-1}\mathbf{W}_m^*\mathbf{n}_m$ for

$$\boldsymbol{\Upsilon}_m = \sqrt{P_{\mathrm{t}}} \begin{bmatrix} \mathbf{s}[1]^{\mathsf{T}}[\mathbf{F}_m^{\mathsf{T}}]_{:,1}\mathbf{W}_m^* & \dots & \mathbf{s}[1]^{\mathsf{T}}[\mathbf{F}_m^{\mathsf{T}}]_{:,N_{\mathrm{t}}}\mathbf{W}_m^* & \dots & \mathbf{0}^{\mathsf{T}}[\mathbf{F}^{\mathsf{T}}]_{:,1}\mathbf{W}_m^* & \dots & \mathbf{0}^{\mathsf{T}}[\mathbf{F}_m^{\mathsf{T}}]_{:,N_{\mathrm{t}}}\mathbf{W}_m^* \\ \vdots & \vdots & \vdots & \dots & \vdots & \vdots & \vdots \\ \mathbf{s}[Q]^{\mathsf{T}}[\mathbf{F}_m^{\mathsf{T}}]_{:,1}\mathbf{W}_m^* & \dots & \mathbf{s}[Q]^{\mathsf{T}}[\mathbf{F}_m^{\mathsf{T}}]_{:,N_{\mathrm{t}}}\mathbf{W}_m^* & \dots & \mathbf{s}[Q-(N_{\mathrm{d}}-1)]^{\mathsf{T}}[\mathbf{F}_m^{\mathsf{T}}]_{:,1}\mathbf{W}_m^* & \dots & \mathbf{s}[Q-(N_{\mathrm{d}}-1)]^{\mathsf{T}}[\mathbf{F}_m^{\mathsf{T}}]_{:,N_{\mathrm{t}}}\mathbf{W}_m^* \end{bmatrix}. \tag{3.10}$$

$$[\boldsymbol{\Psi}]_{:,f_{\mathbf{j}}} = \Big[\big[\mathbf{p}(\ddot{\imath}_{j_1})\big]_1 \Big[\bar{\mathbf{a}}(\ddot{\phi}_{j_2}^{\shortparallel}, \ddot{\phi}_{j_3}^{\perp})\Big]_1 \mathbf{a}(\ddot{\theta}_{j_4}^{\shortparallel}, \ddot{\theta}_{j_5}^{\perp}); \dots; \big[\mathbf{p}(\ddot{\imath}_{j_1})\big]_1 \Big[\bar{\mathbf{a}}(\ddot{\phi}_{j_2}^{\shortparallel}, \ddot{\phi}_{j_3}^{\perp})\Big]_{N_{\mathrm{t}}} \mathbf{a}(\ddot{\theta}_{j_4}^{\shortparallel}, \ddot{\theta}_{j_5}^{\perp}); \dots;$$
$$\big[\mathbf{p}(\ddot{\imath}_{j_1})\big]_{N_{\mathrm{d}}} \Big[\bar{\mathbf{a}}(\ddot{\phi}_{j_2}^{\shortparallel}, \ddot{\phi}_{j_3}^{\perp})\Big]_1 \mathbf{a}(\ddot{\theta}_{j_4}^{\shortparallel}, \ddot{\theta}_{j_5}^{\perp}); \dots; \big[\mathbf{p}(\ddot{\imath}_{j_1})\big]_{N_{\mathrm{d}}} \Big[\bar{\mathbf{a}}(\ddot{\phi}_{j_2}^{\shortparallel}, \ddot{\phi}_{j_3}^{\perp})\Big]_{N_{\mathrm{t}}} \mathbf{a}(\ddot{\theta}_{j_4}^{\shortparallel}, \ddot{\theta}_{j_5}^{\perp}) \Big]. \tag{3.11}$$

simplicity, the whitened collected measurements can be written as

$$\check{\mathbf{Y}}_m = \check{\mathbf{W}}_m^* [\mathbf{H}_0, ..., \mathbf{H}_{N_{\mathrm{d}}-1}] \left((\mathbf{I}_{N_{\mathrm{d}}} \otimes \mathbf{F}_m) \sqrt{P_{\mathrm{t}}} \mathbf{S} \right) + \check{\mathbf{N}}_m, \tag{3.9}$$

where $[\check{\mathbf{Y}}_m]_{:,q} = \check{\mathbf{y}}_m[q]$, $[\check{\mathbf{N}}_m]_{:,q} = \check{\mathbf{n}}_m[q]$, and $[\mathbf{S}]_{:,q} = [\mathbf{s}[q]; \mathbf{s}[q-1]; ...; \mathbf{s}[q-(N_{\mathrm{d}}-1)]]$.

3.3 Channel and Vehicle PO tracking system

This section provides detailed algorithms for joint channel and vehicle PO tracking. The system diagram is shown in Fig. 3.2. We first introduce the F-MOMP algorithm, which accelerates the calculation of the product for the measurement and dictionary matrix and enhances computational efficiency compared to conventional OMP and MOMP algorithms, to realize mmWave channel tracking. Then VO-ChAT employing attention mechanisms predicts the current vehicle orientation based on the channel estimate sequence and orientation history. Following orientation compensation using the predicted orientation, the single-shot position estimate obtained by solving a WLS problem is treated as the initial position estimate to be refined by VP-ChAT. Ultimately, VP-ChAT, the network inspired by the Transformer architecture, leverages historical channel estimate and position estimate sequences to provide the correction of the current single-shot position estimate, realizing precise vehicle position tracking.

$$[\boldsymbol{\Upsilon}_m]_{(q-1)N_\mathrm{s}+1:qN_\mathrm{s},:}[\boldsymbol{\Psi}]_{:,f_\mathbf{j}} = \sqrt{P_\mathrm{t}}\left(\sum_{n_\mathrm{d}=1}^{N_\mathrm{d}}[\mathbf{p}(\ddot{\imath}_{j_1})]_{n_\mathrm{d}}\mathbf{s}[q-(n_\mathrm{d}-1)]^\mathsf{T}\right)\left(\sum_{n_\mathrm{t}=1}^{N_\mathrm{t}}[\mathbf{F}_m^\mathsf{T}]_{:,n_\mathrm{t}}\left[\bar{\mathbf{a}}(\ddot{\phi}_{j_2}^{\shortparallel},\ddot{\phi}_{j_3}^{\perp})\right]_{n_\mathrm{t}}\right)\left(\sum_{n_\mathrm{r}=1}^{N_\mathrm{r}}[\mathbf{W}_m^*]_{:,n_\mathrm{r}}\left[\mathbf{a}(\ddot{\theta}_{j_4}^{\shortparallel},\ddot{\theta}_{j_5}^{\perp})\right]_{n_\mathrm{r}}\right)$$

$$= \left[\sqrt{P_\mathrm{t}}[\mathbf{s}[q],\mathbf{s}[q-1],\ldots,\mathbf{0}]\,\mathbf{p}(\ddot{\imath}_{j_1})\right]^\mathsf{T}\left[\mathbf{F}_m^\mathsf{T}\bar{\mathbf{a}}(\ddot{\phi}_{j_2}^{\shortparallel},\ddot{\phi}_{j_3}^{\perp})\right]\left[\mathbf{W}_m^*\mathbf{a}(\ddot{\theta}_{j_4}^{\shortparallel},\ddot{\theta}_{j_5}^{\perp})\right] \tag{3.12}$$

$$= [\zeta_q^\mathrm{S}(j_1)]^\mathsf{T}[\zeta_m^\mathrm{F}(j_2,j_3)][\zeta_m^\mathrm{W}(j_4,j_5)] \in \mathbb{C}^{N_\mathrm{s}\times 1}. \tag{3.13}$$

3.3.1 F-MOMP Based Channel Tracking

Before diving into the proposed F-MOMP channel tracking algorithm, we present a concise overview of the conventional OMP algorithm for channel estimation, followed by an explanation of how MOMP addresses the computational complexity issue of OMP through dimensional operations. Relevant notations are introduced throughout the discussion. Based on $\mathrm{vec}(\mathbf{AXB}) = (\mathbf{B}^\mathsf{T}\otimes\mathbf{A})\mathrm{vec}(\mathbf{X})$, (3.9) can be written in the form

$$\mathrm{vec}(\breve{\mathbf{Y}}_m) = \boldsymbol{\Upsilon}_m\mathrm{vec}([\mathbf{H}_0,\ldots,\mathbf{H}_{N_\mathrm{d}-1}]) + \mathrm{vec}(\breve{\mathbf{N}}_m),$$

where $\boldsymbol{\Upsilon}_m = ((\mathbf{I}_{N_\mathrm{d}}\otimes\mathbf{F}_m)\sqrt{P_\mathrm{t}}\mathbf{S})^\mathsf{T}\otimes\breve{\mathbf{W}}_m^* \in \mathbb{C}^{QN_\mathrm{s}\times N_\mathrm{d}N_\mathrm{t}N_\mathrm{r}}$ is the measurement matrix. The channel can be represented as $\mathrm{vec}([\mathbf{H}_0,\ldots,\mathbf{H}_{N_\mathrm{d}-1}]) = \boldsymbol{\Psi}\mathbf{x}$ leveraging its sparsity, where $\boldsymbol{\Psi}$ is the dictionary formulated as

$$\boldsymbol{\Psi} = \mathbf{A}_\mathrm{d}\otimes(\overline{\mathbf{A}}_\mathrm{t}\otimes\mathbf{A}_\mathrm{r}) \in \mathbb{C}^{N_\mathrm{r}N_\mathrm{t}N_\mathrm{d}\times N_\mathrm{r}^\mathrm{a}N_\mathrm{t}^\mathrm{a}N_\mathrm{d}^\mathrm{a}}, \tag{3.14}$$

where $\mathbf{A}_\mathrm{d} = \left[\mathbf{p}(\ddot{\imath}_1),\ldots,\mathbf{p}(\ddot{\imath}_{N_\mathrm{d}^\mathrm{a}})\right]$ is the dictionary for the delay evaluated on the grid values $\{\ddot{\imath}_{j_1}|\,j_1 = 1,\ldots,N_\mathrm{d}^\mathrm{a}\}$, and $\mathbf{p}(t) = [f_\mathrm{p}(0\cdot T_\mathrm{s}-t),\ldots,f_\mathrm{p}((N_\mathrm{d}-1)T_\mathrm{s}-t)]^\mathsf{T} \in \mathbb{R}^{N_\mathrm{d}\times 1}$ is a sampled version of $f_\mathrm{p}(\cdot)$ mentioned in (3.6); $\mathbf{A}_\mathrm{t} = \mathbf{A}_\mathrm{t}^{\shortparallel}\otimes\mathbf{A}_\mathrm{t}^{\perp}$ is the dictionary to evaluate azimuth and elevation AoDs with $\mathbf{A}_\mathrm{t}^{\shortparallel} = \left[\mathbf{a}(\ddot{\phi}_1^{\shortparallel}),\ldots,\mathbf{a}(\ddot{\phi}_{N_2^\mathrm{a}}^{\shortparallel})\right] \in \mathbb{C}^{N_\mathrm{t}^\mathrm{x}\times N_2^\mathrm{a}}$ considering grids $\left\{\ddot{\phi}_{j_2}^{\shortparallel}|\,j_2 = 1,\ldots,N_2^\mathrm{a}\right\}$ and $\mathbf{A}_\mathrm{t}^{\perp} = \left[\mathbf{a}(\ddot{\phi}_1^{\perp}),\ldots,\mathbf{a}(\ddot{\phi}_{N_3^\mathrm{a}}^{\perp})\right] \in \mathbb{C}^{N_\mathrm{t}^\mathrm{y}\times N_3^\mathrm{a}}$ considering grids $\{\ddot{\phi}_{j_3}^{\perp}|\,j_3 = 1,\ldots,N_3^\mathrm{a}\}$; and $\mathbf{A}_\mathrm{r} = \mathbf{A}_\mathrm{r}^{\shortparallel}\otimes\mathbf{A}_\mathrm{r}^{\perp}$ is the dictionary to evaluate azimuth and elevation AoAs, where $\mathbf{A}_\mathrm{r}^{\shortparallel} \in \mathbb{C}^{N_\mathrm{r}^\mathrm{x}\times N_4^\mathrm{a}}$ and $\mathbf{A}_\mathrm{r}^{\perp} \in \mathbb{C}^{N_\mathrm{r}^\mathrm{y}\times N_5^\mathrm{a}}$ are constructed similarly as $\mathbf{A}_\mathrm{t}^{\shortparallel}$ and $\mathbf{A}_\mathrm{t}^{\perp}$, respectively. Therefore, $N_1^\mathrm{a} = N_\mathrm{d}^\mathrm{a}$, $N_\mathrm{t}^\mathrm{a} = N_2^\mathrm{a}N_3^\mathrm{a}$, and

$N_{\mathrm{r}}^{\mathrm{a}} = N_4^{\mathrm{a}} N_5^{\mathrm{a}}$. In addition, $\mathbf{x} \in \mathbb{C}^{N_{\mathrm{r}}^{\mathrm{a}} N_{\mathrm{t}}^{\mathrm{a}} N_{\mathrm{d}}^{\mathrm{a}} \times 1}$ is the sparse vector to be estimated the supports of which are the complex gains. To solve the following sparse recovery problem for channel estimation:

$$\min_{\mathbf{x}} \left(\sum_{m=1}^{M} \left\| \mathrm{vec}(\check{\mathbf{Y}}_m) - \boldsymbol{\Upsilon}_m \boldsymbol{\Psi} \mathbf{x} \right\|^2 \right), \tag{3.15}$$

conventional OMP iteratively finds the supports of $\mathbf{x}$, denoted as a set $\mathfrak{x} = \{ j \mid [\mathbf{x}]_j \neq 0 \}, |\mathfrak{x}| \leq N_{\mathrm{est}}$ where N_{est} is the number of channel components, based on peaks of the correlation with the residual calculated from the subspace projection. Once the supports are determined, the corresponding atoms in $\boldsymbol{\Psi}$, i.e., $\left\{ [\boldsymbol{\Psi}]_{:,\ell_{\mathrm{s}}} | \ell_{\mathrm{s}} \in \mathfrak{x} \right\}$, indicate the estimated delays and angles. The searching space size of the algorithm is $\prod_{k=1}^{5} N_k^{\mathrm{a}}$, and with large antenna array and wide bandwidth for fine angular and delay domain resolutions, the resulted complexity $\mathcal{O}\left(N_{\mathrm{est}} N_{\mathrm{s}} Q \prod_{k=1}^{5} N_k^{\mathrm{s}} N_k^{\mathrm{a}} \right)$, where $N_1^{\mathrm{s}} = N_{\mathrm{d}}$, $N_2^{\mathrm{s}} = N_{\mathrm{t}}^{\mathrm{x}}$, $N_3^{\mathrm{s}} = N_{\mathrm{t}}^{\mathrm{y}}$, $N_4^{\mathrm{s}} = N_{\mathrm{r}}^{\mathrm{x}}$, and $N_5^{\mathrm{s}} = N_{\mathrm{r}}^{\mathrm{y}}$, becomes prohibitive. To cope with the complexity issue, MOMP first formulates the problem with multidimensional operations:

$$\min_{\mathbf{X}} \left(\sum_{m=1}^{M} \left\| \mathrm{vec}(\check{\mathbf{Y}}_m) - \sum_{\mathbf{i} \in \mathfrak{I}} \sum_{\mathbf{j} \in \mathfrak{J}} [\boldsymbol{\Phi}_m]_{:,\mathbf{i}} \left(\prod_{k=1}^{5} [\boldsymbol{\Psi}_k]_{i_k, j_k} \right) [\mathbf{X}]_{\mathbf{j}} \right\|^2 \right), \tag{3.16}$$

where $\mathbf{i} = (i_1, ..., i_5) \in \mathbb{N}_+^5$ and $\mathbf{j} = (j_1, ..., j_5) \in \mathbb{N}_+^5$ are multidimensional indices, and $\mathfrak{I} = \left\{ \mathbf{i} | i_k = 1, ..., N_k^{\mathrm{s}} \right\}$ and $\mathfrak{J} = \left\{ \mathbf{j} | j_k = 1, ..., N_k^{\mathrm{a}} \right\}$ are the index sets. The measurement tensor $\boldsymbol{\Phi}_m \in \mathbb{C}^{Q N_{\mathrm{s}} \otimes_{k=1}^{5} N_k^{\mathrm{s}}}$ relates to $\boldsymbol{\Upsilon}_m$ as $[\boldsymbol{\Phi}_m]_{:,\mathbf{i}} = [\boldsymbol{\Upsilon}_m]_{:,f_{\mathbf{i}}}$ where $f_{\mathbf{i}} = \left(\sum_{k=1}^{4} (i_k - 1) (\prod_{k'=k+1}^{5} N_{k'}^{\mathrm{s}}) \right) + i_5$. The single dictionary $\boldsymbol{\Psi}$ calculated from the Kronecker product as in (3.14) is separated into five independent dictionaries $\boldsymbol{\Psi}_k \in \mathbb{C}^{N_k^{\mathrm{s}} \times N_k^{\mathrm{a}}}$ for the five dimensions associated with delay, azimuth and elevation AoD, and azimuth and elevation AoA, i.e., $\boldsymbol{\Psi}_1 = \mathbf{A}_{\mathrm{d}}$, $\boldsymbol{\Psi}_2 = \mathbf{A}_{\mathrm{t}}^{\shortparallel}$, $\boldsymbol{\Psi}_3 = \mathbf{A}_{\mathrm{t}}^{\perp}$, $\boldsymbol{\Psi}_4 = \mathbf{A}_{\mathrm{r}}^{\shortparallel}$, and $\boldsymbol{\Psi}_5 = \mathbf{A}_{\mathrm{r}}^{\perp}$. Finally, the sparse vector $\mathbf{x}$ becomes the sparse tensor $\mathbf{X} \in \mathbb{C}^{\otimes_{k=1}^{5} N_k^{\mathrm{a}}}$ to be estimated, where $[\mathbf{X}]_{\mathbf{j}} = [\mathbf{x}]_{f_{\mathbf{j}}}$ with $f_{\mathbf{j}} = \left(\sum_{k=1}^{4} (j_k - 1) (\prod_{k'=k+1}^{5} N_{k'}^{\mathrm{a}}) \right) + j_5$. To solve (3.16), alternating maximization is adopted to estimate the parameters for each dimension independently, with the cost of N_{iter} iterations to refine the estimates per dimension. Thus, the computational complexity is reduced to $\mathcal{O}\left(N_{\mathrm{est}} N_{\mathrm{s}} Q N_{\mathrm{iter}} (\sum_{k=1}^{5} N_k^{\mathrm{a}}) (\prod_{k=1}^{5} N_k^{\mathrm{s}}) \right)$ compared with the conventional OMP, where the product term $\prod_{k=1}^{5} N_k^{\mathrm{a}}$ is transformed into a summation $N_{\mathrm{iter}} (\sum_{k=1}^{5} N_k^{\mathrm{a}})$. However,

the complexity can still explode when employing large antenna arrays and wide bandwidth as in the product term $\prod_{k=1}^{5} N_k^{\mathrm{s}}$. We hereby propose the F-MOMP algorithm that transforms the term into a summation for complexity reduction, while sticking with alternating maximization to determine the estimates for each dimension.

