Sparse Recovery with Attention: A Hybrid Data/Model Driven Solution for High Accuracy Position and Channel Tracking at mmWave

Yun Chen[†], Nuria González-Prelcic[†], Takayuki Shimizu[‡], Hongshen Lu[‡], and Chinmay Mahabal[‡]
[†] North Carolina State University, Email: {ychen273, ngprelcic}@ncsu.edu
[‡] Toyota Motor North America, Email: {takayuki.shimizu, hongsheng.lu, chinmay.mahabal}@toyota.com

Abstract—In this paper, we propose first a mmWave channel tracking algorithm based on multidimensional orthogonal matching pursuit algorithm (MOMP) using reduced sparsifying dictionaries, which exploits information from channel estimates in previous frames. Then, we present an algorithm to obtain the vehicle's initial location for the current frame by solving a system of geometric equations that leverage the estimated path parameters. Next, we design an attention network that analyzes the series of channel estimates, the vehicle's trajectory, and the initial estimate of the position associated with the current frame, to generate a refined, high accuracy position estimate. The proposed system is evaluated through numerical experiments using realistic mmWave channel series generated by ray-tracing. The experimental results show that our system provides a 2D position tracking error below 20 cm, significantly outperforming previous work based on Bayesian filtering.

Index Terms—V2X communication, mmWave MIMO, joint localization and communication, attention network.

I. Introduction

Wireless communication networks are introducing sensing into the functionalities offered to their users. High accuracy localization services are relevant for several vertical industries. In particular, highly/fully automated driving applications could be facilitated if the vehicles' positions were known by the network with an accuracy in the order of cm [1].

One way to obtain accurate location information in mmWave networks is based on exploiting the geometric relationships between the mmWave MIMO channel parameters and the position of the vehicle [2]. High accuracy single shot joint localization and channel estimation for initial access in vehicular systems has been addressed in recent work (see for example [3], [4] and references therein). Once the link has been established, both the accuracy of the channel and position estimates could be further improved. The work on joint channel and position tracking is, however, scarce, both in general and for the automotive application in particular.

Channel tracking methods exploiting mmWave channel sparsity and compressed sensing (CS) are introduced in [5], where an extended Kalman filter (EKF) exploits a known channel evolution model. A Kuhn-Munkres approach for

channel tracking is exploited in [6], while [7] proposes to use deep learning to refine tracking results. There are also studies focusing on joint channel tracking and localization [8]–[10]. Filters like EKF [8], particle filters [9], and Poisson multi-Bernoulli mixture (PMBM) filters [10] are exploited, which can be applied independently or interactively to the channel tracking and position tracking process.

The aforementioned methods have certain limitations when applied to vehicular systems: 1) they rely on unrealistic channel evolution models that assume a constant evolving rate which does not match practical vehicular systems; 2) they consider the channel as containing only line-of-sight (LOS) and first order non-line-of-sight (NLOS) paths, without specifying any mechanism to identify and discard estimated second order reflections, not exploited for localization; 3) the clock offset between the transmitter (TX) and receiver (RX) is neglected; and 4) no procedures to track both angle and delay channel parameters are provided, which are required for localization when the vehicle has a single active link to a base station (BS).

In this work, we focus on channel and position tracking in a realistic urban environment. First, we propose a low complexity channel tracking method based on multidimensional orthogonal matching pursuit (MOMP) [11]. Then, we design an attention network, V-ChATNet, that provides a refined, high accuracy tracking of the vehicle's position for LOS and NLOS settings. The inputs to V-ChATNet are the initial location estimate obtained from a geometric mapping of the tracked channel parameters and the series of previous channel and position estimates. It identifies the channel evolution patterns, associates the channel estimates with the localization results, and provides the location corrections to keep the location error below 0.2 m for 95% of the time.