The F-MOMP algorithm reduces the complexity of calculating the product of measurement and dictionary matrices. We first expand $\boldsymbol{\Upsilon}_m$ and $\boldsymbol{\Psi}$ as (3.10) and (3.11), and based on the element-wise correspondence for the product operation, the product of $[\boldsymbol{\Upsilon}_m]_{(q-1)N_{\mathrm{s}}+1:qN_{\mathrm{s}},:}$ and $[\boldsymbol{\Psi}]_{:,f_{\mathbf{j}}}$ can be derived by factoring and then computing the product for each factor, as detailed in (3.12). Therefore, the sparse recovery problem can be written as

$$\min_{\mathbf{X}} \sum_{m=1}^{M} \sum_{q=1}^{Q} \left\| \check{\mathbf{y}}_m[q] - \sum_{\mathbf{j}\in\mathfrak{J}} \left[\zeta_q^{\mathrm{S}}(j_1)\right]^{\mathsf{T}} \left[\zeta_m^{\mathrm{F}}(j_2,j_3)\right] \left[\zeta_m^{\mathrm{W}}(j_4,j_5)\right] \mathbf{X}_{\mathbf{j}} \right\|^2, \tag{3.17}$$

where

$$\zeta_q^{\mathrm{S}}(j_1) = \sqrt{P_{\mathrm{t}}} \left[\mathbf{s}[q], \mathbf{s}[q-1], \ldots, \mathbf{0}\right] \mathbf{p}(\ddot{\imath}_{j_1}) \in \mathbb{C}^{N_{\mathrm{s}}\times 1}; \tag{3.18}$$

$$\zeta_m^{\mathrm{F}}(j_2,j_3) = \mathbf{F}_m^{\mathsf{T}} \bar{\mathbf{a}}(\ddot{\boldsymbol{\phi}}_{j_2}^{\shortparallel}, \ddot{\boldsymbol{\phi}}_{j_3}^{\perp}) \in \mathbb{C}^{N_{\mathrm{s}}\times 1}; \tag{3.19}$$

$$\zeta_m^{\mathrm{W}}(j_4,j_5) = \mathbf{W}_m^{*} \mathbf{a}(\ddot{\boldsymbol{\theta}}_{j_4}^{\shortparallel}, \ddot{\boldsymbol{\theta}}_{j_5}^{\perp}) \in \mathbb{C}^{N_{\mathrm{s}}\times 1}, \tag{3.20}$$

and $\mathbf{X} \in \mathbb{C}^{\otimes_{k=1}^{5} N_k^{\mathrm{a}}}$ defined the same as that in (3.16) is the sparse tensor to be estimated with the MOMP algorithm. In the conventional MOMP, there is a step for estimate initialization for each dimension based on cost function approximation, then the alternating maximization algorithm is adopted to iteratively refine the estimates per dimension. However, in the channel tracking scenario, the channel estimates at time τ_{n-1} can be used as the estimate initialization at time τ_n, i.e., the estimated support set at τ_{n-1}:

$$\hat{\mathfrak{J}}_{\mathrm{sup}}^{(\tau_{n-1})} = \left\{ \hat{\mathbf{j}}_1^{(\tau_{n-1})}, \ldots, \hat{\mathbf{j}}_{N_{\mathrm{est}}}^{(\tau_{n-1})} \right\}, \tag{3.21}$$

where $\hat{\mathbf{j}}_{n_{\text{est}}}^{(\tau_{n-1})} = \left(\hat{j}_{1,n_{\text{est}}}^{(\tau_{n-1})}, ..., \hat{j}_{5,n_{\text{est}}}^{(\tau_{n-1})}\right)$, provides the initialization for $\hat{j}_{k,n_{\text{est}}}^{(\tau_n)}$ at time τ_n as $\hat{j}_{k,n_{\text{est}}}^{(\tau_n)} \leftarrow \hat{j}_{k,n_{\text{est}}}^{(\tau_{n-1})}$. In addition, the dictionaries at τ_n are constructed based on historical estimates as well. Let independent dictionaries $\mathbf{\Psi}_k$ defined previously be the full dictionaries used usually for initial access stages, we define the reduced dictionaries as $\mathbf{\Psi}_{k,n_{\text{est}}}^{(\tau_n)}$ for the n_{est}-th channel component at τ_n, where the atoms from the full dictionaries corresponding to the previous estimates and their neighboring atoms are included:

$$\mathbf{\Psi}_{k,n_{\text{est}}}^{(\tau_n)} = [\mathbf{\Psi}_k]_{:,\hat{j}_{k,n_{\text{est}}}^{(\tau_{n-1})} - g_k : \hat{j}_{k,n_{\text{est}}}^{(\tau_{n-1})} + g_k}, \tag{3.22}$$

where g_k is the number of spanning grids for the neighboring atoms depending on the grid resolution of each $\mathbf{\Psi}_k$. Even with the reduced dictionaries, directly applying the OMP algorithm remains computationally intensive considering the high dictionary resolutions required for precise localization, especially when increasing g_k for a larger searching space. Hence, we rely on the alternating maximization algorithm as in MOMP to iteratively estimate the support of each dimension for the n_{est}-th channel component by solving the following optimization problem (the upper right time index "(τ_n)" is omitted for simplicity):

$$\max_{j_{k,n_{\text{est}}}} \sum_{m=1}^{M} \frac{\left|\left(\mathbf{\Upsilon}_m[\mathbf{\Psi}]_{:,f_{\mathbf{j}_{n_{\text{est}}}}}\right)^* \text{vec}(\check{\mathbf{Y}}_m^{\text{res}})\right|}{\left\|\mathbf{\Upsilon}_m[\mathbf{\Psi}]_{:,f_{\mathbf{j}_{n_{\text{est}}}}}\right\|_2} \tag{3.23}$$

$$\text{s.t.} \quad j_{k,n_{\text{est}}} \in \mathfrak{J}_{k,n_{\text{est}}},\ \mathbf{j}_{n_{\text{est}}} \notin \hat{\mathfrak{J}}_{\text{sup}} \tag{3.24}$$

where $\mathfrak{J}_{k,n_{\text{est}}} = \left\{\hat{j}_{k,n_{\text{est}}}^{(\tau_{n-1})} - g_k, ..., \hat{j}_{k,n_{\text{est}}}^{(\tau_{n-1})} + g_k\right\}$, and $\check{\mathbf{Y}}_m^{\text{res}}$ –which is initialized using $\check{\mathbf{Y}}_m$– represents the residual after the subspace projection using the estimated supports. Specifically, in each optimization iteration $n_{\text{iter}} \leq N_{\text{iter}}$, the algorithm sequentially optimizes the estimate $\hat{j}_{k,n_{\text{est}}}$ while fixing estimates of other dimensions $\hat{j}_{k',n_{\text{est}}}$, $k' \neq k$, until every $\hat{j}_{k,n_{\text{est}}}$ is obtained. Thereafter, delay and angle estimates of each path are determined by indexing the grid values in the dictionaries using $\hat{\mathbf{j}}_{n_{\text{est}}}$. Finally, the estimated complex gain for each path $\hat{\alpha}_{n_{\text{est}}}$

is acquired based on the estimated sparse vector as $\hat{\alpha}_{n_{\text{est}}} = [\hat{\mathbf{x}}]_{n_{\text{est}}}$. The pseudo codes of the F-MOMP algorithm are presented in Algorithm 3. The algorithm results in a complexity of $\mathcal{O}\left(N_{\text{est}}N_{\text{s}}QN_{\text{iter}}(\sum_{k=1}^{5}N_k^{\text{a}})(N_1^{\text{a}}+N_2^{\text{s}}N_3^{\text{s}}+N_4^{\text{s}}N_5^{\text{s}})\right)$, which reduces the complexity by turning the multiplication term into the summation comparing with the MOMP [105, 117], as specified in Table 3.3, while allows simultaneously estimating parameters across the five dimensions for delay, azimuth and elevation AoDs, and azimuth and elevation AoAs. We denote the estimated channel at τ_n containing N_{est} estimated paths without the compensation for the time-varying clock offset $t_{\text{off}}^{(\tau_n)}$ and orientation $\varpi^{(\tau_n)}$ as

$$\hat{\underline{\mathbf{Z}}}_{\tau_n} = \left[\hat{\boldsymbol{\alpha}}_{\tau_n}, \hat{\underline{\mathbf{t}}}_{\tau_n}, \hat{\underline{\boldsymbol{\theta}}}^{\text{az}}_{\tau_n}, \hat{\boldsymbol{\theta}}^{\text{el}}_{\tau_n}, \hat{\boldsymbol{\phi}}^{\text{az}}_{\tau_n}, \hat{\boldsymbol{\phi}}^{\text{el}}_{\tau_n}\right] \in \mathbb{R}^{N_{\text{est}}\times 6}, \tag{3.25}$$

where $\hat{\boldsymbol{\alpha}}_{\tau_n} = \left[\left|\hat{\alpha}_1^{(\tau_n)}\right|, ..., \left|\hat{\alpha}_{N_{\text{est}}}^{(\tau_n)}\right|\right]^{\mathsf{T}}$ (the phase is irrelevant to acquire the position and orientation estimation [121]), $\hat{\underline{\mathbf{t}}}_{\tau_n} = \left[\hat{t}_1^{(\tau_n)} - \hat{t}_{\text{off}}^{(\tau_n)}, ..., \hat{t}_{N_{\text{est}}}^{(\tau_n)} - \hat{t}_{\text{off}}^{(\tau_n)}\right]^{\mathsf{T}}$, $\hat{\underline{\boldsymbol{\theta}}}^{\text{az}}_{\tau_n} = \left[\hat{\theta}_1^{\text{az}(\tau_n)} - \hat{\varpi}^{(\tau_n)}, ..., \hat{\theta}_{N_{\text{est}}}^{\text{az}(\tau_n)} - \hat{\varpi}^{(\tau_n)}\right]^{\mathsf{T}}$, $\hat{\boldsymbol{\theta}}^{\text{el}}_{\tau_n} = \left[\hat{\theta}_1^{\text{el}(\tau_n)}, ..., \hat{\theta}_{N_{\text{est}}}^{\text{el}(\tau_n)}\right]^{\mathsf{T}}$, $\hat{\boldsymbol{\phi}}^{\text{az}}_{\tau_n} = \left[\hat{\phi}_1^{\text{az}(\tau_n)}, ..., \hat{\phi}_{N_{\text{est}}}^{\text{az}(\tau_n)}\right]^{\mathsf{T}}$, and $\hat{\boldsymbol{\phi}}^{\text{el}}_{\tau_n} = \left[\hat{\phi}_1^{\text{el}(\tau_n)}, ..., \hat{\phi}_{N_{\text{est}}}^{\text{el}(\tau_n)}\right]^{\mathsf{T}}$.

3.3.2 VO-ChAT for Vehicle Orientation Tracking

The absence of orientation knowledge limits the vehicle's estimated angles to relative values w.r.t the planar array rather than the absolute values in the global coordinate system. Consequently, precisely determining the vehicle's position becomes infeasible. While orientation changes, influenced by multiple factors including road curvature, lane changes, and other

Table 3.3. Complexity comparisons for various channel estimation algorithms. Runtimes were measured using the Python time module on a Dell XPS 15 laptop (i9-13900H, 65 GB RAM). OMP fails to run with the current system settings due to excessive complexity.

Method	Complexity	Runtime (s)
Conventional OMP [80, 109]	$\mathcal{O}\left(N_{\text{est}}N_{\text{s}}Q\prod_{k=1}^{5}N_k^{\text{s}}N_k^{\text{a}}\right)$	N/A
MOMP [105, 117]	$\mathcal{O}\left(N_{\text{est}}N_{\text{s}}QN_{\text{iter}}(\sum_{k=1}^{5}N_k^{\text{a}})(\prod_{k=1}^{5}N_k^{\text{s}})\right)$	12.3 ± 1.6
F-MOMP (proposed)	$\mathcal{O}\left(N_{\text{est}}N_{\text{s}}QN_{\text{iter}}(\sum_{k=1}^{5}N_k^{\text{a}})(N_{\text{d}}+N_2^{\text{s}}N_3^{\text{s}}+N_4^{\text{s}}N_5^{\text{s}})\right)$	0.62 ± 0.13

driving dynamics [147], are challenging to model accurately, we turn to DL, specifically, the attention schemes, that have been broadly studied in prior work to address context-aware

Algorithm 3. F-MOMP for channel tracking.

Input:
Vectorized received signals $\breve{\boldsymbol{\gamma}} \leftarrow \left[\text{vec}(\breve{\mathbf{Y}}_1^{(\tau_n)}); ...; \text{vec}(\breve{\mathbf{Y}}_M^{(\tau_n)})\right]$;
Previous estimated supports $\hat{\mathfrak{J}}_{\text{sup}}^{(\tau_{n-1})}$ as in (3.21);
The number of channel components N_{est};
Initialize:
The estimated support set $\hat{\mathfrak{J}}_{\text{sup}}^{(\tau_n)} \leftarrow \emptyset$;
The subspace projection residual $\breve{\boldsymbol{\gamma}}^{\text{res}} \leftarrow \breve{\boldsymbol{\gamma}}$;
for $n_{\text{est}} = 1 : N_{\text{est}}$ **do**
  for $k = 1 : 5$ **do**
    Initialize support estimates $\hat{j}_{k,n_{\text{est}}}^{(\tau_n)} \leftarrow \hat{j}_{k,n_{\text{est}}}^{(\tau_{n-1})}$;
    Construct reduced dictionaries $\boldsymbol{\Psi}_{k,n_{\text{est}}}^{(\tau_n)}$ as in (3.22);
    Form index sets $\mathfrak{J}_k \leftarrow \left\{\hat{j}_{k,n_{\text{est}}}^{(\tau_n)} - g_k, ..., \hat{j}_{k,n_{\text{est}}}^{(\tau_n)} + g_k\right\}$;
  end for
  % Factor calculation
  $\boldsymbol{\Xi}_{j_1}^{\text{S}} \leftarrow \left[\zeta_1^{\text{S}}(j_1)^{\mathsf{T}}; ...; \zeta_Q^{\text{S}}(j_1)^{\mathsf{T}}\right]$ for $j_1 \in \mathfrak{J}_1$;
  $\boldsymbol{\Xi}_{j_2,j_3}^{\text{F}} \leftarrow \left[\zeta_1^{\text{F}}(j_2, j_3), ..., \zeta_M^{\text{F}}(j_2, j_3)\right]$ for $j_2 \in \mathfrak{J}_2$, $j_3 \in \mathfrak{J}_3$;
  $\boldsymbol{\Xi}_{j_4,j_5}^{\text{W}} \leftarrow \left[\zeta_1^{\text{W}}(j_4, j_5), ..., \zeta_M^{\text{W}}(j_4, j_5)\right]$ for $j_4 \in \mathfrak{J}_4$, $j_5 \in \mathfrak{J}_5$;
  for $n_{\text{iter}} = 1 : N_{\text{iter}}$ **do**
    for $k = 1 : 5$ **do**
      $\mathfrak{J} \leftarrow \{\mathbf{j} | j_k \in \mathfrak{J}_k, j_{k'} = \hat{j}_{k',\text{est}}^{(\tau_n)}, k' \neq k\} \backslash \hat{\mathfrak{J}}_{\text{sup}}^{(\tau_n)}$;
      $\boldsymbol{\xi}_{\mathbf{j}} \leftarrow \text{vec}\left(\left(\boldsymbol{\Xi}_{j_1}^{\text{S}} \boldsymbol{\Xi}_{j_2,j_3}^{\text{F}}\right) \odot \boldsymbol{\Xi}_{j_4,j_5}^{\text{W}}\right), \mathbf{j} \in \mathfrak{J}$, as in (3.12);
      $\hat{j}_{k,n_{\text{est}}}^{(\tau_n)} \leftarrow \arg\max_{j_k} \frac{|\boldsymbol{\xi}_{\mathbf{j}}^* \breve{\boldsymbol{\gamma}}^{\text{res}}|}{\|\boldsymbol{\xi}_{\mathbf{j}}\|_2}$ for solving (3.23);
    end for
  end for
  Collect support estimates $\hat{\mathbf{j}}_{n_{\text{est}}}^{(\tau_n)} = \left(\hat{j}_{1,n_{\text{est}}}^{(\tau_n)}, ..., \hat{j}_{5,n_{\text{est}}}^{(\tau_n)}\right)$;
  Retrieve channel parameters $\hat{t}_{n_{\text{est}}}^{(\tau_n)} = \ddot{t}_{\hat{j}_{1,n_{\text{est}}}^{(\tau_n)}}$, $\hat{\phi}_{n_{\text{est}}}^{\shortparallel(\tau_n)} = \ddot{\phi}_{\hat{j}_{2,n_{\text{est}}}^{(\tau_n)}}^{\shortparallel}$, $\hat{\phi}_{n_{\text{est}}}^{\perp(\tau_n)} = \ddot{\phi}_{\hat{j}_{3,n_{\text{est}}}^{(\tau_n)}}^{\perp}$, $\hat{\theta}_{n_{\text{est}}}^{\shortparallel(\tau_n)} = \ddot{\theta}_{\hat{j}_{4,n_{\text{est}}}^{(\tau_n)}}^{\shortparallel}$, and $\hat{\theta}_{n_{\text{est}}}^{\perp(\tau_n)} = \ddot{\theta}_{\hat{j}_{5,n_{\text{est}}}^{(\tau_n)}}^{\perp}$;
  Update support set $\hat{\mathfrak{J}}_{\text{sup}}^{(\tau_n)} \leftarrow \hat{\mathfrak{J}}_{\text{sup}}^{(\tau_n)} \cup \left\{\hat{\mathbf{j}}_{n_{\text{est}}}^{(\tau_n)}\right\}$;
  % Subspace projection and residual update
  $\hat{\mathbf{x}} \leftarrow \left[\boldsymbol{\xi}_{\hat{\mathbf{j}}_1^{(\tau_n)}}, ..., \boldsymbol{\xi}_{\hat{\mathbf{j}}_{n_{\text{est}}}^{(\tau_n)}}\right]^{\dagger} \breve{\boldsymbol{\gamma}}$;
  $\breve{\boldsymbol{\gamma}}^{\text{res}} \leftarrow \breve{\boldsymbol{\gamma}} - \left[\boldsymbol{\xi}_{\hat{\mathbf{j}}_1^{(\tau_n)}}, ..., \boldsymbol{\xi}_{\hat{\mathbf{j}}_{n_{\text{est}}}^{(\tau_n)}}\right] \hat{\mathbf{x}}$;
end for
Retrieve path complex gains where $\hat{\alpha}_{n_{\text{est}}} = [\hat{\mathbf{x}}]_{n_{\text{est}}}$;
Output: $\hat{\mathfrak{J}}_{\text{sup}}^{(\tau_n)}$ and estimated channel parameters for each path.

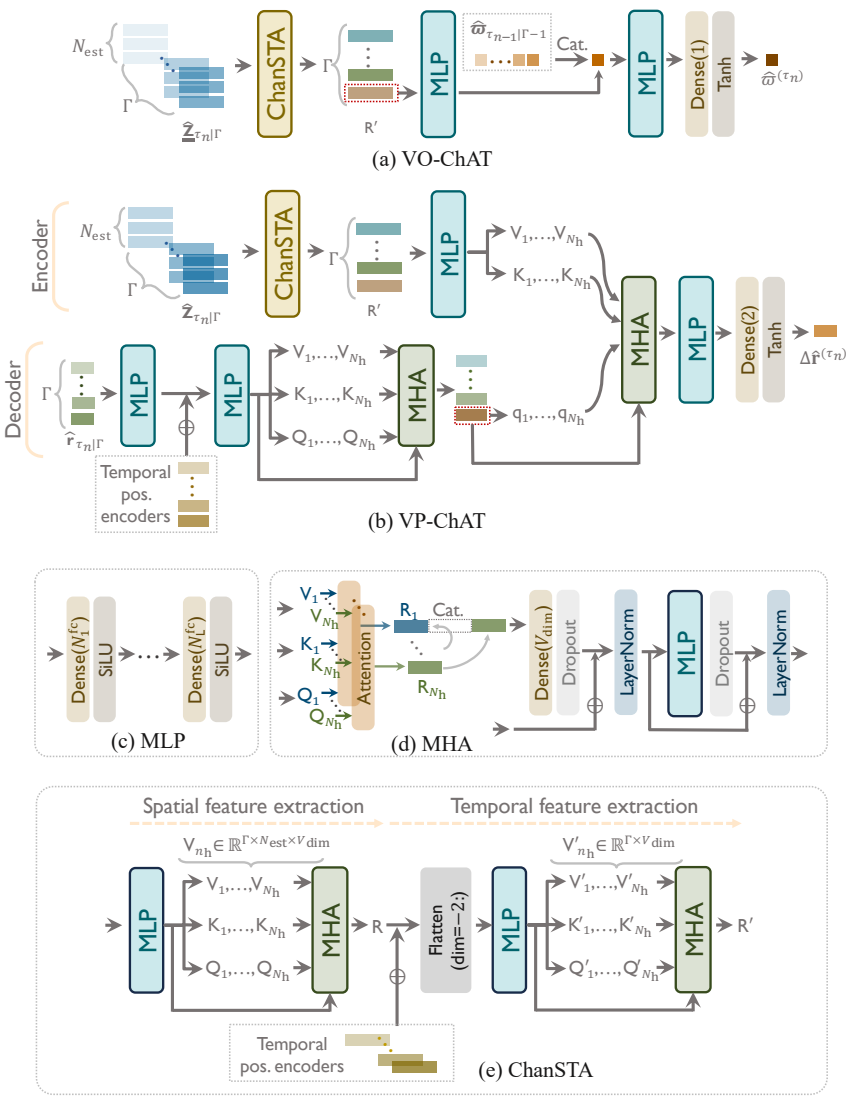


Figure 3.3. (a) VO-ChAT architecture for vehicle orientation tracking; (b) VP-ChAT architecture exchanging information between the estimated channel sequence and the trajectory to refine single-shot position estimates; (c) multilayer perceptron (MLP) layer construction; (d) multi-head attention (MHA) module design; (e) ChanSTA module performing spatial then temporal feature extraction on the channel sequence.

problems [148]. Our task of PO tracking is inherently a time series forecasting problem, thus we adopt Transformer-inspired network architectures due to its demonstrated effectiveness in capturing complex long-range dependencies and global contextual information for accurate sequential prediction tasks [149]. For orientation tracking with VO-ChAT, as illustrated in Fig. 3.3a, the core self-attention mechanism dynamically assesses the importance and confidence

of estimated channel MPCs, and integrates channel variations across time to infer the updates in orientation. VO-ChAT can be formulated as (3.26) which takes in the estimated channel sequence and previous orientation estimates to infer the current orientation

$$\hat{\boldsymbol{\omega}}^{(\tau_n)} = \text{VO-ChAT}\left(\hat{\underline{\mathbf{Z}}}_{\tau_n|\Gamma}, \hat{\boldsymbol{\omega}}_{\tau_{n-1}|\Gamma-1}; \mathbf{W}^{\text{o}}\right), \tag{3.26}$$

where Γ is the length of history information, $\hat{\underline{\mathbf{Z}}}_{\tau_n|\Gamma} = \text{stack}\left(\hat{\underline{\mathbf{Z}}}_{\tau_{n-\Gamma+1}}, ..., \hat{\underline{\mathbf{Z}}}_{\tau_n}\right) \in \mathbb{R}^{\Gamma \times N_{\text{est}} \times 6}$ is the estimated channel sequence, $\hat{\boldsymbol{\omega}}_{\tau_{n-1}|\Gamma-1} = \left[\hat{\boldsymbol{\omega}}^{(\tau_{n-\Gamma+1})}, ..., \hat{\boldsymbol{\omega}}^{(\tau_{n-1})}\right] \in \mathbb{R}^{\Gamma-1}$ is the vector containing previous determined orientations from $\tau_{n-\Gamma+1}$ to τ_{n-1}, and $\mathbf{W}^{\text{o}}$ represents the learnable network matrices.

VO-ChAT firstly processes $\hat{\underline{\mathbf{Z}}}_{\tau_n|\Gamma}$ with a channel spatial and temporal attention (ChanSTA) module depicted in Fig. 3.3e, which is composed of two self multi-head attention (MHA) blocks (Fig. 3.3d), one for extracting spatial features w.r.t the estimated N_{est} paths at each time step, and the other one for analyzing the temporal channel evolution features for the input channel sequence of length Γ. Specifically, $\hat{\underline{\mathbf{Z}}}_{\tau_n|\Gamma}$ goes through a multilayer perceptron (MLP) module consisting of dense layers and activation layers to obtain three types of abstract representations: *Value* $\{\mathbf{V}_1, ..., \mathbf{V}_{N_{\text{h}}}\}$ with $\mathbf{V}_{n_{\text{h}}} \in \mathbb{R}^{\Gamma \times N_{\text{est}} \times V_{\text{dim}}}$, *Key* $\{\mathbf{K}_1, ..., \mathbf{K}_{N_{\text{h}}}\}$ with $\mathbf{K}_{n_{\text{h}}} \in \mathbb{R}^{\Gamma \times N_{\text{est}} \times K_{\text{dim}}}$, and *Query* $\{\mathbf{Q}_1, ..., \mathbf{Q}_{N_{\text{h}}}\}$ with $\mathbf{Q}_{n_{\text{h}}} \in \mathbb{R}^{\Gamma \times N_{\text{est}} \times Q_{\text{dim}}}$, where V_{dim}, K_{dim}, and $Q_{\text{dim}} = K_{\text{dim}}$ are the embedding dimensions for each type of the representations, and N_{h} is the number of attention heads in the MHA mechanism. The attention operation for the n_{h}-th head to extract path-wise spatial features is mathematically formulated as

$$[\mathbf{R}_{n_{\text{h}}}]_{\text{i,k,:}} = \text{Attention}\Big([\mathbf{Q}_{n_{\text{h}}}]_{\text{i,k,:}}, [\mathbf{K}_{n_{\text{h}}}]_{\text{i,:,:}}, [\mathbf{V}_{n_{\text{h}}}]_{\text{i,:,:}}\Big) \tag{3.27}$$

$$= \text{Softmax}\left(\frac{[\mathbf{Q}_{n_{\text{h}}}]_{\text{i,k,:}}, [\mathbf{K}_{n_{\text{h}}}]^{\mathsf{T}}_{\text{i,:,:}}}{\sqrt{K_{\text{dim}}}}\right)[\mathbf{V}_{n_{\text{h}}}]_{\text{i,:,:}}. \tag{3.28}$$

Here, $\mathbf{R}_{n_{\text{h}}}$ has the same shape as $\mathbf{V}_{n_{\text{h}}}$, and each row of $[\mathbf{R}_{n_{\text{h}}}]_{\text{i,:,:}}$ corresponds to an estimated channel path factoring in other paths' information within the same estimation time frame, i.e.,

paths with higher estimation confidence at each time step should be prioritized for subsequent processing. As shown in Fig. 3.3d, outputs from individual attention heads are concatenated and processed through a dense layer resulting in a dimension of V_{dim}, and then connected to the normalization layers to ensure consistent feature scaling and effective information propagation through the network. Let $\mathbf{R} \in \mathbb{R}^{\Gamma \times N_{\text{est}} \times V_{\text{dim}}}$ represent the output from the spatial feature extraction block, position encoding is then applied to $\mathbf{R}$ to preserve the chronological order of the estimated channels before temporal feature extraction. The resulting tensor is flattened along the last two dimensions with the out shape of $\Gamma \times N_{\text{est}} V_{\text{dim}}$. Subsequent processing through a MLP module generates three types of channel representations, $\{\mathbf{V}'_1, ..., \mathbf{V}'_{N_{\text{h}}}\}$ with $\mathbf{V}'_{n_{\text{h}}} \in \mathbb{R}^{\Gamma \times V_{\text{dim}}}$, $\{\mathbf{K}'_1, ..., \mathbf{K}'_{N_{\text{h}}}\}$ with $\mathbf{K}'_{n_{\text{h}}} \in \mathbb{R}^{\Gamma \times K_{\text{dim}}}$, and $\{\mathbf{Q}'_1, ..., \mathbf{Q}'_{N_{\text{h}}}\}$ with $\mathbf{Q}'_{n_{\text{h}}} \in \mathbb{R}^{\Gamma \times Q_{\text{dim}}}$, where each row of the representation matrices corresponds to information from a specific time step. Thereafter, the three types of representations are processed through the MHA module where the attention operation for the n_{h}-th attention head becomes $[\mathbf{R}'_{n_{\text{h}}}]_{\text{i},:} = \text{Attention}\left([\mathbf{Q}'_{n_{\text{h}}}]_{\text{i},:}, \mathbf{K}'_{n_{\text{h}}}, \mathbf{V}'_{n_{\text{h}}}\right)$. The resulting output passes through a dense layer and is residually connected to the normalization layers, similar to the structure of the preceding spatial feature extraction block. The output from the temporal feature extraction block is the representation $\mathbf{R}' \in \mathbb{R}^{\Gamma \times V_{\text{dim}}}$ which emphasizes more accurately estimated channels within the sequence and incorporates temporal evolution features. To acquire the orientation estimate at the current time step, the final row of $\mathbf{R}'$, i.e., $[\mathbf{R}']_{-1,:}$ indicating the current information, is selected to go through MLP layers, along with the concatenated previous orientation estimates $\hat{\boldsymbol{\varpi}}_{\tau_{n-1}|\Gamma-1}$, to produce the current orientation estimate $\hat{\boldsymbol{\varpi}}^{(\tau_n)}$. Notably, the channel estimates provide essential information about the propagation environment, and enable the network to adjust and mitigate sequential orientation prediction errors. In summary, this approach achieves orientation tracking by incorporating channel and orientation histories, leveraging the temporal consistency between channel variations and orientation changes, and capturing the inherent relationship between vehicle motion and channel evolution. This decoupled spatial-then-temporal design is motivated by the observation that spatial multipath structure changes slowly across time steps, while temporal evolution captures vehicle motion dynamics.

The algorithm for orientation tracking is summarized in Algorithm 4.

3.3.3 Geometric Transformation for Single-Shot Localization

Once the vehicle orientation $\hat{\boldsymbol{\varpi}}^{(\tau_n)}$ is determined, the estimated relative azimuth AoAs can be compensated to acquire the angles in the global coordinate system as $\hat{\boldsymbol{\theta}}^{\mathrm{az}}_{\tau_n} = \underline{\hat{\boldsymbol{\theta}}}^{\mathrm{az}}_{\tau_n} + \hat{\boldsymbol{\varpi}}^{(\tau_n)} = \left[\hat{\theta}^{\mathrm{az}(\tau_n)}_1, ..., \hat{\theta}^{\mathrm{az}(\tau_n)}_{N_{\mathrm{est}}}\right]^{\mathsf{T}}$. To derive the vehicle's position, we first leverage DoD and DoA in the form of unitary vectors, denoted as $\overrightarrow{\varphi_\ell} = [\cos(\phi^{\mathrm{el}}_\ell)\cos(\phi^{\mathrm{az}}_\ell), \cos(\phi^{\mathrm{el}}_\ell)\sin(\phi^{\mathrm{az}}_\ell), \sin(\phi^{\mathrm{el}}_\ell)]^{\mathsf{T}}$ and $\overrightarrow{\vartheta_\ell} = [\cos(\theta^{\mathrm{el}}_\ell)\cos(\theta^{\mathrm{az}}_\ell), \cos(\theta^{\mathrm{el}}_\ell)\sin(\theta^{\mathrm{az}}_\ell), \sin(\theta^{\mathrm{el}}_\ell)]^{\mathsf{T}}$, and formulate the geometric relationship between the BS and the vehicle for each first order reflection ℓ satisfying

$$\mathbf{r}^{(\tau_n)}_{\mathrm{v}} + d^{\vartheta(\tau_n)}_\ell \cdot \overrightarrow{\vartheta}^{(\tau_n)}_\ell = \mathbf{r}_{\mathrm{B}} + d^{\varphi(\tau_n)}_\ell \cdot \overrightarrow{\varphi}^{(\tau_n)}_\ell; \tag{3.29}$$

$$d^{\vartheta(\tau_n)}_\ell + d^{\varphi(\tau_n)}_\ell = \left(t^{(\tau_n)}_\ell + t^{(\tau_n)}_{\mathrm{off}}\right) \cdot v_{\mathrm{c}}, \tag{3.30}$$

where $\mathbf{r}^{(\tau_n)}_{\mathrm{v}}$ is the vehicle's 3D position at τ_n, $\mathbf{r}_{\mathrm{B}}$ is the known array position on the BS, $d^{\vartheta(\tau_n)}_\ell$ is the distance between the vehicle and the scattering point, $d^{\varphi(\tau_n)}_\ell$ is the distance between the scattering point and the BS, and v_{c} is the light speed. Note that (3.29) and (3.30) hold for LOS situations as well assuming a pseudo scattering point in the middle of the LOS path. Before resolving (3.29) and (3.30) to determine the vehicle's position, it is imperative to identify and select the LOS/first order reflections, as it allows for the exclusion of higher order MPCs that will introduce errors in the following localization process. While our previous work [121] employs a neural network for path

Algorithm 4. Orientation Tracking using VO-ChAT.

1: **Input:** Previous channel estimate and orientation estimate sequences $\{\underline{\hat{\mathbf{Z}}}_{\tau_{n-1}|\Gamma-1}, \hat{\boldsymbol{\varpi}}_{\tau_{n-1}|\Gamma-1}\}$;
2: **Output:** Current slot orientation estimate: $\hat{\boldsymbol{\varpi}}^{(\tau_n)}$;
3: **for** each tracking slot τ_n **do**
4: Obtain channel estimate $\underline{\hat{\mathbf{Z}}}_{\tau_n}$ via F-MOMP as in (3.25);
5: Acquire $\hat{\mathbf{Z}}_{\tau_n|\Gamma} = \left[\hat{\mathbf{Z}}_{\tau_{n-1}|\Gamma-1}, \underline{\hat{\mathbf{Z}}}_{\tau_n}\right]$ by concatenation;
6: Apply VO-ChAT algorithm for $\hat{\boldsymbol{\varpi}}^{(\tau_n)}$ as in (3.26);
7: **end for**
8: **return** $\hat{\boldsymbol{\varpi}}^{(\tau_n)}$

order classification, we leverage the tracking scenario's inherent advantages here. Specifically, we assume the height of the vehicle array is known and remained consistent along a trajectory as $\left[\mathbf{r}_{\mathrm{v}}^{(\tau_n)}\right]_3 = h_{\mathrm{v}}$, and substitute $d_\ell^{\varphi(\tau_n)} = \left(h_{\mathrm{v}} + d_\ell^{\vartheta(\tau_n)} \cdot \left[\overrightarrow{\vartheta}_\ell^{(\tau_n)}\right]_3 - [\mathbf{r}_{\mathrm{B}}]_3\right) / \left[\overrightarrow{\varphi}_\ell^{(\tau_n)}\right]_3$ into (3.30) to derive $\hat{d}_\ell^{\vartheta(\tau_n)} = \frac{\left[\hat{\overrightarrow{\varphi}}_\ell^{(\tau_n)}\right]_3 \cdot \left(\hat{t}_\ell^{(\tau_n)} + \hat{t}_{\mathrm{off}}^{(\tau_0)}\right) \cdot v_{\mathrm{c}} + [\mathbf{r}_{\mathrm{B}}]_3 - h_{\mathrm{v}}}{\left[\hat{\overrightarrow{\varphi}}_\ell^{(\tau_n)}\right]_3 + \left[\hat{\overrightarrow{\vartheta}}_\ell^{(\tau_n)}\right]_3}$, where $\hat{t}_{\mathrm{off}}^{(\tau_0)}$ is the clock offset estimated during the initial access stage [121] and the subsequent time-varying clock offsets $t_{\mathrm{off}}^{(\tau_n)}$ should be attributed to small drifts. Then path ℓ is discarded for localization if $\hat{d}_\ell^{\vartheta(\tau_n)} \leq [\mathbf{r}_{\mathrm{B}}]_3$. In addition, the estimated path gain should be above a threshold to guarantee the channel tracking accuracy, i.e., an estimated path is also discarded if $|\hat{\alpha}_\ell| \leq |\alpha_{\mathrm{th}}|$. Afterwards, for all the selected paths $\ell \in \mathfrak{L}$, where $\mathfrak{L}$ is the set containing the estimated LOS and/or first order reflections, we substitute $d_\ell^{\varphi(\tau_n)} = v_{\mathrm{c}} t_\ell^{(\tau_n)} + v_{\mathrm{c}} t_{\mathrm{off}}^{(\tau_n)} - d_\ell^{\vartheta(\tau_n)}$ into (3.29) as

$$\begin{aligned} &\mathbf{r}_{\mathrm{v}}^{(\tau_n)} + d_\ell^{\vartheta(\tau_n)} \left(\overrightarrow{\vartheta}_\ell^{(\tau_n)} + \overrightarrow{\varphi}_\ell^{(\tau_n)}\right) - v_{\mathrm{c}} t_{\mathrm{off}}^{(\tau_n)} \overrightarrow{\varphi}_\ell^{(\tau_n)} \\ &\qquad = \mathbf{r}_{\mathrm{B}} + v_{\mathrm{c}} t_\ell^{(\tau_n)} \overrightarrow{\varphi}_\ell^{(\tau_n)}. \end{aligned} \tag{3.31}$$

Therefore, a WLS estimation problem can be formulated:

$$\begin{bmatrix} w_1 \mathbf{B}_1 \\ \vdots \\ w_{|\mathfrak{L}|} \mathbf{B}_{|\mathfrak{L}|} \end{bmatrix} \mathbf{o} = \begin{bmatrix} w_1 \mathbf{b}_1 \\ \vdots \\ w_{|\mathfrak{L}|} \mathbf{b}_{|\mathfrak{L}|} \end{bmatrix}, \tag{3.32}$$

where w_ℓ is the weight assigned to path ℓ proportional to its estimated gain $|\hat{\alpha}_\ell|$ in decibel, $\mathbf{B}_\ell = \left[\mathbf{B}'_\ell, \mathbf{B}''_\ell\right] \in \mathbb{R}^{3\times(3+|\mathfrak{L}|)}$ with $\mathbf{B}'_\ell = \begin{bmatrix} \mathbf{I}_2 & -\hat{\overrightarrow{\varphi}}_\ell^{(\tau_n)} \\ \mathbf{0}_{1\times 2} & \end{bmatrix}$ and $\mathbf{B}''_\ell \in \mathbb{R}^{3\times|\mathfrak{L}|}$ containing columns of zeros except its ℓ-th column given by $\left[\mathbf{B}''_\ell\right]_{:,\ell} = \hat{\overrightarrow{\vartheta}}_\ell^{(\tau_n)} + \hat{\overrightarrow{\varphi}}_\ell^{(\tau_n)}$, $\mathbf{b}_\ell = \mathbf{r}_{\mathrm{B}} - [0,0,h_{\mathrm{v}}]^{\mathsf{T}} + v_{\mathrm{c}} \hat{t}_\ell^{(\tau_n)} \hat{\overrightarrow{\varphi}}_\ell^{(\tau_n)}$, and the vector containing the unknown variables to be estimated with the LS estimation algorithm

is defined as

$$\mathbf{o} = \left[[\mathbf{r}_{\mathrm{v}}^{(\tau_n)}]_{1:2}^{\mathsf{T}}, v_{\mathrm{c}} t_{\mathrm{off}}^{(\tau_n)}, d_1^{\vartheta(\tau_n)}, ..., d_{|\mathfrak{L}|}^{\vartheta(\tau_n)} \right]^{\mathsf{T}}. \tag{3.33}$$

By solving (3.32), the single-shot 2D localization result is given by $\hat{\mathbf{r}}_{\mathrm{v}_{||}}^{(\tau_n)} = [\hat{\mathbf{o}}]_{1:2}$, the clock offset is determined as $\hat{t}_{\mathrm{off}}^{(\tau_n)} = \frac{[\hat{\mathbf{o}}]_3}{v_{\mathrm{c}}}$, and the estimated absolute ToAs are accordingly obtained as $\hat{\mathbf{t}}_{\tau_n} = \hat{\underline{\mathbf{t}}}_{\tau_n} + \hat{t}_{\mathrm{off}}^{(\tau_n)} = \left[\hat{t}_1^{(\tau_n)}, ..., \hat{t}_{N_{\mathrm{est}}}^{(\tau_n)} \right]$.

We denote $\hat{\mathbf{Z}}_{\tau_n|\Gamma} = \mathrm{stack}\left(\hat{\mathbf{Z}}_{\tau_{n-\Gamma+1}}, ..., \hat{\mathbf{Z}}_{\tau_n} \right)$ for the subsequent position tracking task, where $\hat{\mathbf{Z}}_{\tau_n} = \left[\hat{\boldsymbol{\alpha}}_{\tau_n}, \hat{\mathbf{t}}_{\tau_n}, \hat{\boldsymbol{\theta}}_{\tau_n}^{\mathrm{az}}, \hat{\boldsymbol{\theta}}_{\tau_n}^{\mathrm{el}}, \hat{\boldsymbol{\phi}}_{\tau_n}^{\mathrm{az}}, \hat{\boldsymbol{\phi}}_{\tau_n}^{\mathrm{el}} \right] \in \mathbb{R}^{N_{\mathrm{est}} \times 6}$ with the time offset and orientation compensated should be distinguished from $\hat{\underline{\mathbf{Z}}}_{\tau_n}$ defined in (3.25).