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). $\mathbf{X}^\mathsf{T}$ and $\bar{\mathbf{X}}$ are the transpose and conjugate of $\mathbf{X}$. $[\mathbf{X}, \mathbf{Y}]$ and $[\mathbf{X}; \mathbf{Y}]$ are the horizontal and vertical concatenation of $\mathbf{X}$ and $\mathbf{Y}$. $\mathbf{X} \otimes \mathbf{Y}$ is the Kronecker product of $\mathbf{X}$ and $\mathbf{Y}$.

II. System Model

We consider a mmWave vehicular communication system where the BS is equipped with a uniform rectangular array

(URA) of size $N_{\mathrm{t}} = N_{\mathrm{t}}^{\mathrm{x}} \times N_{\mathrm{t}}^{\mathrm{y}}$, while the vehicle has 4 smaller URAs distributed on the hardtop as in [4], each of them of size $N_{\mathrm{r}} = N_{\mathrm{r}}^{\mathrm{x}} \times N_{\mathrm{r}}^{\mathrm{y}}$ elements. A hybrid MIMO architecture is adopted to transmit N_s data streams. The q-th time instance of the transmitted signal is denoted as $\mathbf{s}[q] \in \mathbb{C}^{N_s \times 1}$, with $\mathbb{E}[\mathbf{s}[q]\mathbf{s}[q]^*] = \frac{1}{N_s}\mathbf{I}_{N_s}$. The hybrid precoder and combiner are defined as $\mathbf{F} = \mathbf{F}_{\mathrm{RF}}\mathbf{F}_{\mathrm{BB}} \in \mathbb{C}^{N_{\mathrm{t}} \times N_s}$ and $\mathbf{W} = \mathbf{W}_{\mathrm{RF}}\mathbf{W}_{\mathrm{BB}} \in \mathbb{C}^{N_{\mathrm{r}} \times N_s}$, 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. The d-th tap of the MIMO channel with L paths can be formulated as

$$\mathbf{H}_d = \sum_{\ell=1}^{L} \alpha_\ell f_{\mathrm{p}}\left(dT_s - (t_\ell - t_{\mathrm{off}})\right) \mathbf{a}_{\mathrm{r}}(\boldsymbol{\theta}_\ell)\mathbf{a}_{\mathrm{t}}(\boldsymbol{\phi}_\ell)^*, \quad (1)$$

where t_{off} is the unknown clock offset between the TX and RX, α_ℓ and t_ℓ are the complex gain and the time of arrival (ToA) of the l-th path, T_s is the sampling interval, f_{p} is the pulse shaping function, and $\mathbf{a}_{\mathrm{r}}(\boldsymbol{\theta}_\ell)$ and $\mathbf{a}_{\mathrm{t}}(\boldsymbol{\phi}_\ell)$ represent the array responses of the l-th path evaluated at the direction-of-arrival (DoA) $\boldsymbol{\theta}_\ell$, and the direction-of-departure (DoD) $\boldsymbol{\phi}_\ell$. Note that $\mathbf{a}_{\mathrm{r}}(\boldsymbol{\theta}_\ell) = \mathbf{a}_{\mathrm{r}}(\theta_\ell^{\mathrm{az}}) \otimes \mathbf{a}_{\mathrm{r}}(\theta_\ell^{\mathrm{el}})$, and $\mathbf{a}_{\mathrm{t}}(\boldsymbol{\phi}_\ell) = \mathbf{a}_{\mathrm{t}}(\phi_\ell^{\mathrm{az}}) \otimes \mathbf{a}_{\mathrm{t}}(\phi_\ell^{\mathrm{el}})$, with $\theta_\ell^{\mathrm{az}}$, and $\theta_\ell^{\mathrm{el}}$ the DoA in azimuth and elevation, and ϕ_ℓ^{az} and ϕ_ℓ^{el} the DoD in azimuth and elevation. To simplify calculations, $\boldsymbol{\theta}_\ell$ and $\boldsymbol{\phi}_\ell$ are defined as unitary direction vectors, i.e., $\boldsymbol{\theta}_\ell = [\cos\theta_\ell^{\mathrm{el}}\cos\theta_\ell^{\mathrm{az}}, \cos\theta_\ell^{\mathrm{el}}\sin\theta_\ell^{\mathrm{az}}, \sin\theta^{\mathrm{el}}]^{\mathsf{T}}$; a similar definition applies to $\boldsymbol{\phi}_\ell$. Assuming the channel has N_{d} taps, the q-th instance of the received signal is

$$\mathbf{y}[q] = \mathbf{W}^* \sum_{d=0}^{N_{\mathrm{d}}-1} \sqrt{P_{\mathrm{t}}}\mathbf{H}_d\mathbf{F}\mathbf{s}[q-d] + \mathbf{W}^*\mathbf{n}[q], \quad (2)$$

where P_{t} is the transmitted power, and $\mathbf{n}[q] \sim \mathcal{N}(\mathbf{0}, \sigma_{\mathbf{n}}^2\mathbf{I}_{N_s})$ is modeled as additive white Gaussian noise (AWGN). We define the whitened received signal as $\check{\mathbf{y}}[q] = \mathbf{L}^{-1}\mathbf{y}[q]$, where $\mathbf{L}$ is computed via Cholesky decomposition of $\mathbf{W}^*\mathbf{W} = \mathbf{L}\mathbf{L}^*$. This way, the resulting noise term in the whitened received signal can be modeled as AWGN, i.e. $\mathbb{E}[\check{\mathbf{n}}[q]\check{\mathbf{n}}[q]^*] = \sigma_{\mathbf{n}}^2\mathbb{E}[\mathbf{L}^{-1}\mathbf{W}^*\mathbf{W}(\mathbf{L}^{-1})^*] = \sigma_{\mathbf{n}}^2\mathbf{I}$. Accordingly, the whitened received signal can be written as

$$\check{\mathbf{Y}} = \check{\mathbf{W}}^*[\mathbf{H}_0, ..., \mathbf{H}_{N_{\mathrm{d}}-1}]\left((\mathbf{I}_{N_{\mathrm{d}}} \otimes \mathbf{F})\mathbf{S}\right) + \check{\mathbf{N}}, \quad (3)$$

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

III. POSITION AND CHANNEL TRACKING SYSTEM

As discussed in Section I, previous work on high accuracy localization requires either delay and angular information from a single BS, or communication with several BSs to obtain an estimate of the position. In our system model we consider a communication link between a vehicle and a single mmWave BS, so delay and angular parameters of the channel need to be tracked. The design we propose in this Section tackles the problem of tracking the channel and the vehicle's position, while the channel estimation for initial access and initial localization could be realized by other methods in previous work, such as [3], [4]. The block diagram of our
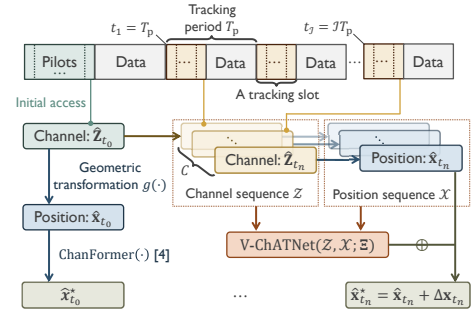

Fig. 1: Diagram of the position and channel tracking system.