3.3.4 VP-ChAT for Vehicle Position Tracking

While solving (3.32) yields the single-shot localization results, incorporating historical trajectory information and calibrating $\hat{\mathbf{r}}_{\mathrm{v}_{||}}^{(\tau_n)}$ is beneficial to enhance the accuracy. To this end, we propose a second network VP-ChAT built upon the architecture of VO-ChAT, as illustrated in Fig. 3.3b. Our approach conceptualizes the position tracking task as a sequence-to-sequence translation problem. The encoder, which maintains a self-attention mechanism similar to that in VO-ChAT, is designed to capture intricate spatial and temporal multipath features from the mmWave channel. A decoder is applied to process the vehicle's trajectory, and given the strong correlation between vehicle trajectory and channel dynamics, the cross-attention module can effectively align and translate the encoded channel features into precise location features. In VP-ChAT, the ChanSTA module, structurally identical to that of VO-ChAT, together with an additional MLP module serves as an encoder, which captures the complex multipath characteristics and their temporal evolution. Concurrently, the decoder processes the trajectory information within the given time period, then generates the *query* representations of the position information to perform cross MHA with the encoder outputs to request for a correction of the

current single-shot position estimate, i.e.,

$$\Delta\hat{\mathbf{r}}^{(\tau_n)} = \text{VP-ChAT}\left(\hat{\mathbf{Z}}_{\tau_n|\Gamma}, \hat{\mathbf{r}}_{\tau_n|\Gamma}; \mathbf{W}^{\text{p}}\right), \tag{3.34}$$

where $\Delta\hat{\mathbf{r}}^{(\tau_n)}$ is the correction vector for $\hat{\mathbf{r}}_{\text{V}_{\text{II}}}^{(\tau_n)}$ so that the corrected position is $\widetilde{\mathbf{r}}_{\text{V}_{\text{II}}}^{(\tau_n)} = \hat{\mathbf{r}}_{\text{V}_{\text{II}}}^{(\tau_n)} + \Delta\hat{\mathbf{r}}^{(\tau_n)}$, $\hat{\mathbf{Z}}_{\tau_n|\Gamma}$, the channel sequence with orientation and clock offset compensated (as opposed to $\underline{\hat{\mathbf{Z}}}_{\tau_n|\Gamma}$ which retains the uncompensated offsets), serves as the encoder input, $\hat{\mathbf{r}}_{\tau_n|\Gamma}$ is the decoder input comprising the historical corrected position estimates $\widetilde{\mathbf{r}}_{\text{V}_{\text{II}}}^{(\tau_{n'})}$ for $n' = n-\Gamma+1, ..., n-1$ and the current single-shot position estimate $\hat{\mathbf{r}}_{\text{V}_{\text{II}}}^{(\tau_n)}$, denoted as $\hat{\mathbf{r}}_{\tau_n|\Gamma} = \left[\widetilde{\mathbf{r}}_{\text{V}_{\text{II}}}^{(\tau_{n-\Gamma+1})}; ...; \widetilde{\mathbf{r}}_{\text{V}_{\text{II}}}^{(\tau_{n-1})}; \hat{\mathbf{r}}_{\text{V}_{\text{II}}}^{(\tau_n)}\right] \in \mathbb{R}^{\Gamma\times 2}$, and $\mathbf{W}^{\text{p}}$ is the learnable network parameters. In detail, the encoder extracts the spatial and temporal evolution features of the estimated channels similar to that in the orientation tracking task, and generates N_{h} pairs of *value* and *key* representations of the estimated channel sequence, denoted as $\{\mathbf{V}_1, ..., \mathbf{V}_{N_{\text{h}}}\}$ with $\mathbf{V}_{n_{\text{h}}} \in \mathbb{R}^{\Gamma\times V_{\text{dim}}}$ and $\{\mathbf{K}_1, ..., \mathbf{K}_{N_{\text{h}}}\}$ with $\mathbf{K}_{n_{\text{h}}} \in \mathbb{R}^{\Gamma\times K_{\text{dim}}}$ for the following cross attention operations. At the decoder, $\hat{\mathbf{r}}_{\tau_n|\Gamma}$ is processed by MLP layers for feature expansion, followed by the positional encoding to preserve chronological order. The resulting tensor is processed to generate the three types of representations for self MHA operations, which extracts the vehicle's moving patterns and produces a sequence of dimension $\Gamma \times V_{\text{dim}}$, where each row represents the position information at a specific time step while incorporating contextual information from the entire input trajectory sequence. To facilitate current position correction, the final row of the sequence indicating the information from the current time step is transformed into N_{h} *query* vectors, denoted as $\{\mathbf{q}_1, ..., \mathbf{q}_{N_{\text{h}}}\}$, to perform cross MHA with the encoder outputs, i.e., Attention $\left(\mathbf{q}_{n_{\text{h}}}, \mathbf{K}_{n_{\text{h}}}, \mathbf{V}_{n_{\text{h}}}\right)$ for $n_{\text{h}} = 1, ..., N_{\text{h}}$. Finally, the output from the cross MHA mechanism undergoes additional processing through MLP layers and yields the current position correction vector. In summary, the encoder-decoder architecture of VP-ChAT maintains temporal coherence for the channel and trajectory sequences, and the cross attention mechanisms establish the connections among the channel evolution, the vehicle's trajectory, and system errors introduced by the channel estimation and localization methods, hereby achieving

Algorithm 5. Vehicle 2D position tracking using VP-ChAT.

Input: Channel estimate sequence (with orientation compensated) $\hat{\mathbf{Z}}_{\tau_n|\Gamma}$; Previous position tracking results $\left[\widetilde{\mathbf{r}}_{\mathrm{V}_{\mathrm{II}}}^{(\tau_{n-\Gamma+1})};\ldots;\widetilde{\mathbf{r}}_{\mathrm{V}_{\mathrm{II}}}^{(\tau_{n-1})}\right]$;
Output: Corrected current position estimate $\widetilde{\mathbf{r}}_{\mathrm{V}_{\mathrm{II}}}^{(\tau_n)}$;
for each tracking slot τ_n **do**
    Derive single-shot location estimate $\hat{\mathbf{r}}_{\mathrm{V}_{\mathrm{II}}}^{(\tau_n)}$ as in (3.32);
    Acquire $\hat{\mathbf{r}}_{\tau_n|\Gamma} = \left[\widetilde{\mathbf{r}}_{\mathrm{V}_{\mathrm{II}}}^{(\tau_{n-\Gamma+1})};\ldots;\widetilde{\mathbf{r}}_{\mathrm{V}_{\mathrm{II}}}^{(\tau_{n-1})};\hat{\mathbf{r}}_{\mathrm{V}_{\mathrm{II}}}^{(\tau_n)}\right]$ by concatenation;
    Apply VP-ChAT for current location correction $\Delta\hat{\mathbf{r}}^{(\tau_n)}$ as in (3.34);
    Correct location estimate by $\widetilde{\mathbf{r}}_{\mathrm{V}_{\mathrm{II}}}^{(\tau_n)} = \hat{\mathbf{r}}_{\mathrm{V}_{\mathrm{II}}}^{(\tau_n)} + \Delta\hat{\mathbf{r}}^{(\tau_n)}$;
end for
return $\widetilde{\mathbf{r}}_{\mathrm{V}_{\mathrm{II}}}^{(\tau_n)}$

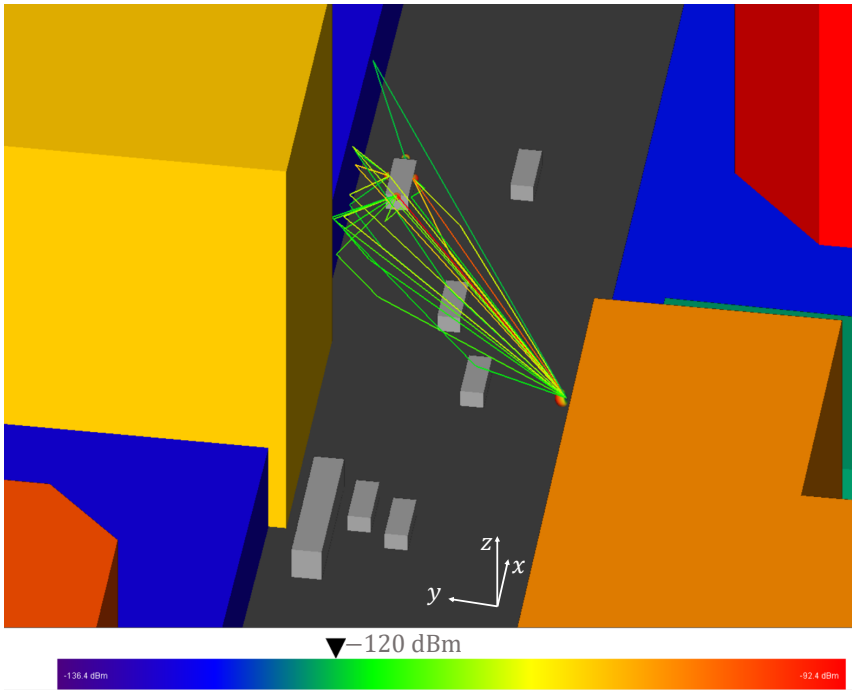


Figure 3.4. Ray-tracing simulation example in the urban canyon environment. The MPCs with gains ≥ -120 dBm are plotted.

precise position refinement. The position tracking algorithm is summarized in Algorithm 5.

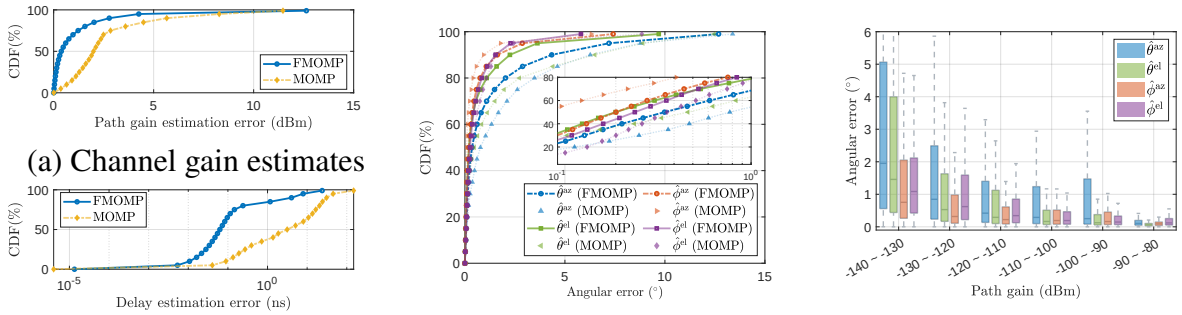
3.4 Simulation Results

This section presents the mmWave vehicular system setups, followed by the analysis of the experimental results. All experiments use ray-tracing simulated channels from a realistic urban canyon environment described below. We first present the channel tracking performance using the F-MOMP algorithm to demonstrate its accuracy for vehicle localization. Subsequently, we present the orientation prediction results using VO-ChAT and evaluate the vehicle position tracking performance using VP-ChAT after orientation compensation, comparing these results with SOTA localization methods with mmWave communication channels.

As the ray-tracing simulation snapshot depicted in Fig. 3.4, we consider an urban canyon environment within a rectangular cuboid with opposite corners located at $(x,y,z) = [-13,-123,0]$ m and $[231,85,56]$ m, respectively. The environmental configurations including the surface materials follow the settings in [150]. The cars and trucks are distributed across four lanes in the center according to the 3rd Generation Partnership Project (3GPP) standard technical report [40], and move at the speed limits assigned to each lane: 60, 50, 25, and 15 km/h. We pick an active vehicle driving at 60 km/h on the first lane for the tracking experiment, with its orientation dynamically adjusted according to the driver behavior model by setting $\mathbf{r}^{\star}(\tau + T_{\mathrm{la}})$ on the lane centerline with looking ahead time $T_{\mathrm{la}} = 0.5$ s, driver gain $K_{\mathrm{e}} = 2$, leading time constant $T_{\mathrm{e}} = 0.2$, the mean and variance of the wheel steering rate $\dot{\omega} = 1.3$ rad/s and $\sigma_{\omega}^2 = 0.17^2$. Ray-tracing simulations are conducted at a carrier frequency of $f_{\mathrm{c}} = 73$ GHz with the snapshots captured at $\Delta\tau = 10$ ms intervals until the active vehicle reaches the lane end. We generate 32 trajectories where the vehicles start from randomly selected positions on each lane, each of which contains simulation results of ~ 250 snapshots.

3.4.1 F-MOMP for Channel Tracking

We consider the communication architecture where a $N_{\mathrm{t}}^{\mathrm{x}} \times N_{\mathrm{t}}^{\mathrm{y}} = 16 \times 16$ URA and four $N_{\mathrm{r}}^{\mathrm{x}} \times N_{\mathrm{r}}^{\mathrm{y}} = 12 \times 12$ URAs are deployed at the BS and the vehicle, respectively. In every tracking



(a) Channel gain estimates

(b) ToA estimates

(c) Angle estimates

(d) Angle est. vs. path gains

Figure 3.5. Channel tracking performance comparison. (a)–(c) F-MOMP vs. MOMP with $P_{\mathrm{t}} = 45$ dBm, noise ~ -84 dBm, and $N_{\mathrm{est}} = 5$ estimated paths per channel; (d) angular estimation accuracy vs. path gains using F-MOMP, showing sufficient paths with adequate accuracy for reliable localization.

period, the BS transmits $N_{\mathrm{s}} = 4$ data streams with a length of $Q = 36$ drawn from a Hadamard matrix of order 2^6, with a transmitted power of $P_{\mathrm{t}} = 45$ dBm. A raised-cosine filter with a roll-off factor of 0.4 is used as the pulse shaping function. The system operates at the carrier frequency of $f_{\mathrm{c}} = 73$ GHz, with a bandwidth of $B_{\mathrm{c}} = 1$ GHz. Based on the simulated channel properties and the bandwidth, the number of channel taps is fixed to $N_{\mathrm{d}} = 32$. The analog precoders and combiners are constructed based on the historical channel angle estimates, i.e., the beams point toward the directions aligning with the previously estimated DoAs and DoDs. The vehicle receives $M = 40$ measurements to track channel parameters under the assumption of $N_{\mathrm{est}} = 5$ paths per channel. The resolution for the delay reduced dictionary $\mathbf{\Psi}_1$ is set to 0.25 ns, and the angular reduced dictionaries $\mathbf{\Psi}_k$ ($k = 2, \ldots, 5$) are constructed with a resolution of 0.25°. For all dictionaries, the number of search grids is set to $g_k = 8$. Furthermore, the number of iterations for the MOMP algorithm is set to $N_{\mathrm{iter}} = 4$ to ensure convergence with low computational complexity. The channel tracking results are shown in Fig. 3.5, where the errors are calculated between the estimated paths and their closest counterparts in the true channel. For both F-MOMP and MOMP, the delay errors are below 5 ns for 95% of the situations, and the 95-th percentile values of the angular errors are 7.1°, 3.6°, 2.3°, and 2.8° for the estimated azimuth and elevation AoAs, and azimuth and elevation AoDs, respectively. MOMP performance

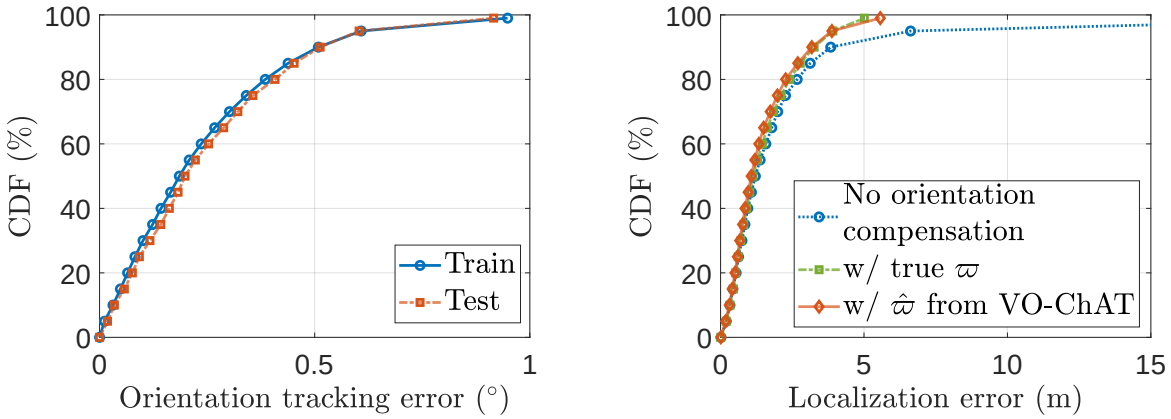


Figure 3.6. (a) VO-ChAT tracking unknown orientation for trajectories in $\mathcal{S}_{\text{tr}}$ and $\mathcal{S}_{\text{te}}$; (b) single-shot geometric localization, where the VO-ChAT predicted $\hat{\boldsymbol{\varpi}}$ achieves performance comparable to using the true $\boldsymbol{\varpi}$.

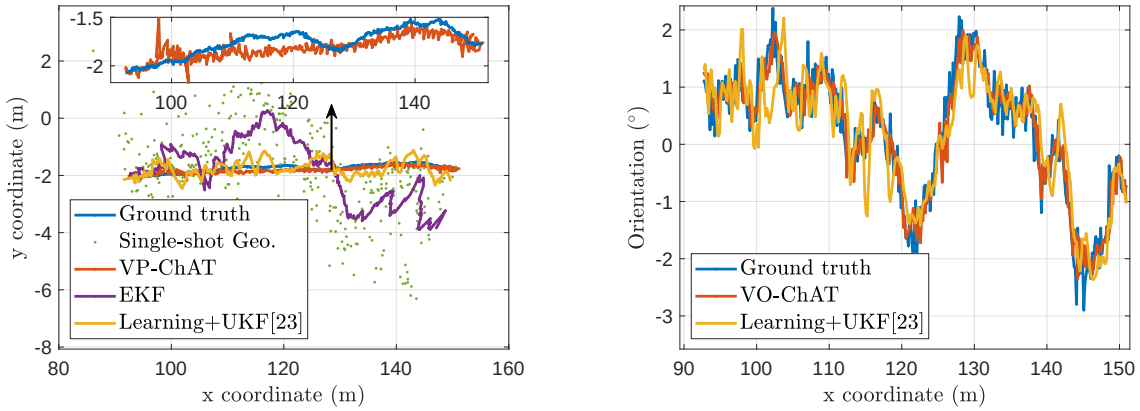
with identical system settings is shown in Fig. 3.5 (a)-(c) for comparison, exhibiting degraded path gain and delay estimation accuracy. Angular estimation results are comparable for azimuth and elevation AoD (90th percentile error $\leq 1.5°$), while F-MOMP demonstrates better azimuth and elevation AoA estimation performance compared to MOMP's 90th percentile error of $\sim 6.5°$. This improvement can stem from F-MOMP's factoring operation enabling broader dictionary grid coverage for capturing true angle values. The estimation for departure angles has higher accuracy due to the larger antenna array employed at the BS. In addition, paths with higher gain magnitude allow higher angle estimation accuracy, as shown in Fig. 3.5d, which motivates us to assign weights proportional to the estimated path gains to prioritize more reliable paths during the localization phase to enhance the accuracy.

3.4.2 VO-ChAT for Orientation Tracking and Single-Shot Localization after Orientation Compensation

We consider the tracking length of $\Gamma = 8$, and the input has a dimension of $8 \times 5 \times 6$. According to the design in Fig. 3.3, there are six MLP modules in VO-ChAT, denoted as MLP_{i} for $\text{i} = 1, ..., 6$: four modules within the ChanSTA component, one before and one after the

concatenation of historical orientation estimates. Each MLP module consists of dense layers with neuron configurations as $(32, 128)$ for MLP_1, $(128, 128)$ for MLP_i, $\text{i} = 2, 3, 4$, $(32, 8, 1)$ for MLP_5, and $(16, 16)$ for MLP_6, where each tuple represents the number of neurons per dense layer. The SiLU activation function [151] is applied after each FC layer to introduce non-linearity into the network. We consider a single head and two heads for the two MHA modules, respectively, and the embedding dimensions are set to $K_{\text{dim}} = Q_{\text{dim}} = 32$ for keys and queries, and $V_{\text{dim}} = 128$ for values, for both the MHA modules. A dropout rate of 0.2 is applied across all dropout layers in the MHA modules. Among the 32 trajectories in the database, VO-ChAT is trained on 24 trajectories, denoted as $\mathcal{S}_{\text{tr}}$ which is split 0.85/0.15 for training and validation, and tested on the other 8 trajectories denoted as $\mathcal{S}_{\text{te}}$. The network training employs MSE loss with the Adam optimizer for 500 epochs, incorporating early stopping based on validation performance to prevent overfitting. The learning rate is 0.001 with a decay rate of 0.95 every 80 epochs. The orientation tracking performance, presented in Fig. 3.6a, demonstrates the network's generalization capability. On the training set, the 50th, 80th, and 95th percentile errors are 0.18°, 0.38°, and 0.61°, respectively. Comparable errors of 0.20°, 0.41°, and 0.60° are observed on the testing set.

After the orientation compensation to retrieve the angle values in the global coordinate system, we acquire the single-shot geometric localization results employing weighted path contributions, where the weight for each path ℓ is computed as $w_\ell = |\hat{\alpha}_\ell| - \min\{|\hat{\alpha}_1|, ..., |\hat{\alpha}_{|\mathfrak{L}|}|\} + \varepsilon_{|\alpha|}$, where $\varepsilon_{|\alpha|} = 2$ is the positive constant that ensures non-zero weights for all paths. As presented in Fig. 3.6b, the localization errors are below 1.06 m, 2.26 m, and 3.88 m for 50%, 80%, and 95% of the cases, respectively. The results are compared to the situation with no orientation compensation and with perfect orientation knowledge, where using the predicted $\hat{\varpi}$ achieves comparable performance to that with the true ϖ, while without orientation compensation the 95-th percentile accuracy is 6.62 m.