proposed channel and position tracking system is shown in Fig.1. The channel tracking period is set to T_{p}, within which the channel is considered invariant. Every T_{p}, the channel is tracked using the procedure described in Section III-A. For every $T_n = nT_p$, the channel is fully re-estimated using the procedure described in [3], [4] to consider the case where significant changes occur over time (for example, a cluster that appears or disappears). The estimated channel is denoted as $\hat{\mathbf{Z}} \in \mathbb{R}^{N_{\mathrm{est}} \times 6}$, where N_{est} is the number of estimated paths, and the ℓ-th row contains the estimated parameters for the ℓ-th path, i.e. $\hat{\mathbf{z}}_\ell = [\hat{\alpha}_\ell, \hat{\tau}_\ell, \hat{\theta}_\ell^{\mathrm{az}}, \hat{\theta}_\ell^{\mathrm{el}}, \hat{\phi}_\ell^{\mathrm{az}}, \hat{\phi}_\ell^{\mathrm{el}}]$. We exploit the estimated time difference of arrival (TDoA) $\hat{\tau}_\ell = \hat{t}_\ell - t_{\mathrm{off}} - (\hat{t}^{\mathrm{min}} - t_{\mathrm{off}}) = \hat{t}_\ell - \hat{t}^{\mathrm{min}}$ for localization instead of the ToA, where $\hat{t}^{\mathrm{min}} = \min\{\hat{t}_\ell | \ell = 1, ..., N_{\mathrm{est}}\}$. The estimated paths have to satisfy the requirements for localization, i.e., the number of first order reflections has to be ≥ 1 in LOS channels or ≥ 3 in the NLOS case [3]. Otherwise, the vehicle cannot be located. We use *PathNet* [4] to determine the path orders and identify the LoS and first order reflections exploited for localization. Then, the initial 3D location estimate at the n-th time slot $t_n = nT_{\mathrm{p}}$ is defined as $\hat{\mathbf{x}}'_{t_n} = g\left(\hat{\mathbf{Z}}_{t_n}\right)$, where $g(\cdot)$ is the solution to the geometric system of equations defined in [4]. Since we are interested in the 2D position of the cars driving on the road, an attention network, V-ChATNet, designed in Section III, is then applied to refine the 2D vehicle location estimate $\hat{\mathbf{x}}_{t_n}$, where $\hat{\mathbf{x}}_{t_n} = \left[g\left(\hat{\mathbf{Z}}_{t_n}\right)\right]_{:2}$ if the vehicle can be located accordingly to the aforementioned criteria. Otherwise, $\hat{\mathbf{x}}_{t_n} = \hat{\mathbf{x}}_{t_{n-1}} + T_{\mathrm{p}}\hat{\mathbf{v}}_{t_{n-1}}$, where $\hat{\mathbf{v}}_{t_{n-1}}$ is the rough speed read from the speedometer at t_{n-1}. A sequence of historical channel estimates $\mathcal{Z}$ and 2D location estimates $\mathcal{X}$ are the input to the network, while the output is the correction of the current location estimate $\Delta\mathbf{x}_{t_n} = [\Delta x, \Delta y]$. The final location estimate is computed as $\hat{\mathbf{x}}^\star_{t_n} = \hat{\mathbf{x}}_{t_n} + \Delta\mathbf{x}_{t_n}$. In the reminder of this Section, we describe the details of the channel tracking algorithm and the attention network for position refinement.

A. MOMP-based Channel Tracking

Solutions for mmWave channel tracking proposed in previous work do not consider delay tracking, which disables the possibility of using these methods in a position tracking scenario where there is communication with a single BS. Second, they exploit a theoretical evolution model for the
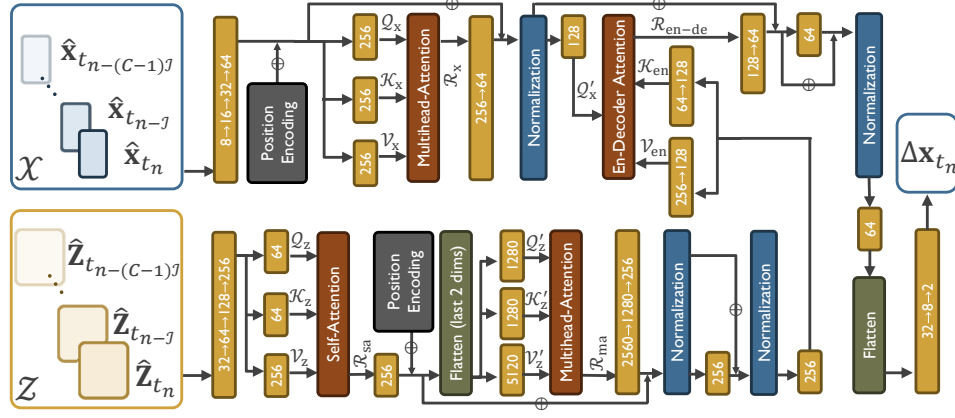
Fig. 2: Architecture of V-ChATNet which takes historical channel estimates and localization results from geometric transformations as the input, and outputs the correction for the current location estimate.