(a) **(b)**

Figure 3.7. An example of (a) position tracking performance, and (b) orientation tracking performance based on a trajectory from $\mathcal{S}_{\text{te}}$.

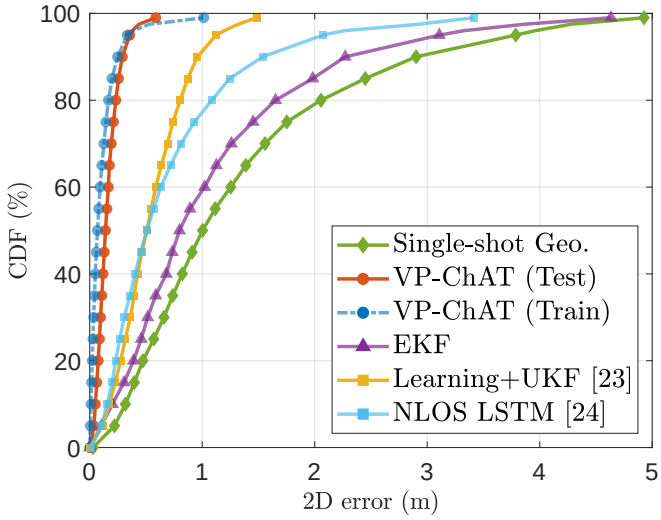


Figure 3.8. CDF of the position tracking error for VP-ChAT and relevant SOTA methods. VP-ChAT significantly outperforms prior work, achieving 0.26 m accuracy for 80% of the test trajectories.

Table 3.4. MLP module specifications for the encoder and decoder of VP-ChAT.

Module	**Encoder**					**Decoder**		
	MLP_1	MLP_2	MLP_3	MLP_4	MLP_5	MLP_1	MLP_2	MLP_3
Layer specification	(32, 128)	(128, 128)	(128, 128)	(128, 128)	(32)	(8)	(32)	(32)

3.4.3 VP-ChAT for Position Tracking

VP-ChAT employs an encoder-decoder architecture where the encoder processes estimated channel information using the same structure as VO-ChAT, i.e., the input channel sequence has a length of $\Gamma = 8$ and its dimension is $8 \times 5 \times 6$, while the decoder analyzes the input position sequence with a length of $\Gamma = 8$ to generate position corrections for the current single-shot position estimate. The encoder ahead of the cross-MHA module comprises five MLP modules with layer configurations specified in Table 3.4, following the same notations defined for VO-ChAT for simplicity. The encoder adopts single-head attention for the first MHA and two-head attention for the second MHA module. Besides, the decoder employs MLP modules with FC layer configurations specified in Table 3.4, processing position evolution information through single-head self-attention. The following single-head cross-attention between the decoder-generated query and encoder-produced value-key pair is implemented considering the inside MLP with a single layer of 32 neurons, and the final MLP module consists of a single layer of 8 neurons. Similar to training VO-ChAT, the Adam optimizer and MSE loss are considered for training on $\mathcal{S}_{\text{tr}}$ for 1000 epochs with early stopping.

An example of tracking performance on a trajectory from $\mathcal{S}_{\text{te}}$ is presented in Fig. 3.7a, where the VP-ChAT tracked positions align with the ground truth with deviations of ≤ 0.47 m, outperforming tracking using EKF and that in [134]. In addition, Fig. 3.8 illustrates the position tracking performance for trajectories within training and testing sets. For the training set, the average tracking accuracy is 0.1 m, with 95th percentile errors at 0.33 m. On the testing set it realizes sub-meter localization across all trajectories, achieving localization errors of 0.15 m, 0.26 m, and 0.36 m at the 50th, 80th, and 95th percentiles, respectively. These results indicate strong generalization capabilities of VP-ChAT, as its performance on unseen test data closely mirrors its performance on the training data. For comparison, we implement an EKF as the baseline, considering the state vector of $[x, y, v_{\text{v}}, \varpi]^{\mathsf{T}}$. The algorithm has a computational complexity of $\mathcal{O}(N^3)$, which is not overly costly for moderate state dimensions. In addition, we reproduce

the algorithm proposed in [134], which considers a similar urban driving scenario, addresses clock offset using RTT, and identifies higher-order reflections via a learning method trained on 3.6 million data samples, followed by vehicle PO tracking using a UKF with a computational complexity of $\mathcal{O}(N^3)$. While [134] assumes idealized channel parameters, our implementation adopts channel estimates obtained through F-MOMP. For the DL-based tracking module, we compared network complexity with [135], which employs an LSTM network for position tracking, as shown in Table 3.5. Although the proposed VP-ChAT network requires 5.4 MFLOPs compared to 2 MFLOPs, it achieves higher accuracy with improving tracking performance from 2.1 m to 0.36 m at the 95th percentile.

Table 3.5. Comparison of model complexity and efficiency, where model size is evaluated by the number of parameters.

Model	Model Size (M)	Memory (MB)	MFLOPs
VP-ChAT (Proposed)	0.23	0.88	5.40
NLOS LSTM [135]	0.30	1.15	2.00

3.5 Collaborative Vehicle Positioning via V2I/V2V Sidelink Exploitation

The single-vehicle framework developed in the preceding sections assumes a persistent V2I LOS link. In practice, vehicles frequently enter NLOS regions in dense urban canyons, where V2I localization performance degrades substantially. Existing collaborative localization methods [152–154] address multi-vehicle scenarios but suffer from several limitations: (1) reliance on measurements from multiple BSs; (2) neglect of unknown inter-vehicle clock offsets, which critically limits ranging accuracy; (3) neglect of antenna array orientation offsets; (4) reliance on simplified 2D models or perfect channel parameter assumptions; and (5) the high complexity of time-domain channel estimation accounting for realistic filtering effects is left unaddressed. This section extends the framework to a multi-vehicle collaborative scenario: vehicles with a LOS to the BS act as *anchors* and are localized through the V2I downlink,

while NLOS vehicles become *targets* whose positions are inferred through V2V sidelinks. The BS coordinates resource allocation and scheduling across the vehicle group, while the anchor vehicles manage the aggregation and processing of collaborative positioning information. This purely model-based geometric framework complements the neural network-based tracking of the previous sections; it explicitly handles per-vehicle clock and orientation offsets using ESPRIT-D channel estimates that account for realistic filtering effects, making it robust for vehicles that intermittently lose V2I coverage.

3.5.1 System Model Extension for Multi-Vehicle Scenarios

We consider a group of three vehicles sharing the same roadway, of which at least one has a LOS channel to the BS at any given time frame. The V2I channel for vehicle u is as in (3.6), parameterized by its per-vehicle clock offset $t_{\text{off,u}}^{(\tau)}$ and array orientation offset $\Delta\varpi_u^{(\tau)}$ relative to the global coordinate frame. When the targets enter NLOS, the V2V sidelink is activated. The sidelink channel from vehicle u_i to vehicle u_k at frame τ is modeled as

$$\begin{aligned}\dot{\mathbf{H}}_n^{(\tau)} = \sum_{\ell=1}^{\dot{L}} \dot{\alpha}_\ell^{(\tau)} f_{\text{p}}\Big(nT_{\text{s}} - \Big(\dot{t}_\ell^{(\tau)} - (t_{\text{off,u}_\text{k}}^{(\tau)} - t_{\text{off,u}_\text{i}}^{(\tau)})\Big)\Big) \cdot \\ \mathbf{a}\Big(\dot{\theta}_{\text{az},\ell}^{(\tau)} - \Delta\varpi_{u_k}^{(\tau)}, \dot{\theta}_{\text{el},\ell}^{(\tau)}\Big)\, \mathbf{a}\Big(\dot{\phi}_{\text{az},\ell}^{(\tau)} - \Delta\varpi_{u_i}^{(\tau)}, \dot{\phi}_{\text{el},\ell}^{(\tau)}\Big)^*,\end{aligned} \tag{3.35}$$

where dots denote sidelink channel parameters ($\dot{\alpha}_\ell^{(\tau)}$, $\dot{t}_\ell^{(\tau)}$, $\dot{\theta}_\ell^{(\tau)}$, $\dot{\phi}_\ell^{(\tau)}$), $t_{\text{off,u}_\text{k}}^{(\tau)} - t_{\text{off,u}_\text{i}}^{(\tau)}$ is the inter-vehicle relative clock offset, and $\Delta\varpi_{u_k}^{(\tau)}$ and $\Delta\varpi_{u_i}^{(\tau)}$ are the orientation offsets at the receiver and transmitter, respectively. The sidelink is modeled as LOS due to the close proximity among vehicles within a group. The received sidelink signal follows the same form as (3.8):

$$\dot{\mathbf{y}}_{\tau,m}[q] = \dot{\mathbf{W}}_{\tau,m}^* \sum_{n=0}^{\dot{N}_{\text{d}}-1} \sqrt{\dot{P}}\dot{\mathbf{H}}_n^{(\tau)} \dot{\mathbf{F}}_{\tau,m}\mathbf{s}[q-n] + \dot{\mathbf{W}}_{\tau,m}^* \dot{\mathbf{n}}_{\tau,m}[q], \tag{3.36}$$

where $\dot{P}$ is the V2V transmit power, $\dot{N}_{\text{d}}$ is the number of sidelink channel taps, $\dot{\mathbf{F}}_{\tau,m}$ and $\dot{\mathbf{W}}_{\tau,m}$ are the V2V precoder and combiner designed based on previous channel angle estimates, and $\dot{\mathbf{n}}_{\tau,m}[q]$

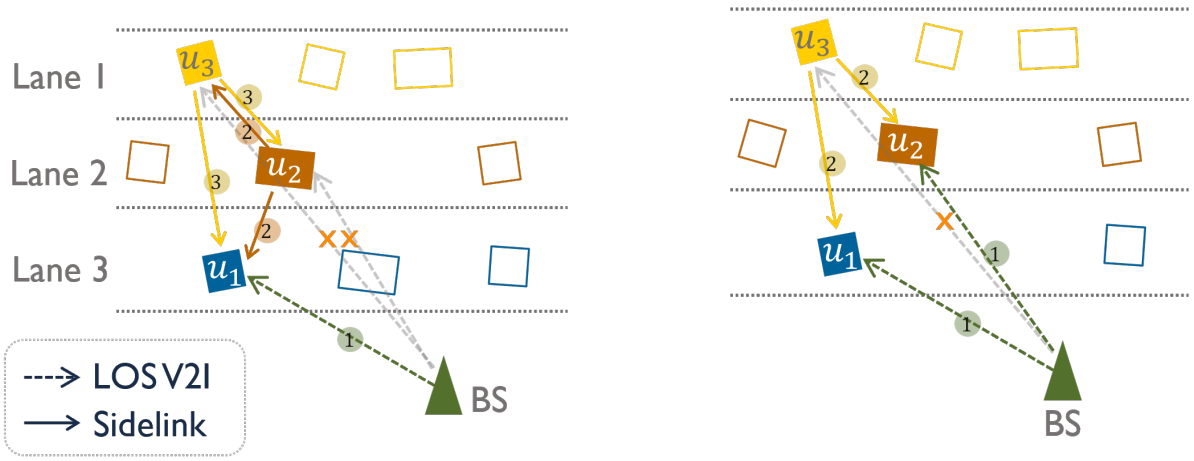


Figure 3.9. Sky view of collaborative positioning: (a) single-anchor scenario (u_1 localized via V2I, u_2 and u_3 as NLOS targets); (b) multi-anchor scenario (u_1 and u_2 as anchors, u_3 as target). Circled numbers indicate the ordering of link parameter estimation.

is the additive white Gaussian noise on the sidelink. The BS coordinates scheduling across the vehicle group; the anchor vehicles aggregate the positioning information for processing.

3.5.2 Collaborative Localization Framework

Sidelink channel parameters are estimated using ESPRIT-D (Sec. 2.3.2 in Chapter 2), adapted to the V2V setting. This algorithm accounts for the system filtering effects and provides high-resolution AoA, AoD, and delay estimates for the sidelink LOS path. The beams are steered toward the previous angle estimates to reduce pilot overhead. As the localization geometry is time-stationary within each estimation interval, the time index τ is omitted in the following for clarity.

Single-Anchor Scenario

Let u_1 be the anchor (V2I LOS) and u_2, u_3 be the targets (Fig. 3.9a). Each target transmits pilots in turn to the anchor and to the other target. For a V2V link from u_i to u_k, the estimated

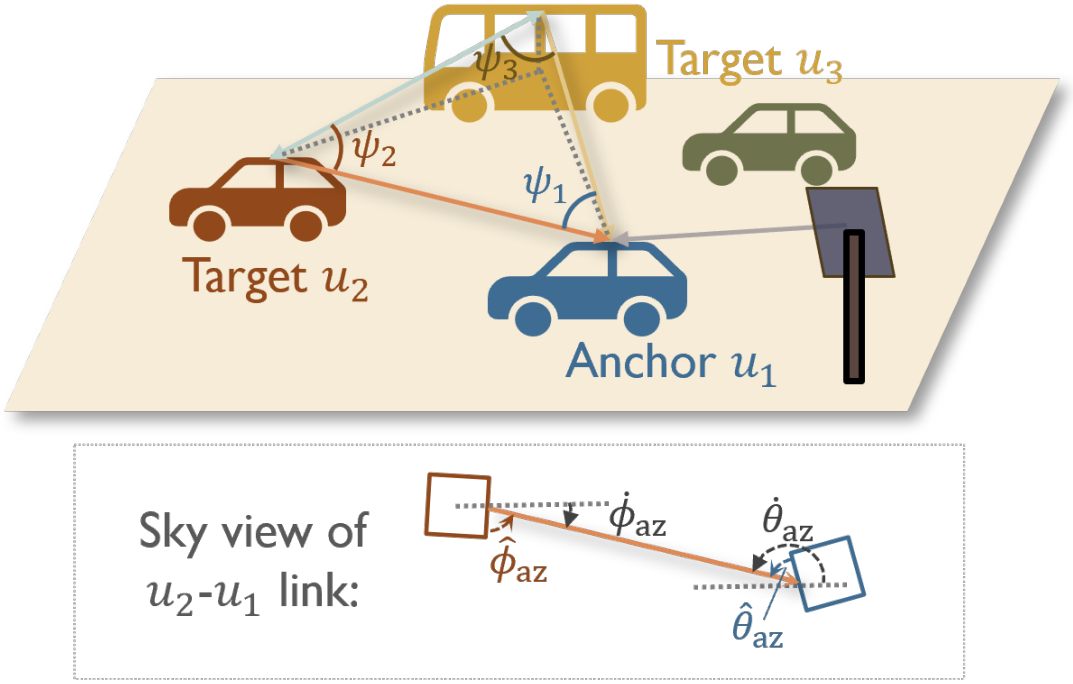


Figure 3.10. Illustration of the V2V sidelink channel geometry exploited for collaborative localization.

AoA and AoD at the receiver satisfy, for the LOS path:

$$\begin{cases} \cos(\hat{\phi}_{\text{az},i} + \Delta\varpi_{u_i}) = -\cos(\hat{\theta}_{\text{az},k} + \Delta\varpi_{u_k}) \\ \sin(\hat{\phi}_{\text{az},i} + \Delta\varpi_{u_i}) = -\sin(\hat{\theta}_{\text{az},k} + \Delta\varpi_{u_k}) \end{cases}, \tag{3.37}$$

reflecting the geometric anti-parallel relationship between AoD and AoA for a LOS path. Let $\Delta\varpi_{ik} = \Delta\varpi_{u_i} - \Delta\varpi_{u_k}$; then (3.37) yields the linear system

$$\underbrace{\begin{bmatrix} \cos\hat{\phi}_{\text{az},i} & -\sin\hat{\phi}_{\text{az},i} \\ \sin\hat{\phi}_{\text{az},i} & \cos\hat{\phi}_{\text{az},i} \end{bmatrix}}_{\mathbf{A}_{\text{ori}}} \begin{bmatrix} \cos\Delta\varpi_{ik} \\ \sin\Delta\varpi_{ik} \end{bmatrix} = \begin{bmatrix} -\cos\hat{\theta}_{\text{az},k} \\ -\sin\hat{\theta}_{\text{az},k} \end{bmatrix}, \tag{3.38}$$

from which $\Delta\varpi_{ik}$ is obtained via LS estimation. Relative orientation offsets $\Delta\varpi_{23}$, $\Delta\varpi_{21}$, and $\Delta\varpi_{31}$ are estimated from three distinct transmissions. The absolute offsets $\Delta\varpi_{u_2}$ and $\Delta\varpi_{u_3}$ are

then recovered using the known $\Delta\varpi_{u_1}$ (from the V2I AoD) by solving the over-determined system

$$\begin{bmatrix} 1 & -1 \\ 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \Delta\varpi_{u_2} \\ \Delta\varpi_{u_3} \end{bmatrix} = \begin{bmatrix} \Delta\varpi_{23} \\ \Delta\varpi_{21} + \Delta\varpi_{u_1} \\ \Delta\varpi_{31} + \Delta\varpi_{u_1} \end{bmatrix}. \tag{3.39}$$

With all angles expressed in the global coordinate frame, the inter-vehicle distance $\hat{d}_{23}$ is obtained from the RTT between u_2 and u_3: $\hat{d}_{23} = c(\hat{t}_{23} + \hat{t}_{32})/2$. The law of sines applied to the triangle formed by the three vehicles gives

$$\frac{(\hat{t}_{21} + t_{\text{off,u}_1} - t_{\text{off,u}_2}) \cdot c}{\sin\psi_3} = \frac{(\hat{t}_{31} + t_{\text{off,u}_1} - t_{\text{off,u}_3}) \cdot c}{\sin\psi_2} = \frac{\hat{d}_{23}}{\sin\psi_1}, \tag{3.40}$$

where $\psi_i = \cos^{-1}(\langle \boldsymbol{\phi}_{i1}, \boldsymbol{\phi}_{ik} \rangle)$ for $i,k \in \{2,3\}$, $i \neq k$, and $\psi_1 = \cos^{-1}(\langle \boldsymbol{\theta}_{21}, \boldsymbol{\theta}_{31} \rangle)$, with $\boldsymbol{\phi}_{ik}$ and $\boldsymbol{\theta}_{ik}$ being the unit DoD and DoA direction vectors. This yields the linear system $\mathbf{Ax} = \mathbf{b}$ for $\mathbf{x} = \left[t_{\text{off,u}_2}, t_{\text{off,u}_3}\right]^{\mathsf{T}}$:

$$\mathbf{A} = \begin{bmatrix} \mathbf{I}_2 \\ \sin\psi_2 \;\; -\sin\psi_3 \end{bmatrix}; \; \mathbf{b} = \begin{bmatrix} \hat{t}_{21} + t_{\text{off,u}_1} - \frac{\hat{d}_{23}\sin\psi_2}{c\sin\psi_1} \\ \hat{t}_{31} + t_{\text{off,u}_1} - \frac{\hat{d}_{23}\sin\psi_3}{c\sin\psi_1} \\ \sum_{i=2}^{3}(-1)^i \sin\psi_i (\hat{t}_{i1} + t_{\text{off,u}_1}) \end{bmatrix}. \tag{3.41}$$

Clock offset estimates are averaged over the trajectory to reduce the impact of channel estimation errors. Finally, the target positions are obtained by geometric propagation from the anchor:

$$\hat{\mathbf{p}}_{u_i} = \hat{\mathbf{p}}_{u_1} + \boldsymbol{\theta}_{i1} \cdot c \left(\hat{t}_{i1} + t_{\text{off,u}_1} - t_{\text{off,u}_i}\right), \quad i = 2,3. \tag{3.42}$$

Multi-Anchor Scenario

When two vehicles (u_1 and u_2) are in LOS and act as anchors (Fig. 3.9b), the target u_3 transmits pilots to both. Each anchor independently estimates the V2V channel parameters. The relative orientation offsets $\Delta\varpi_{31}$ and $\Delta\varpi_{32}$ are computed from (3.38) using the respective anchor

receptions, and the absolute orientation offset of the target is obtained by averaging:

$$\Delta\varpi_{u_3} = \tfrac{1}{2}\left(\Delta\varpi_{31} + \Delta\varpi_{u_1} + \Delta\varpi_{32} + \Delta\varpi_{u_2}\right). \tag{3.43}$$

The clock offset $t_{\mathrm{off,u_3}}$ is derived analogously to (3.40) using the two anchor delays and their known clock offsets:

$$t_{\mathrm{off,u_3}} = \frac{\sum_{i=1}^{2}(-1)^2 \sin\psi_i (\hat{t}_{3i} - t_{\mathrm{off,u_i}})}{\sin\psi_1 - \sin\psi_2}. \tag{3.44}$$

Two independent position estimates for u_3 are computed via (3.42) from each anchor, and their average is taken as the final estimate.

Kalman Filter for Temporal Tracking

A KF is applied independently per vehicle to improve positioning accuracy across frames, particularly during brief NLOS periods. The state of vehicle u_i at frame τ is $\mathbf{s}_{u_i}^{(\tau)} = \left[x_{u_i}^{(\tau)},\, v_{\mathrm{x},u_i}^{(\tau)},\, y_{u_i}^{(\tau)},\, v_{\mathrm{y},u_i}^{(\tau)}\right]^{\mathsf{T}}$, where $v_{\mathrm{x},u_i}^{(\tau)} \sim \mathcal{N}(\dot{v}_{\mathrm{x},u_i}, \sigma_{\mathrm{x},u_i}^2)$ and $v_{\mathrm{y},u_i}^{(\tau)} \sim \mathcal{N}(\dot{v}_{\mathrm{y},u_i}, \sigma_{\mathrm{y},u_i}^2)$ are the velocity components, modeled as Gaussian with vehicle-specific mean and variance [155]. The measurement vector is the 2D position estimate $\mathbf{z}_{u_i}^{(\tau)} = [\hat{\mathbf{p}}_{u_i}^{(\tau)}]_{1:2}$. During fully NLOS periods where no sidelink observation is available, the KF relies on the state transition function alone, sustaining coverage across the entire trajectory.

3.5.3 Simulation Results

The simulation environment is the same urban canyon described in Sec. 3.4. Three vehicles — two cars (1.6 m height) and one truck (3.8 m height) — move at speeds of 20, 40, and 60 mph. For the V2V links, the BS array is 24×24 and vehicle arrays are 8×8; transmit power is 10 dBm for V2V and 40 dBm for V2I. Pilot length $Q = 38$, snapshot interval 25 ms; raised cosine pulse shaping with roll-off 0.6. Ten trajectory sets with different starting positions are used, each containing approximately 150 snapshots.

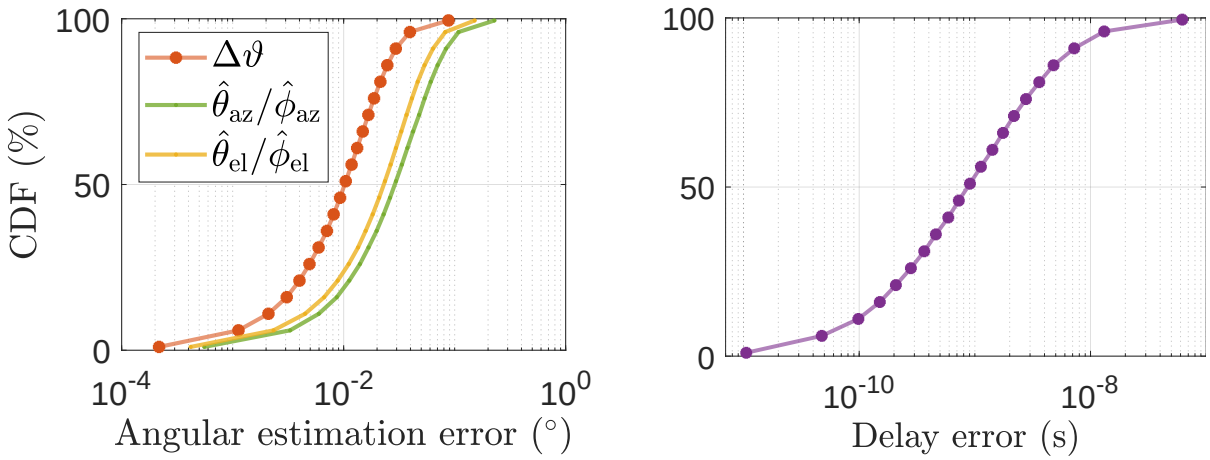


Figure 3.11. Sidelink channel estimation results: (a) CDF of angular estimation errors for 3D AoA/AoD and orientation offsets; (b) CDF of delay estimation errors after clock offset compensation.

Fig. 3.11 shows the sidelink parameter estimation accuracy. ESPRIT-D achieves angular errors below 0.1° for 95% of cases. Orientation offset estimation errors remain consistently below 0.05° due to the geometric constraints in (3.38). Clock offset estimation errors stay below 10 ns, and the compensated delay errors are likewise within 10 ns, confirming the accuracy of the ranging stage.

Fig. 3.12 shows an example trajectory for one vehicle group. Without V2V collaboration, V2I coverage reaches only 98%, 45%, and 18% of time frames for u_1, u_2, and u_3, respectively. After introducing collaborative sidelink positioning, all vehicles achieve 98% coverage, with the remaining 2% relying on the KF state transition. The NLOS periods (gray regions) exhibit no localization accuracy penalty compared to the LOS periods.

Fig. 3.13 presents the CDF of 2D localization error across all testing trajectories. Single-snapshot collaborative positioning achieves an average error of 0.4 m for NLOS targets, with errors below 1.2 m for 90% of cases. After applying the KF, the average improves to 0.3 m, with sub-meter accuracy for 90% of cases. V2I NLOS localization [121] — which requires ≥ 3 first-order reflections — achieves only 1.2 m average, highlighting the advantage of the

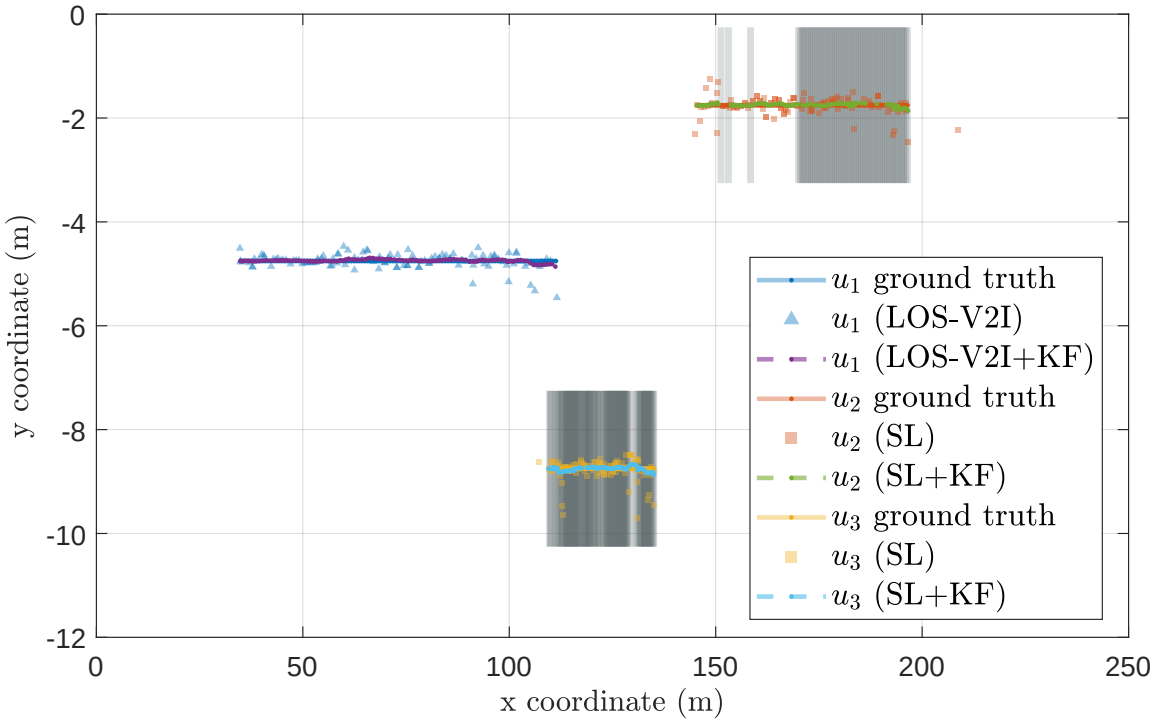


Figure 3.12. Sky view of collaborative positioning results for a group of three vehicles on three lanes. Gray regions indicate NLOS periods where V2I coverage is lost and V2V sidelinks are activated. Sub-meter positioning is maintained throughout the trajectory.

Table 3.6. Comparison of cooperative localization schemes in urban driving scenarios. The proposed geometric-model approach outperforms all baselines, with over 4× improvement in average accuracy over GNSS-based methods.

Ref.	Data type	Localization scheme	Avg. error (m)	90%-ile error (m)
[152]	V2I/SL-AoA + V2I/SL-ToA	Factor Graph + BP	1.2	1.5
[153]	GNSS + SL-ToA	Factor Graph + BP	2.9	3.5
[154]	GNSS + SL-AoA	Graph Laplacian with LMS	1.4	2.5
Proposed	V2I/SL-AoA + V2I/SL-ToA	Collaborative geometric loc. + KF	**0.3**	**0.7**

collaborative V2V approach.

Table 3.6 compares the proposed method against representative V2X cooperative localization schemes. Methods exploiting mmWave V2I/V2V channels [152] outperform GNSS-based solutions [153, 154]. The proposed approach achieves 1 m improvement in average accuracy over [152] and a 50% improvement at the 90th percentile. This is primarily attributable to two

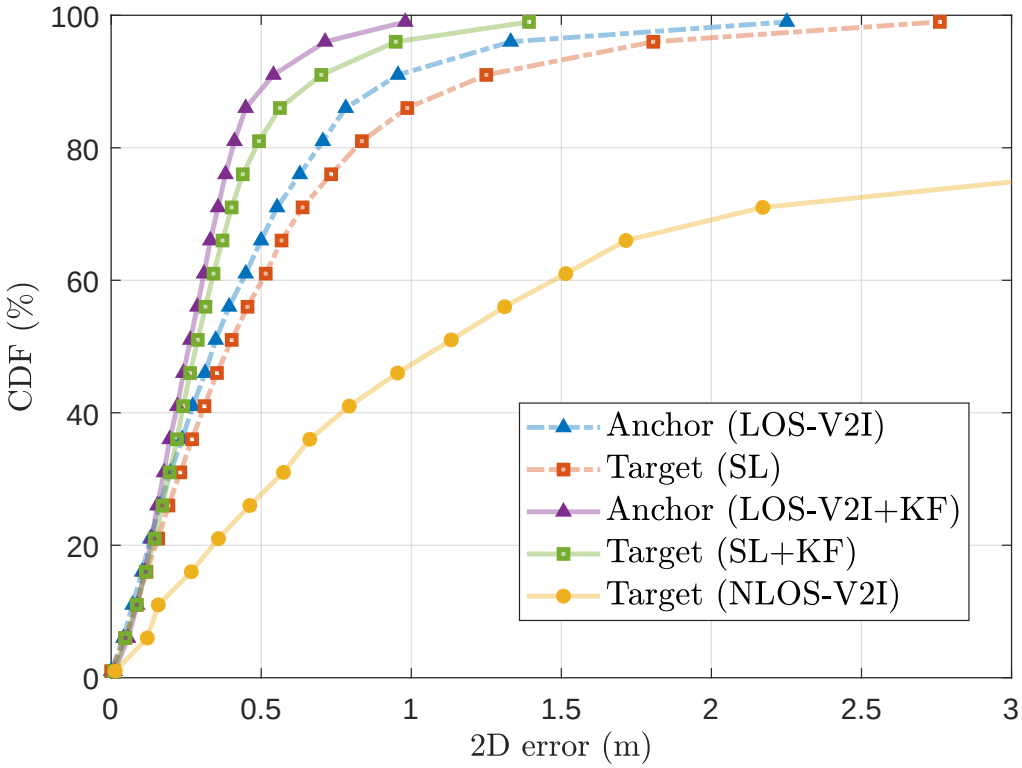


Figure 3.13. CDF of the 2D localization error (m) for collaborative positioning of NLOS targets, with and without Kalman filtering, compared to V2I NLOS localization.

innovations: (1) explicit model-based estimation of per-vehicle clock offsets, which is neglected by all baselines, and (2) use of high-resolution mmWave channel estimates from ESPRIT-D accounting for realistic filtering effects, as opposed to simplified or ideal channel parameters assumed in prior work.

The model-based collaborative framework is complementary to the data-driven tracking networks of the previous sections. While VP-ChAT provides precise single-vehicle position corrections through learned channel-trajectory correlations, the collaborative extension handles multi-vehicle NLOS scenarios where V2I information is unavailable, relying on interpretable geometric transformations without any training data requirement. Together, the two approaches form a complete positioning framework covering both single-vehicle and multi-vehicle operating conditions in realistic urban deployments.

3.6 Conclusion

We developed a hybrid model/data-driven framework for mmWave communication channel tracking and precise vehicle PO tracking in urban environments, adopting realistic system models that account for factors often neglected in prior studies. First, we introduced a low-complexity time domain channel tracking algorithm, F-MOMP, to accurately estimate multipath parameters with delay and angular errors below 0.1 ns and 2° for 80% of cases, sufficiently supporting vehicle localization. Then, VO-ChAT, employing an attention mechanism to process channel estimate sequences, tracks the vehicle's orientation with errors below 0.5° in 80% of cases. Thereafter, we formulated a WLS problem using the selected LOS and first-order channel paths to realize single-shot localization. Finally, VP-ChAT, built upon the Transformer architecture, leverages the channel and position estimation sequence to provide the correction for the single-shot position estimate, achieving the tracking accuracy of 15 cm and 26 cm at the 50th and 80th percentiles, respectively.

We further extended the framework to multi-vehicle collaborative positioning, where a model-based geometric approach enables NLOS vehicles to be localized via V2V sidelinks. By explicitly estimating per-vehicle clock offsets and orientation offsets from sidelink channel parameters and applying a Kalman filter for temporal smoothing, the proposed collaborative framework achieves an average 2D localization error of 0.3 m for NLOS targets, with sub-meter accuracy for 90% of cases — outperforming existing cooperative localization schemes by a wide margin.

The results demonstrate that the hybrid model/data-driven approaches for precise vehicle tracking with mmWave channel estimates in complex urban environments are effective, with deep learning modules integrated when model-based methods exhibit limitations. The network architecture integrates two intuitive principles: 1) a Transformer-based architecture to capture long-range temporal dependencies, and 2) decoupled branches for channel parameter estimation and pose regression to mitigate mutual interference. In contrast to billion-parameter language models

[156], our lightweight framework contains only 0.23 M trainable parameters, demonstrating data efficiency.

Several extensions follow naturally from the framework. The collaborative positioning solver decomposes per V2V pair, so a fully distributed implementation in which each vehicle exchanges sidelink measurements only with its neighbors and updates its local estimate via consensus or distributed primal/dual iterations is the natural multi-agent counterpart of the centralized scheme presented here. The current evaluation reports observability empirically through CDFs across many ray-tracing scenarios; the local geometric solve is non-observable in the degenerate case where collaborating vehicles are colinear with the NLOS target, which motivates a formal observability and stability analysis as a control-theoretic extension. From a communication-resource perspective, since each pair only exchanges tens of bytes per update, an event- or self-triggered messaging policy where a vehicle transmits only when its local innovation exceeds a threshold is a principled way to handle bandwidth and intermittent V2V links without sacrificing tracking accuracy. Finally, the F-MOMP tracking layer carries an implicit per-frame data association through warm-started reduced dictionaries; a multipath-SLAM-style factor graph [116] that persistently maps scatterers across frames and treats path-to-path association probabilistically would directly improve performance in NLOS-anchor cases and is a clean bridge between the proposed pipeline and message-passing localization frameworks.

Acknowledgements

Chapter 3, in part, is a reprint, with permission, of the material as it appears in the papers: *[Best Student Paper]* Y. Chen, N. González-Prelcic, T. Shimizu, H. Lu, C. Mahabal, "Sparse Recovery with Attention: A Hybrid Data/Model Driven Solution for High Accuracy Position and Channel Tracking at mmWave", in the 2023 IEEE 24th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC 2023), Shanghai, China, Sep., 2023; Y. Chen, N. González-Prelcic, T. Shimizu, C. Mahabal, "V2X-Enabled Collaborative Orientation Estimation and Position Tracking in mmWave Networks", 2024 58th Asilomar Conference on Signals, Systems, and Computers. IEEE, 2024, Oct., 2024; and Y. Chen, N. González-Prelcic,

T. Shimizu, C. Mahabal, “A Hybrid Model/Data-Driven Solution to Channel, Position and Orientation Tracking in mmWave Vehicular Systems.”, IEEE J. Sel. Areas Commun., vol. 44, pp. 927–941, Mar. 2026, doi: 10.1109/JSAC.2025.3612354. The dissertation author was the primary investigator and author of these papers, which were partially supported by the NSF under grant no. 2433782, no. 2147955, and in part by funds from the federal agency and industry partners as specified in the Resilient & Intelligent NextG Systems (RINGS) program, and by a gift from Toyota Motor North America, Inc., USA.

Chapter 4

Conclusion and Future Work

This dissertation presents a comprehensive hybrid model/data-driven framework to address fundamental challenges in mmWave vehicular networks operating within the ISAC paradigm. By strategically integrating domain-specific signal processing models with deep learning architectures, this work demonstrates that careful co-design of sensing and communication functions can overcome the limitations of purely model-based or purely data-driven approaches, achieving robust performance in realistic propagation environments with reduced training overhead and maintained interpretability.

4.1 Summary of Contributions

Radar-Aided Multiuser Link Configuration. In Chapter 1, we addressed the critical challenge of beam training overhead in dynamic multiuser vehicular environments. By exploiting the spatial congruence between automotive radar and communication channels, we developed a passive radar-aided framework wherein the RSU opportunistically senses existing FMCW transmissions to configure mmWave downlink links. A key contribution is the mixing filter bank architecture with CFAR detection, which reliably isolates individual vehicle radar signals from multiuser FMCW interference without prior knowledge of each vehicle's chirp parameters. Three distinct DNN architectures (mapping radar covariances to communication APS, eigenvectors, and covariance vectors) effectively learn and compensate for frequency and geometry mismatches

between the 76 GHz radar and 73 GHz communication bands. The covariance-prediction-assisted strategy achieves up to 32$\times$ reduction in beam search space (from 64 to 2 beams at the RSU) and 21.9% rate improvement over radar-only baselines, demonstrating that out-of-band spatial information can be effectively leveraged for rapid link establishment in both LOS and NLOS conditions.

Learning to Localize with Attention. Chapter 2 tackled the complementary problem of high-accuracy vehicle localization from single-snapshot channel observations. Recognizing that pure geometric methods suffer from error propagation when high-order reflections are misclassified, while pure learning methods lack robustness to unseen environments, we proposed a hybrid pipeline integrating low-complexity joint channel estimation and localization with attention-based location estimate refinement. The two-stage MOMP and ESPRIT-D algorithms provide high-resolution 3D channel parameter estimates while accounting for realistic filtering effects and unknown clock offsets. *PathNet* introduces a physics-informed classification strategy to identify geometrically useful paths, while *ChanFormer* employs cross-attention mechanisms to evaluate path reliability and spatial congruence, refining initial geometric estimates. This architecture achieves 28 cm accuracy for 80% of users in LOS and sub-meter accuracy for 55% of users in NLOS, representing an order-of-magnitude improvement over conventional fingerprinting and maximum likelihood approaches.

Channel, Position and Orientation Tracking. In Chapter 3, we extended these snapshot-based solutions to continuous tracking scenarios, addressing the practical challenges of unknown vehicle orientations and temporal channel variations. The *F-MOMP* algorithm exploits temporal correlation and factored dictionary structures to enable efficient high-resolution channel tracking with 0.1 ns delay accuracy. For orientation-agnostic scenarios, *VO-ChAT* leverages spatial-temporal attention over channel sequences to track vehicle headings with errors below 0.5°, while *VP-ChAT* employs cross-attention between channel evolution and trajectory history to refine single-shot position estimates. The collaborative positioning framework further demonstrates that V2V sidelinks can effectively localize NLOS vehicles through model-based geometric

transformations and Kalman filtering, achieving 0.3 m average localization error for NLOS targets with sub-meter accuracy for 90% of cases.

Overall, this work demonstrates that combining physical-layer knowledge with data-driven deep learning provides a systematic and effective methodology for meeting the low-latency, high-reliability, and high-accuracy requirements of 6G vehicular ISAC systems.

4.2 Future Work

Several directions naturally extend the contributions of this dissertation. On the algorithmic side, the radar-aided framework can be extended from initial access to continuous beam tracking by exploiting the temporal consistency of FMCW covariance estimates across frames, and generalized to support automotive radar waveforms beyond FMCW, such as OFDM-based radar. The single-BS localization approach can be improved for NLOS conditions by fusing channel parameters from multiple RSUs and by utilizing higher-order reflections as additional geometric anchors, rather than discarding them as in the current *PathNet* design. Extending these channel estimation and localization methods to the upper mid-band (FR3, 7–24 GHz), where 6G deployments are expected, is also a natural next step, potentially leveraging cross-frequency information from available mmWave observations to aid localization where FR3 measurements are sparse.