channel parameters that can be hardly met by a realistic vehicular channel, which leads to inaccurate estimations of the channel parameters. To overcome these limitations, we propose a channel tracking method that incorporates delay tracking without exploiting any rigid parameter evolution model. Our only assumption will be that the parameters will change smoothly, without considering any particular mathematical form. We will exploit this idea and the sparsity of the mmWave channel to develop a tracking procedure that relies on the recently defined MOMP algorithm [11], [12]. This algorithm solves a sparse recovery problem for channel estimation, without relying on sparsifying dictionaries based on Kronecker products, so that it can operate with large and planar arrays without incurring in prohibitive complexity and memory requirements, as discussed in [4], [11], [12]. The MOMP algorithm solves the optimization problem

$$
\min_{\mathbf{X}} \left(\left\| \mathbf{Y} - \sum_{\mathbf{i} \in \mathfrak{I}} \sum_{\mathbf{j} \in \mathfrak{J}} [\mathbf{\Phi}]_{:,\mathbf{i}} \left(\prod_{k=1}^{N_{\mathrm{D}}} [\mathbf{\Psi}_k]_{i_k,j_k} \right) [\mathbf{X}]_{\mathbf{j},:} \right\|_{\mathrm{F}}^2 \right), \quad (4)
$$

where $\mathbf{Y} = [\mathrm{vec}(\check{\mathbf{Y}}_1); ...; \mathrm{vec}(\check{\mathbf{Y}}_M)]$ collects M measurements using different pairs of training precoders and combiners, $\mathbf{F}_m$ and $\mathbf{W}_m$, for every training frame m, with $m = 1, ..., M$; $\mathbf{\Phi} \in \mathbb{C}^{MQN_s \times \otimes_{k=1}^{N_{\mathrm{D}}} N_k^s}$, where $[\mathbf{\Phi}]_{(m-1)QN_s+q,\mathbf{i}} = [(\mathbf{S}_m^{\mathsf{T}}(\mathbf{I}_{N_{\mathrm{d}}} \otimes \mathbf{F}_m^{\mathsf{T}})) \otimes \check{\mathbf{W}}_m^*]_{q,I}$ is the measurement tensor, where $I = (i_{N_{\mathrm{D}}} - 1) \prod_{k=1}^{N_{\mathrm{D}}-1} N_k^s + \sum_{k=1}^{N_{\mathrm{D}}-2} \left((i_k - 1) \prod_{k'=k+1}^{N_{\mathrm{D}}-1} N_{k'}^s \right) + i_{N_{\mathrm{D}}-1}$, N_{D} is the number of independent dimensions, and N_k^s is the length of the response vector along k-th dimension; $\mathbf{\Psi}_k \in \mathbb{C}^{N_k^s \times N_k^a}$ is the dictionary along dimension k, where N_k^a is the size of k-th dictionary depending on the resolutions; $\mathfrak{I} = \{\mathbf{i} = [i_1, ..., i_{N_{\mathrm{D}}}] \mid i_k \leq N_k^s\}$ and $\mathfrak{J} = \{\mathbf{j} = [j_1, ..., j_{N_{\mathrm{D}}}] \mid j_k \leq N_k^a\}$ are the sets for indices; and $\mathbf{X} \in \mathbb{C}^{\otimes_{k=1}^{N_{\mathrm{D}}} N_k^a \times 1}$ is the sparse tensor whose supports indicate the entries of the dictionaries to determine the channel parameters. MOMP-based channel estimation for initial access is introduced in [4], [11], [12], where the angular dictionaries for estimating the DoA/DoD span the

whole range of $[0, \pi]$, with a resolution of $\frac{\pi}{N_k^a}$, and the delay ranges from 0 s to $N_{\mathrm{d}} T_s$, with a resolution of $\frac{N_{\mathrm{d}} T_s}{N_3^a}$. For the tracking case, the dictionary for channel estimation at time t_n can be constructed, however, with the information from t_{n-1} to reduce the dictionary dimensions, reduce complexity and facilitate the estimation process. Let the estimated channel at time t_{n-1} be $\hat{\mathbf{Z}}_{t'_n}$, where the ℓ-th estimated path is $\hat{\mathbf{z}}_{t'_n,\ell} = [\hat{\alpha}_{t'_n,\ell}, \hat{\tau}_{t'_n,\ell}, \hat{\theta}_{t'_n,\ell}^{\mathrm{az}}, \hat{\theta}_{t'_n,\ell}^{\mathrm{el}}, \hat{\phi}_{t'_n,\ell}^{\mathrm{az}}, \hat{\phi}_{t'_n,\ell}^{\mathrm{el}}]$. The reduced angular dictionaries at t_n are determined as:

$$
\begin{cases}
\mathbf{\Psi}_{1,t_n} = [\bar{\mathbf{A}}(\hat{\phi}_{t'_n,1}^{\mathrm{az}}) \cup ... \cup \bar{\mathbf{A}}(\hat{\phi}_{t'_n,N_{\mathrm{est}}}^{\mathrm{az}})] \\
\mathbf{\Psi}_{2,t_n} = [\bar{\mathbf{A}}(\hat{\phi}_{t'_n,1}^{\mathrm{el}}) \cup ... \cup \bar{\mathbf{A}}(\hat{\phi}_{t'_n,N_{\mathrm{est}}}^{\mathrm{el}})] \\
\mathbf{\Psi}_{3,t_n} = [\mathbf{A}(\hat{\theta}_{t'_n,1}^{\mathrm{az}}) \cup ... \cup \mathbf{A}(\hat{\theta}_{t'_n,N_{\mathrm{est}}}^{\mathrm{az}})] \\
\mathbf{\Psi}_{4,t_n} = [\mathbf{A}(\hat{\theta}_{t'_n,1}^{\mathrm{el}}) \cup ... \cup \mathbf{A}(\hat{\theta}_{t'_n,N_{\mathrm{est}}}^{\mathrm{el}})]
\end{cases}, \quad (5)
$$

where $\mathbf{A}(\varphi) = \{\mathbf{a}(\varphi - \omega)\mathbf{a}(\varphi - \omega + \Delta\omega), ..., \mathbf{a}(\varphi + \omega)\}$ covers an angular sector of width 2ω with a resolution of $\Delta\omega$, and $\bar{\mathbf{A}}(\varphi)$ stands for the matrix that contains the conjugate of the entries in $\mathbf{A}(\varphi)$. Usually, ω is selected based on the beam width, assuming that the optimal beam pair remains the same throughout the tracking period. Furthermore, the reduced delay dictionary is determined as

$$
\mathbf{\Psi}_{5,t_n} = \left[\mathbf{p}_{\mathrm{d}}\left(\hat{t}_{t'_n}^{\mathrm{min}}\right), \mathbf{p}_{\mathrm{d}}\left(\hat{t}_{t'_n}^{\mathrm{min}} + \Delta\tau\right), ..., \mathbf{p}_{\mathrm{d}}\left(\hat{t}_{t'_n}^{\mathrm{min}} + \hat{\tau}_{t'_n}^{\mathrm{max}} + \varepsilon\right) \right], \quad (6)
$$