From a deployment perspective, the collaborative positioning framework should be scaled to large, dynamic vehicular networks where the roles of anchors and targets change over time based on V2I and V2V link availability. A distributed implementation, where each vehicle estimates its own position using only local neighbor measurements, would be more practical and scalable than a centralized solution. In parallel, running the proposed algorithms on automotive-grade embedded hardware requires exploring model compression and quantization to meet the latency and power constraints of onboard processors.

From a network design perspective, the attention-based architectures proposed in this

dissertation could be extended to output uncertainty estimates alongside position and orientation predictions, for example by incorporating Bayesian layers or conformal prediction heads. Such uncertainty-aware designs would be directly useful when the localization output feeds a downstream safety-critical controller, allowing the vehicle system to act more conservatively when confidence is low. More broadly, incorporating probabilistic graphical models into the hybrid framework, where geometric constraints and learned components interact through structured belief propagation, could yield a more principled and interpretable integration of the model-based and data-driven layers.

Finally, all results in this dissertation are based on ray-tracing simulations. Validating the proposed methods through over-the-air experiments with real automotive radars and mmWave hardware would be an important step toward practical deployment and would produce publicly available datasets for the broader vehicular ISAC research community.

Bibliography

[1] N. González-Prelcic, D. Tagliaferri, M. F. Keskin, H. Wymeersch, and L. Song, "Six Integration Avenues for ISAC in 6G and Beyond," *IEEE Vehicular Technology Magazine*, vol. 20, no. 1, pp. 18–39, 2025.

[2] Z. Wei, H. Qu, Y. Wang, X. Yuan, H. Wu, Y. Du, K. Han, N. Zhang, and Z. Feng, "Integrated sensing and communication signals toward 5G-A and 6G: A survey," *IEEE Internet of Things Journal*, vol. 10, no. 13, pp. 11 068–11 092, 2023.

[3] 3GPP, "Release 18 Description; Summary of Rel-18 Work Items," 3rd Generation Partnership Project (3GPP), Technical report (TR) 21.918, Apr., 2025, version 18.0.0. [Online]. Available: https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=4229

[4] Z. Du, F. Liu, Y. Li, W. Yuan, Y. Cui, Z. Zhang, C. Masouros, and B. Ai, "Toward ISAC-Empowered Vehicular Networks: Framework, Advances, and Opportunities," *IEEE Transactions on Wireless Communications*, vol. 32, no. 2, pp. 222–229, 2025.

[5] 3GPP, "Service requirements for enhanced V2X scenarios," 3rd Generation Partnership Project (3GPP), Technical Specification (TS) 22.186, Apr., 2024, version 18.0.1. [Online]. Available: https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=3180

[6] F. Pedraza and G. Caire, "Sensing-assisted beam tracking for mmWave V2I communications with analog, hybrid, and digital antenna architectures," *IEEE Transactions on Wireless Communications*, 2024.

[7] A. Ali, N. González-Prelcic, and A. Ghosh, "Passive radar at the roadside unit to configure millimeter wave vehicle-to-infrastructure links," *IEEE Trans. on Vehicular Technology*, vol. 69, no. 12, pp. 14 903–14 917, 2020.

[8] Z. Ye, C. Yu, H. Zhu, Y. He, M. Gao, and G. Yu, "ISAC-assisted collision avoidance mechanism for vehicle-to-infrastructure systems," *IEEE Transactions on Intelligent Vehicles*, 2024.

[9] M. Giordani, M. Polese, A. Roy, D. Castor, and M. Zorzi, "A tutorial on beam management for 3gpp nr at mmwave frequencies," *IEEE Communications Surveys & Tutorials*, vol. 21, no. 1, pp. 173–196, 2018.

[10] F. Dong, F. Liu, S. Lu, Y. Xiong, Q. Zhang, Z. Feng, and F. Gao, "Communication-assisted sensing in 6G networks," *IEEE Journal on Selected Areas in Communications*, 2025.

[11] K. V. Mishra, M. B. Shankar, N. González-Prelcic, M. Valkama, W. Yu, and B. Ottersten, "Editorial introduction to the special issue on learning-based signal processing for integrated sensing and communications," *IEEE Journal of Selected Topics in Signal Processing*, vol. 18, no. 5, pp. 731–736, 2024.

[12] N. Wu, R. Jiang, X. Wang, L. Yang, K. Zhang, W. Yi, and A. Nallanathan, "Ai-enhanced integrated sensing and communications: Advancements, challenges, and prospects," *IEEE Communications Magazine*, vol. 62, no. 9, pp. 144–150, 2024.

[13] Y. Li, F. Liu, Z. Du, W. Yuan, Q. Shi, and C. Masouros, "Frame structure and protocol design for sensing-assisted NR-V2X communications," *IEEE Transactions on Mobile Computing*, 2024.

[14] X. Luo, Q. Lin, R. Zhang, H.-H. Chen, X. Wang, and M. Huang, "ISAC–a survey on its layered architecture, technologies, standardizations, prototypes and testbeds," *IEEE Communications Surveys Tutorials*, 2025.

[15] M. Bayraktar, N. González-Prelcic, H. Chen, and C. J. Zhang, "Truly full-duplex integrated sensing and single-user communication at mmWave," in *2024 IEEE 25th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, 2024, pp. 116–120.

[16] N. González-Prelcic, A. Ali, V. Va, and R. W. Heath, "Millimeter-Wave Communication with Out-of-Band Information," *IEEE Communications Magazine*, vol. 55, no. 12, pp. 140–146, 2017.

[17] WiSeCom Lab., "Radar-Comm. Channel Data Set," 2022. [Online]. Available: https://github.com/WiSeCom-Lab/DL-based-radar-aided-mmWave-V2X

[18] M. Hashemi, C. E. Koksal, and N. B. Shroff, "Out-of-band millimeter wave beamforming and communications to achieve low latency and high energy efficiency in 5G systems," *IEEE Transactions on Communications*, vol. 66, no. 2, pp. 875–888, 2018.

[19] A. Ali, N. González-Prelcic, and R. W. Heath, "Millimeter wave beam-selection using out-of-band spatial information," *IEEE Transactions on Wireless Communications*, vol. 17, no. 2, pp. 1038–1052, 2018.

[20] A. Ali, N. González-Prelcic, and R. W. Heath, "Spatial covariance estimation for millimeter wave hybrid systems using out-of-band information," *IEEE Transactions on Wireless Communications*, vol. 18, no. 12, pp. 5471–5485, 2019, (early access).

[21] N. González-Prelcic, R. Méndez-Rial, and R. W. Heath, "Radar aided beam alignment in mmWave V2I communications supporting antenna diversity," in *2016 Information Theory and Applications Workshop (ITA)*, 2016, pp. 1–7.

[22] F. Liu, W. Yuan, C. Masouros, and J. Yuan, "Radar-assisted predictive beamforming for vehicular links: communication served by sensing," *IEEE Transactions on Wireless Communications*, vol. 19, no. 11, pp. 7704–7719, 2020.

[23] C. Aydogdu, F. Liu, C. Masouros, H. Wymeersch, and M. Rydström, "Distributed radar-aided vehicle-to-vehicle communication," in *2020 IEEE Radar Conference (RadarConf20)*, 2020, pp. 1–6.

[24] A. Klautau, N. González-Prelcic, and R. W. Heath, "LIDAR data for deep learning-based mmWave beam-selection," *IEEE Wireless Communications Letters*, vol. 8, no. 3, pp. 909–912, 2019.

[25] T. Woodford, X. Zhang, E. Chai, K. Sundaresan, and A. Khojastepour, "Spacebeam: Lidar-driven one-shot mmwave beam management," in *Proceedings of the 19th Annual International Conference on Mobile Systems, Applications, and Services*, ser. MobiSys '21. New York, NY, USA: Association for Computing Machinery, 2021, pp. 389–401. [Online]. Available: https://doi.org/10.1145/3458864.3466864

[26] M. Brambilla, M. Nicoli, S. Savaresi, and U. Spagnolini, "Inertial sensor aided mmWave beam tracking to support cooperative autonomous driving," in *2019 IEEE International Conference on Communications Workshops (ICC Workshops)*, 2019, pp. 1–6.

[27] A. Ali, J. Mo, B. L. Ng, V. Va, and J. C. Zhang, "Orientation-assisted beam management for beyond 5g systems," *IEEE Access*, vol. 9, pp. 51 832–51 846, 2021.

[28] P. Kela, M. Costa, J. Turkka, M. Koivisto, J. Werner, A. Hakkarainen, M. Valkama, R. Jantti, and K. Leppanen, "Location based beamforming in 5G ultra-dense networks," in *2016 IEEE 84th Vehicular Technology Conference (VTC-Fall)*, 2016, pp. 1–7.

[29] W. Miao, C. Luo, G. Min, L. Wu, T. Zhao, and Y. Mi, "Position-based beamforming design for UAV communications in LTE networks," in *ICC 2019 - 2019 IEEE International Conference on Communications (ICC)*, 2019, pp. 1–6.

[30] T.-H. Chou, N. Michelusi, D. J. Love, and J. V. Krogmeier, "Fast position-aided mimo beam training via noisy tensor completion," *IEEE Journal of Selected Topics in Signal Processing*, vol. 15, no. 3, pp. 774–788, 2021.

[31] V. Va, T. Shimizu, G. Bansal, and R. W. Heath, "Online learning for position-aided millimeter wave beam training," *IEEE Access*, vol. 7, pp. 30 507–30 526, 2019.

[32] V. Va, J. Choi, T. Shimizu, G. Bansal, and R. W. Heath, "Inverse multipath fingerprinting for millimeter wave v2i beam alignment," *IEEE Transactions on Vehicular Technology*, vol. 67, no. 5, pp. 4042–4058, 2017.

[33] K. Satyanarayana, M. El-Hajjar, A. A. M. Mourad, and L. Hanzo, "Deep learning aided fingerprint-based beam alignment for mmWave vehicular communication," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 11, pp. 10 858–10 871, 2019.

[34] Y. Wang, A. Klautau, M. Ribero, A. C. K. Soong, and R. W. Heath, "MmWave vehicular beam selection with situational awareness using machine learning," *IEEE Access*, vol. 7, pp. 87 479–87 493, 2019.

[35] G. H. Sim, S. Klos, A. Asadi, A. Klein, and M. Hollick, "An online context-aware machine learning algorithm for 5G mmWave vehicular communications," *IEEE/ACM Transactions on Networking*, vol. 26, no. 6, pp. 2487–2500, 2018.

[36] M. B. Mashhadi, M. Jankowski, T.-Y. Tung, S. Kobus, and D. Gündüz, "Federated mmwave beam selection utilizing lidar data," *IEEE Wireless Communications Letters*, vol. 10, no. 10, pp. 2269–2273, 2021.

[37] A. Ali, N. González-Prelcic, and A. Ghosh, "Millimeter wave V2I beam-training using base-station mounted radar," in *2019 IEEE Radar Conference (RadarConf)*. IEEE, 2019, pp. 1–5.

[38] A. Graff, A. Ali, and N. González-Prelcic, "Measuring radar and communication congruence at millimeter wave frequencies," in *53rd Asilomar Conference on Signals, Systems, and Computers*, 2019, pp. 925–929.

[39] Y. Chen, A. Graff, N. González-Prelcic, and T. Shimizu, "Radar aided mmWave vehicle-to-infrastructure link configuration using deep learning," in *2021 IEEE Global Communications Conference (GLOBECOM)*, 2021, pp. 01–06.

[40] 3GPP, "Study on evaluation methodology of new Vehicle-to-Everything (V2X) use cases for LTE and NR," 3rd Generation Partnership Project (3GPP), Technical report (TR) 37.885, Jun., 2019, version 15.3.0. [Online]. Available: https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=3209

[41] A. Alkhateeb and R. W. Heath Jr., "Frequency selective hybrid precoding for limited feedback millimeter wave systems," *IEEE Transactions on Communications*, vol. 64, no. 5, pp. 1801–1818, May 2016.

[42] E. Bjornson, D. Hammarwall, and B. Ottersten, "Exploiting quantized channel norm feedback through conditional statistics in arbitrarily correlated MIMO systems," *IEEE Transactions on Signal Processing*, vol. 57, no. 10, pp. 4027–4041, 2009.

[43] V. Katkovnik, M.-S. Lee, and Y.-H. Kim, "High-resolution signal processing for a switch antenna array FMCW radar with a single channel receiver," 2002, pp. 543–547.

[44] A. Graff, A. Ali, N. González-Prelcic, and A. Ghosh, "Automotive radar and mmwave mimo v2x communications: Interference or fruitful coexistence?" in *2020 IEEE Radar Conference (RadarConf20)*, 2020, pp. 1–5.

[45] A. Alkhateeb and R. W. Heath, "Frequency selective hybrid precoding for limited feedback millimeter wave systems," *IEEE Transactions on Communications*, vol. 64, no. 5, pp. 1801–1818, 2016.

[46] A. Alkhateeb, O. El Ayach, G. Leus, and R. W. Heath Jr., "Channel estimation and hybrid precoding for millimeter wave cellular systems," *IEEE Journal of Selected Topics in Signal Processing*, vol. 8, no. 5, pp. 831–846, Oct. 2014.

[47] R. Cumplido, C. Torres, and S. Lopez, "On the implementation of an efficient FPGA-based CFAR processor for target detection," in *(ICEEE). 1st International Conference on Electrical and Electronics Engineering, 2004.*, 2004, pp. 214–218.

[48] B. Xu, N. Wang, T. Chen, and M. Li, "Empirical evaluation of rectified activations in convolutional network," *arXiv preprint arXiv:1505.00853*, 2015.

[49] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 12, pp. 2481–2495, 2017.

[50] "Wireless Insite," http://www.remcom.com/wireless-insite.

[51] I. T. U. (ITU), "Effects of building materials and structures on radiowave propagation above about 100 MHz," Tech. Rep. ITU-R P.2040-1, Jul. 2015.

[52] E. S. Li and K. Sarabandi, "Low grazing incidence millimeter-wave scattering models and measurements for various road surfaces," *IEEE Transactions on Antennas and Propagation*, vol. 47, no. 5, pp. 851–861, 1999.

[53] Remcom, "5G mmwave channel modeling with diffuse scattering in an office environment." [Online]. Available: https://www.remcom.com/examples/2017/6/22/5g-mmwave-channel-modeling-with-diffuse-scattering-in-an-office-environment

[54] N. González-Prelcic, M. Furkan Keskin, O. Kaltiokallio, M. Valkama, D. Dardari, X. Shen, Y. Shen, M. Bayraktar, and H. Wymeersch, "The integrated sensing and communication revolution for 6G: Vision, techniques, and applications," *Proceedings of the IEEE*, vol. 112, no. 7, pp. 676–723, 2024.

[55] A. Shahmansoori, G. E. Garcia, G. Destino, G. Seco-Granados, and H. Wymeersch, "Position and orientation estimation through millimeter-wave MIMO in 5G systems," *IEEE Trans. on Wireless Commun/*, vol. 17, no. 3, pp. 1822–1835, 2018.

[56] T. Wild, V. Braun, and H. Viswanathan, "Joint design of communication and sensing for beyond 5G and 6G systems," *IEEE Access*, vol. 9, pp. 30 845–30 857, 2021.

[57] Q. Tao, Z. Hu, Z. Zhou, H. Xiao, and J. Zhang, "SeqPolar: Sequence matching of polarized lidar map with HMM for intelligent vehicle localization," *IEEE Trans. on Vehicular Technology*, 2022.

[58] K. Ćwian, M. R. Nowicki, and P. Skrzypczyński, "GNSS-augmented LiDAR SLAM for accurate vehicle localization in large scale urban environments," in *2022 17th Intl. Conf. on Control, Automation, Robotics and Vision (ICARCV)*. IEEE, 2022, pp. 701–708.

[59] M.-S. Kang, J.-H. Ahn, J.-U. Im, and J.-H. Won, "Lidar-and V2X-based cooperative localization technique for autonomous driving in a GNSS-denied environment," *Remote Sensing*, vol. 14, no. 22, p. 5881, 2022.

[60] A. Schaefer, D. Büscher, J. Vertens, L. Luft, and W. Burgard, "Long-term vehicle localization in urban environments based on pole landmarks extracted from 3-d lidar scans," *Robotics and Autonomous Systems*, vol. 136, p. 103709, 2021.

[61] Y. Liang, S. Müller, D. Schwendner, D. Rolle, D. Ganesch, and I. Schaffer, "A scalable framework for robust vehicle state estimation with a fusion of a low-cost IMU, the GNSS, radar, a camera and lidar," in *2020 IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS)*. IEEE, 2020, pp. 1661–1668.

[62] R. Keating, A. Ghosh, B. Velgaard, D. Michalopoulos, and M. Säily, "The evolution of 5G New Radio positioning technologies," Nokia Bell Labs, Tech. Rep., 02 2021.

[63] F. Gómez-Cuba, G. Feijoo-Rodríguez, and N. González-Prelcic, "Clock and orientation-robust simultaneous radio localization and mapping at millimeter wave bands," in *2023 IEEE Wireless Communications and Networking Conference (WCNC)*, 2023, pp. 1–7.

[64] H. Wymeersch, N. Garcia, H. Kim, G. Seco-Granados, S. Kim, F. Went, and M. Fröhle, "5G mmWave downlink vehicular positioning," in *IEEE Global Commun. Conf. (GLOBECOM)*, 2018, pp. 206–212.

[65] J. Talvitie, M. Koivisto, T. Levanen, M. Valkama, G. Destino, and H. Wymeersch, "High-accuracy joint position and orientation estimation in sparse 5G mmWave channel," in *IEEE Intl. Conf. on Commun. (ICC)*, 2019, pp. 1–7.

[66] F. Wen, J. Kulmer, K. Witrisal, and H. Wymeersch, "5G positioning and mapping with diffuse multipath," *IEEE Trans. on Wireless Communications*, vol. 20, no. 2, pp. 1164–1174, 2020.

[67] F. Jiang, Y. Ge, M. Zhu, H. Wymeersch, and F. Tufvesson, "Low-complexity channel estimation and localization with random beamspace observations," in *ICC 2023-IEEE Intl. Conf. on Communications*. IEEE, 2023, pp. 5985–5990.

[68] F. Jiang, Y. Ge, M. Zhu, and H. Wymeersch, "High-dimensional channel estimation for simultaneous localization and communications," in *2021 IEEE Wireless Communications and Networking Conf. (WCNC)*, 2021, pp. 1–6.

[69] H. Chen, F. Jiang, Y. Ge, H. Kim, and H. Wymeersch, "Doppler-enabled single-antenna localization and mapping without synchronization," in *GLOBECOM 2022-2022 IEEE Global Communications Conf.* IEEE, 2022, pp. 6469–6474.

[70] J. Li, M. F. Da Costa, and U. Mitra, "Joint localization and orientation estimation in millimeter-wave MIMO OFDM systems via atomic norm minimization," *IEEE Trans. on Signal Processing*, vol. 70, pp. 4252–4264, 2022.

[71] Y. Guan, K. Xu, and X. Cheng, "Accurate wideband localization in massive MIMO systems with low-resolution ADCs," *IEEE Trans. on Vehicular Technology*, 2023.

[72] M. A. Nazari, G. Seco-Granados, P. Johannisson, and H. Wymeersch, "MmWave 6D radio localization with a snapshot observation from a single BS," *IEEE Trans. on Vehicular Technology*, 2023.

[73] Z. Gong, X. S. Shen, C. Li, Y. Song, and R. Su, "High-accuracy positioning services for high-speed vehicles in wideband mmWave communications," *IEEE Transactions on Signal Processing*, vol. 71, pp. 3867–3882, 2023.

[74] A. Fascista, A. Coluccia, H. Wymeersch, and G. Seco-Granados, "Downlink single-snapshot localization and mapping with a single-antenna receiver," *IEEE Trans. on Wireless Communications*, vol. 20, no. 7, pp. 4672–4684, 2021.

[75] W. Xu, Y. Xiao, A. Liu, M. Lei, and M.-J. Zhao, "Joint scattering environment sensing and channel estimation based on non-stationary Markov random field," *IEEE Trans. on Wireless Communications*, 2023.

[76] Y. Chen, J. Palacios, N. González-Prelcic, T. Shimizu, and H. Lu, "Joint initial access and localization in millimeter wave vehicular networks: a hybrid model/data driven approach," in *2022 IEEE 12th Sensor Array and Multichannel Signal Processing Workshop (SAM)*, 2022, pp. 355–359.

[77] X. Wu, G. Yang, F. Hou, and S. Ma, "Low-complexity downlink channel estimation for millimeter-wave FDD massive MIMO systems," *IEEE Wireless Communications Letters*, 2019.

[78] A. C. Gurbuz, Y. Yapici, and I. Guvenc, "Sparse channel estimation in millimeter-wave communications via parameter perturbed OMP," in *2018 IEEE Intl. Conf. on Communications Workshops (ICC Workshops)*. IEEE, 2018, pp. 1–6.

[79] A. Mejri, M. Hajjaj, and S. Hasnaoui, "Structured analysis/synthesis compressive sensing-based channel estimation in wideband mmWave large-scale multiple input multiple output systems," *Trans. on Emerging Telecommunications Technologies*, vol. 31, no. 7, p. e3903, 2020.

[80] J. Rodríguez-Fernández, N. González-Prelcic, K. Venugopal, and R. W. Heath, "Frequency-domain compressive channel estimation for frequency-selective hybrid millimeter wave MIMO systems," *IEEE Transactions on Wireless Communications*, vol. 17, no. 5, pp. 2946–2960, 2018.

[81] J. P. González-Coma, J. Rodríguez-Fernández, N. González-Prelcic, L. Castedo, and R. W. Heath, "Channel estimation and hybrid precoding for frequency selective multiuser mmWave MIMO systems," *IEEE Journal of Selected Topics in Signal Processing*, vol. 12, no. 2, pp. 353–367, 2018.

[82] Y. Zhang, M. El-Hajjar, and L.-l. Yang, "Multi-layer sparse Bayesian learning for mmWave channel estimation," *IEEE Trans. on Vehicular Technology*, 2023.

[83] K. Venugopal, A. Alkhateeb, N. Prelcic-González, and R. W. Heath, "Channel estimation for hybrid architecture-based wideband millimeter wave systems," *IEEE Journal on Selected Areas in Communications*, vol. 35, no. 9, pp. 1996–2009, 2017.

[84] J. A. Tropp, A. C. Gilbert, and M. J. Strauss, "Simultaneous sparse approximation via greedy pursuit," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 5, Mar. 2005, pp. 721–724.

[85] J. Palacios, N. González-Prelcic, and C. Rusu, "Multidimensional orthogonal matching pursuit: theory and application to high accuracy joint localization and communication at mmWave," *arXiv preprint arXiv:2208.11600*, 2022.

[86] Y. You and L. Zhang, "Off-grid compressive sensing based channel estimation with non-uniform grid in millimeter wave MIMO system," in *2022 16th European Conf. on Antennas and Propagation (EuCAP)*. IEEE, 2022, pp. 1–5.

[87] K.-H. Liu, W. Zhang, and L. Wan, "Fast convex method for off-grid millimeter-wave/sub-terahertz channel estimation via exploiting joint sparse structure," in *ICC 2022-IEEE Intl. Conf. on Communications*. IEEE, 2022, pp. 919–925.

[88] Y. Yan, C. Shan, J. Zhang, and H. Zhao, "Off-grid channel estimation for OTFS-based mmWave hybrid beamforming systems," *IEEE Communications Letters*, 2023.

[89] B. Qi, W. Wang, and B. Wang, "Off-grid compressive channel estimation for mm-Wave massive MIMO with hybrid precoding," *IEEE Communications Letters*, vol. 23, no. 1, pp. 108–111, 2019.

[90] C. K. Anjinappa, Y. Zhou, Y. Yapici, D. Baron, and I. Guvenc, "Channel estimation in mmWave hybrid MIMO system via off-grid dirichlet kernels," in *2019 IEEE Global Communications Conf. (GLOBECOM)*, 2019, pp. 1–6.

[91] C. K. Anjinappa, A. C. Gürbüz, Y. Yapıcı, and I. Güvenç, "Off-grid aware channel and covariance estimation in mmwave networks," *IEEE Trans. on Communications*, vol. 68, no. 6, pp. 3908–3921, 2020.

[92] Y. You, C. Zhang, and L. Zhang, "Bayesian matching pursuit based estimation of off-grid channel for millimeter wave massive MIMO system," *IEEE Trans. on Vehicular Technology*, vol. 71, no. 11, pp. 11 603–11 614, 2022.

[93] J. Rodriguez-Fernandez, N. González-Prelcic, and R. W. Heath, "A compressive sensing-maximum likelihood approach for off-grid wideband channel estimation at mmWave," in *2017 IEEE 7th Intl. Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*, 2017, pp. 1–5.

[94] F. Wen, H. C. So, and H. Wymeersch, "Tensor decomposition-based beamspace esprit algorithm for multidimensional harmonic retrieval," in *ICASSP 2020-2020 IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 4572–4576.

[95] J. Zhang, D. Rakhimov, and M. Haardt, "Gridless channel estimation for hybrid mmwave mimo systems via tensor-esprit algorithms in dft beamspace," *IEEE Journal of Selected Topics in Signal Processing*, vol. 15, no. 3, pp. 816–831, 2021.

[96] J. Gao, C. Zhong, G. Y. Li, J. B. Soriaga, and A. Behboodi, "Deep learning-based channel estimation for wideband hybrid mmWave massive MIMO," *IEEE Trans. on Communications*, 2023.

[97] S. Liu and X. Huang, "Sparsity-aware channel estimation for mmWave massive MIMO: A deep CNN-based approach," *China Communications*, vol. 18, no. 6, pp. 162–171, 2021.

[98] W. Ma, C. Qi, Z. Zhang, and J. Cheng, "Sparse channel estimation and hybrid precoding using deep learning for millimeter wave massive MIMO," *IEEE Trans. on Communications*, vol. 68, no. 5, pp. 2838–2849, 2020.

[99] J. Gao, D. Wu, F. Yin, Q. Kong, L. Xu, and S. Cui, "MetaLoc: Learning to learn wireless localization," *IEEE Journal on Selected Areas in Communications*, 2023.

[100] A. Salihu, S. Schwarz, and M. Rupp, "Attention aided CSI wireless localization," in *2022 IEEE 23rd Intl. Workshop on Signal Processing Advances in Wireless Communication (SPAWC)*. IEEE, 2022, pp. 1–5.

[101] N. Lv, F. Wen, Y. Chen, and Z. Wang, "A deep learning-based end-to-end algorithm for 5g positioning," *IEEE Sensors Letters*, vol. 6, no. 4, pp. 1–4, 2022.

[102] J. Gante, G. Falcao, and L. Sousa, "Deep learning architectures for accurate millimeter wave positioning in 5G," *Neural Processing Letters*, vol. 51, no. 1, pp. 487–514, 2020.

[103] X. Wang, M. Patil, C. Yang, S. Mao, and P. A. Patel, "Deep convolutional Gaussian processes for mmwave outdoor localization," in *ICASSP 2021-2021 IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 8323–8327.

[104] C. Wu, X. Yi, W. Wang, L. You, Q. Huang, X. Gao, and Q. Liu, "Learning to localize: A 3D CNN approach to user positioning in massive MIMO-OFDM systems," *IEEE Trans. on Wireless Communications*, vol. 20, no. 7, pp. 4556–4570, 2021.

[105] J. Palacios, N. González-Prelcic, and C. Rusu, "Low complexity joint position and channel estimation at millimeter wave based on multidimensional orthogonal matching pursuit," in *2022 30th European Signal Processing Conference (EUSIPCO)*, 2022, pp. 1002–1006.

[106] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[107] Y. Chen, "Learning to Localize with Attention: from sparse mmWave channel estimates from a single BS to high accuracy 3D location," Jan. 2024. [Online]. Available: https://github.com/WiSeCom-Lab/ChanFormer.git

[108] R. Méndez-Rial, C. Rusu, N. González-Prelcic, A. Alkhateeb, and R. W. Heath, "Hybrid MIMO architectures for millimeter wave communications: Phase shifters or switches?" *IEEE access*, vol. 4, pp. 247–267, 2016.

[109] K. Venugopal, A. Alkhateeb, N. González Prelcic, and R. W. Heath, "Channel estimation for hybrid architecture-based wideband millimeter wave systems," *IEEE Journal on Selected Areas in Communications*, vol. 35, no. 9, pp. 1996–2009, 2017.

[110] I. Kisil, G. G. Calvi, B. Scalzo Dees, and D. P. Mandic, "Tensor decompositions and practical applications: A hands-on tutorial," in *Recent Trends in Learning From Data: Tutorials from the INNS Big Data and Deep Learning Conference (INNSBDDL2019)*. Springer, 2020, pp. 69–97.

[111] X. Gong, W. Chen, L. Sun, J. Chen, and B. Ai, "An ESPRIT-based supervised channel estimation method using tensor train decomposition for mmWave 3-D MIMO-OFDM systems," *IEEE Transactions on Signal Processing*, vol. 71, pp. 555–570, 2023.

[112] P. Zhu, H. Lin, J. Li, D. Wang, and X. You, "High-performance channel estimation for mmWave wideband systems with hybrid structures," *IEEE Transactions on Communications*, vol. 71, no. 4, pp. 2503–2516, 2023.

[113] Y. Chen, "ESPRIT-D," Dec. 2023. [Online]. Available: https://github.com/WiSeCom-Lab/ESPRIT-D.git

[114] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[115] A. Shahmansoori, G. E. Garcia, G. Destino, G. Seco-Granados, and H. Wymeersch, "Position and orientation estimation through millimeter-wave mimo in 5G systems," *IEEE Transactions on Wireless Communications*, vol. 17, no. 3, pp. 1822–1835, 2018.

[116] E. Leitinger, A. Venus, B. Teague, and F. Meyer, "Data fusion for multipath-based SLAM: Combining information from multiple propagation paths," *IEEE Transactions on Signal Processing*, vol. 71, pp. 4011–4028, 2023.

[117] J. Palacios, N. González-Prelcic, and C. Rusu, "Multidimensional orthogonal matching pursuit: theory and application to high accuracy joint localization and communication at mmWave," 2022. [Online]. Available: https://arxiv.org/abs/2208.11600

[118] Y. Wang, W. Wang, Y. Wu, J. Liu, Q. Zhang, J. Wang, and W. Fan, "USRP-based multifrequency multiscenario channel measurements and modeling for 5G campus internet of things," *IEEE Internet of Things Journal*, vol. 11, no. 8, pp. 13 865–13 883, 2024.

[119] D. Shakya, M. Ying, T. S. Rappaport, P. Ma, I. Al-Wazani, Y. Wu, Y. Wang, D. Calin, H. Poddar, A. Bazzi, M. Chafii, Y. Xing, and A. Ghosh, "Urban outdoor propagation measurements and channel models at 6.75 GHz FR1(C) and 16.95 GHz FR3 upper mid-band spectrum for 5G and 6G," 2024. [Online]. Available: https://arxiv.org/abs/2410.17539

[120] Y. Chen, J. Palacios, N. González-Prelcic, T. Shimizu, and H. Lu, "Joint initial access and localization in millimeter wave vehicular networks: a hybrid model/data driven approach," in *2022 IEEE 12th Sensor Array and Multichannel Signal Processing Workshop (SAM)*, 2022, pp. 355–359.

[121] Y. Chen, N. González-Prelcic, T. Shimizu, and H. Lu, "Learning to localize with attention: From sparse mmWave channel estimates from a single BS to high accuracy 3D location," *IEEE Transactions on Wireless Communications*, 2024, (under revision).

[122] B. Or, N. Segol, A. Eweida, and M. Freydin, "Learning position from vehicle vibration using an inertial measurement unit," *IEEE Transactions on Intelligent Transportation Systems*, vol. 25, no. 9, pp. 10 766–10 776, 2024.

[123] Y. Ma, T. Wang, X. Bai, H. Yang, Y. Hou, Y. Wang, Y. Qiao, R. Yang, and X. Zhu, "Vision-centric BEV perception: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 12, pp. 10 978–10 997, 2024.

[124] R. Xu, C.-J. Chen, Z. Tu, and M.-H. Yang, "V2X-ViTv2: Improved vision transformers for vehicle-to-everything cooperative perception," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 47, no. 1, pp. 650–662, 2025.

[125] D. Lee, M. Jung, W. Yang, and A. Kim, "LiDAR odometry survey: recent advancements and remaining challenges," *Intelligent Service Robotics*, vol. 17, no. 2, pp. 95–118, 2024.

[126] N. J. Abu-Alrub and N. A. Rawashdeh, "Radar odometry for autonomous ground vehicles: A survey of methods and datasets," *IEEE Transactions on Intelligent Vehicles*, vol. 9, no. 3, pp. 4275–4291, 2024.

[127] G. Wu, F. Zhou, K. Kit Wong, and X.-Y. Li, "A vehicle-mounted radar-vision system for precisely positioning clustering d," *IEEE Journal on Selected Areas in Communications*, vol. 42, no. 10, pp. 2688–2703, 2024.

[128] H. A. Hashim, A. E. E. Eltoukhy, and K. G. Vamvoudakis, "UWB ranging and IMU data fusion: Overview and nonlinear stochastic filter for inertial navigation," *IEEE Transactions on Intelligent Transportation Systems*, vol. 25, no. 1, pp. 359–369, 2024.

[129] Y. Li, Q. Zhao, J. Guo, R. Fang, and J. Zheng, "Multisensor fusion for railway irregularity inspection system: Integration of RTK GNSS, MEMS IMU, odometer, and laser," *IEEE Transactions on Instrumentation and Measurement*, vol. 73, pp. 1–12, 2024.

[130] J. Talvitie, M. Säily, and M. Valkama, "Orientation and location tracking of XR devices: 5G carrier phase-based methods," *IEEE Journal of Selected Topics in Signal Processing*, vol. 17, no. 5, pp. 919–934, 2023.

[131] M. Koivisto, J. Talvitie, E. Rastorgueva-Foi, Y. Lu, and M. Valkama, "Channel parameter estimation and TX positioning with multi-beam fusion in 5G mmWave networks," *IEEE Transactions on Wireless Communications*, vol. 21, no. 5, pp. 3192–3207, 2022.

[132] B. Zhao, K. Hu, F. Wen, S. Cui, and Y. Shen, "TDLoc: Passive localization for MIMO-OFDM system via tensor decomposition," *IEEE Internet of Things Journal*, vol. 10, no. 23, pp. 20 819–20 833, 2023.

[133] T. Wang, Y. Li, J. Liu, K. Hu, and Y. Shen, "Multipath-assisted single-anchor localization via deep variational learning," *IEEE Transactions on Wireless Communications*, vol. 23, no. 8, pp. 9113–9128, 2024.

[134] Q. Bader, S. Saleh, M. Elhabiby, and A. Noureldin, "Leveraging single-bounce reflections and onboard motion sensors for enhanced 5G positioning," *IEEE Transactions on Intelligent Transportation Systems*, vol. 25, no. 12, pp. 20 464–20 477, 2024.

[135] R. Klus, J. Talvitie, J. Equi, G. Fodor, J. Torsner, and M. Valkama, "Robust NLoS localization in 5G mmWave networks: Data-based methods and performance," *IEEE Transactions on Vehicular Technology*, pp. 1–16, 2024.

[136] M. Shamsesalehi, M. A. Attari, M. A. M. Sadr, B. Champagne, and M. Qaraqe, "A BFF-based attention mechanism for trajectory estimation in mmWave MIMO communications," in *2024 IEEE Wireless Communications and Networking Conference (WCNC)*, 2024, pp. 1–6.

[137] A. Venus, E. Leitinger, S. Tertinek, F. Meyer, and K. Witrisal, "Graph-based simultaneous localization and bias tracking," *IEEE Transactions on Wireless Communications*, vol. 23, no. 10, pp. 13 141–13 158, 2024.

[138] J. Gao, J. Fan, S. Zhai, and G. Dai, "Message passing based wireless multipath SLAM with continuous measurements correction," *IEEE Transactions on Signal Processing*, vol. 72, pp. 1691–1705, 2024.

[139] H. Que, J. Yang, C.-K. Wen, S. Xia, X. Li, and S. Jin, "Joint beam management and SLAM for mmWave communication systems," *IEEE Transactions on Communications*, vol. 71, no. 10, pp. 6162–6179, 2023.

[140] J. Yang, C.-K. Wen, J. Xu, H. Que, H. Wei, and S. Jin, "Angle-based SLAM on 5G mmWave systems: Design, implementation, and measurement," *IEEE Internet of Things Journal*, vol. 10, no. 20, pp. 17 755–17 771, 2023.

[141] Y. Ge, O. Kaltiokallio, H. Kim, F. Jiang, J. Talvitie, M. Valkama, L. Svensson, S. Kim, and H. Wymeersch, "A computationally efficient EK-PMBM filter for bistatic mmWave

radio SLAM," *IEEE Journal on Selected Areas in Communications*, vol. 40, no. 7, pp. 2179–2192, 2022.

[142] T. Du, J. Yang, C.-K. Wen, S. Xia, and S. Jin, "General simultaneous localization and mapping scheme for mmWave communication systems," *IEEE Internet of Things Journal*, vol. 11, no. 12, pp. 22 521–22 536, 2024.

[143] O. Kaltiokallio, J. Talvitie, E. Rastorgueva-Foi, H. Wymeersch, and M. Valkama, "Integrated snapshot and filtering-based bistatic radio SLAM in mmWave networks," in *2024 IEEE 25th International Workshop Signal Process. Adv. Wirel. Commun. (SPAWC)*, 2024, pp. 301–305.

[144] Y. Chen, "F-MOMP," Dec. 2024. [Online]. Available: https://github.com/WiSeCom-Lab/F-MOMP.git

[145] ——, "Hybrid Model Data-Driven mmWave PO Tracking," Jul. 2025. [Online]. Available: https://github.com/WiSeCom-Lab/Hybrid-Model-Data-Driven-mmWave-PO-Tracking.git

[146] Y. Chen, N. González-Prelcic, T. Shimizu, H. Lu, and C. Mahabal, "Sparse recovery with attention: A hybrid data/model driven solution for high accuracy position and channel tracking at mmWave," in *2023 IEEE 24th International Workshop Signal Process. Adv. Wirel. Commun. (SPAWC)*, 2023, pp. 491–495.

[147] K. van der El, D. M. Pool, M. M. van Paassen, and M. Mulder, "Modeling driver steering behavior in restricted-preview boundary-avoidance tasks," *Transportation Research Part F: Traffic Psychology and Behaviour*, vol. 94, pp. 362–378, 2023.

[148] A. de Santana Correia and E. L. Colombini, "Attention, please! A survey of neural attention models in deep learning," *Artificial Intelligence Review*, vol. 55, no. 8, pp. 6037–6124, 2022.

[149] A. Zeng, M. Chen, L. Zhang, and Q. Xu, "Are transformers effective for time series forecasting?" in *Proceedings of the AAAI conference on artificial intelligence*, vol. 37, no. 9, 2023, pp. 11 121–11 128.

[150] A. Ali, N. González-Prelcic, and A. Ghosh, "Passive radar at the roadside unit to configure millimeter wave vehicle-to-infrastructure links," *IEEE Transactions on Vehicular Technology*, vol. 69, no. 12, pp. 14 903–14 917, 2020.

[151] S. Elfwing, E. Uchibe, and K. Doya, "Sigmoid-weighted linear units for neural network function approximation in reinforcement learning," *Neural Networks*, vol. 107, pp. 3–11, 2018, special issue on deep reinforcement learning.

[152] W. Meng, X. Chu, Z. Lu, L. Wang, X. Wen, and M. Li, "V2V communication assisted cooperative localization for connected vehicles," in *2021 IEEE Wireless Communications and Networking Conference (WCNC)*. IEEE, 2021, pp. 1–6.

[153] J. Xiong, Z. Xiong, X. Xie, Y. Zhuang, Y. Zheng, S. Xiong, J. W. Cheong, and A. G. Dempster, "Integrity for belief propagation-based cooperative positioning," *IEEE Transactions on Intelligent Transportation Systems*, 2024.

[154] N. Piperigkos, A. S. Lalos, and K. Berberidis, "Graph laplacian diffusion localization of connected and automated vehicles," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 8, pp. 12 176–12 190, 2022.

[155] R. Borsche, A. Klar, and M. Zanella, "Kinetic-controlled hydrodynamics for multilane traffic models," *Physica A: Statistical Mechanics and its Applications*, vol. 587, p. 126486, 2022.

[156] W. X. Zhao, K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou, Y. Min, B. Zhang, J. Zhang, Z. Dong, Y. Du, C. Yang, Y. Chen, Z. Chen, J. Jiang, R. Ren, Y. Li, X. Tang, Z. Liu, P. Liu, J.-Y. Nie, and J.-R. Wen, "A survey of large language models," 2024. [Online]. Available: https://arxiv.org/abs/2303.18223