where $\Delta\tau$ is the delay dictionary resolution, $\hat{t}_{t'_n}^{\mathrm{min}} = \min\{\hat{t}_{t'_n,\ell} \mid \ell = 1, ..., N_{\mathrm{est}}\}$, $\hat{\tau}_{t'_n}^{\mathrm{max}} = \max\{\hat{\tau}_{t'_n,l} \mid l = 1, ..., N_{\mathrm{est}}\}$, ε is the allowed delay extending range, and $\mathbf{p}_{\mathrm{d}}(\cdot) \in \mathbb{C}^{N_{\mathrm{d}} \times 1}$ is the delay response where $[\mathbf{p}_{\mathrm{d}}(t)]_n = f_{\mathrm{p}}((n-1)T_s - t)$. By using the reduced dictionaries, the algorithm focuses on the areas where paths are most likely to exist, thereby solving the sparse recovery problem in (4) with reduced complexity. In addition, the resolutions could be configured to be smaller than those used for the full dictionaries to further improve the accuracy.

B. V-ChATNet for Vehicle Location Tracking

Attention schemes have been broadly studied in prior work to address context-aware problems [13]. Our proposed

V-ChATNet network enables the analysis of past channel estimates, the identification of channel evolution patterns, and the linking of the historical trajectory with the channel features for the correction of the current location estimate. The architecture of V-ChATNet is depicted in Fig. 2.

At the encoder, the network takes in a sequence of channel estimates $\mathcal{Z} = \left[\hat{\mathbf{Z}}_{t_{n-(C-1)\mathcal{I}}}, \hat{\mathbf{Z}}_{t_{n-(C-2)\mathcal{I}}}, ..., \hat{\mathbf{Z}}_{t_{n-\mathcal{I}}}, \hat{\mathbf{Z}}_{t_n}\right] \in \mathbb{R}^{C \times N_{\text{est}} \times 6}$, where C is the length of the input sequence, and $\mathcal{I}$ is the sample interval, which should be appropriately selected to adequately reveal the channel temporal evolution features. $\mathcal{Z}$ is first processed to obtain three abstract representations as *Value* $\mathcal{V}_{\text{z}}$, *Key* $\mathcal{K}_{\text{z}}$, and *Query* $\mathcal{Q}_{\text{z}}$, for self-attention of the paths of each channel. Mathematically,

$$\mathcal{R}_{\text{sa}} = \text{Attention}(\mathcal{Q}_{\text{z}}, \mathcal{K}_{\text{z}}, \mathcal{V}_{\text{z}}) = \text{softmax}(\frac{\mathcal{Q}_{\text{z}}\mathcal{K}_{\text{z}}^{\text{T}}}{\sqrt{\dim(\mathcal{K}_{\text{z}})}})\mathcal{V}_{\text{z}}, \tag{7}$$

where the calculations are for the last dimension of the abstractions. $\mathcal{R}_{\text{sa}}$, which has the same shape as $\mathcal{V}_{\text{z}}$, is the representation where each path incorporates effects from other paths in each channel, i.e., more accurately estimated paths correspond to higher weights. Afterwards, position encoding is applied to record the chronological order of the channels. The following multi-head attention stage extracts the temporal evolution features of the channels, with greater attention given to the channels that are better estimated in the sequence. The resulting representation $\mathcal{R}_{\text{ma}}$ goes through the feed forward and normalization modules and transforms to $\mathcal{V}_{\text{en}}$ and $\mathcal{K}_{\text{en}}$ as the output from the encoder.

The decoder takes the historical vehicle location estimates $\mathcal{X} = \left[\hat{\mathbf{x}}_{t_{n-(C-1)\mathcal{I}}}, \hat{\mathbf{x}}_{t_{n-(C-2)\mathcal{I}}}, ..., \hat{\mathbf{x}}_{t_{n-\mathcal{I}}}, \hat{\mathbf{x}}_{t_n}\right]$ as the input. After feature expansions via fully connected layers, $\mathcal{X}$ results in $\mathcal{V}_{\text{x}}$, $\mathcal{K}_{\text{x}}$, and $\mathcal{Q}_{\text{x}}$ to perform the multihead-attention to extract both the localization and the vehicle moving patterns over the given time period. The resulting representation of the location sequence $\mathcal{R}_{\text{x}}$ is treated as a query $\mathcal{Q}'_{\text{x}}$ to work with $\mathcal{V}_{\text{en}}$ and $\mathcal{K}_{\text{en}}$ for the encoder-decoder multi-head attention process. This helps establish the connections among the channel evolution, vehicle's trajectory, and the system errors introduced by the channel estimation and localization methods. The attention output $\mathcal{R}_{\text{en}-\text{de}}$ passes through the feed forward and normalization modules to provide the correction for the initial location estimate $\Delta\mathbf{x}_{t_n}$, so that the refined location becomes $\hat{\mathbf{x}}^{\star}_{t_n} = \hat{\mathbf{x}}_{t_n} + \Delta\mathbf{x}_{t_n}$.

In summary, V-ChATNet performs a regression task, i.e., $\Delta\mathbf{x}_{t_n} = \text{V-ChaTNet}(\mathcal{Z}, \mathcal{X}; \mathbf{\Xi})$, where $\mathbf{\Xi}$ represents the network parameters to be trained. The loss function for training the network is the mean square error (MSE) loss defined as $\mathcal{L}(\mathbf{\Xi}) = \|\hat{\mathbf{x}}^{\star}_{t_n} - \mathbf{x}_{t_n}\|^2$.

IV. SIMULATION RESULTS

We consider an urban canyon environment where cars and trucks are distributed across 4 lanes and moving 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. A 16×16 URA and 4 12×12
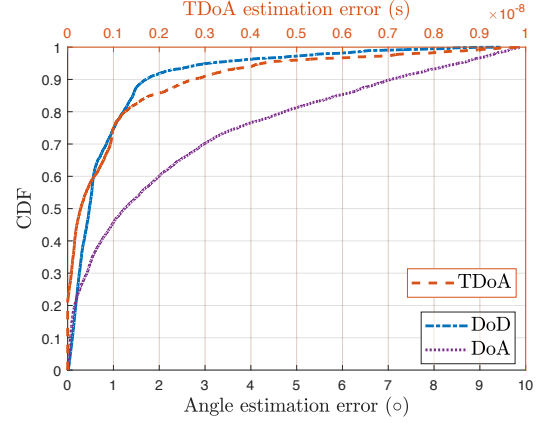

Fig. 3: Performance of channel tracking using reduced dictionaries with the tracking period of $T_{\text{p}} = 0.5$ ms.

URAs are deployed at the BS and the vehicle, respectively. The channel tracking period is set to $T_{\text{p}} = 0.5$ ms. In every tracking period, the BS transmits a data stream drawn from a row of a Hadmard matrix of size $Q = 64$, with a transmitted power of $P_{\text{t}} = 40$ 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 $f_c = 73$ GHz, with a bandwidth of $B_c = 1$ GHz. The dataset containing the channel series is generated by ray-tracing simulations using *Wireless Insite*, which provides realistic channels with higher order reflections, much weaker NLOS paths compared to the LOS, etc. We take 8 scenes where the vehicles' initial positions are different, so that the dataset contains 8 trajectories, with a split of 3:1 to form the training and testing sets for V-ChATNet. The trajectories have a length of 50 m in average, i.e., each set consists of roughly 6000 data samples.

We set $\omega = 15°$ and $\Delta\omega = 0.175°$ for the definition of the reduced angular dictionaries, and $\varepsilon = 0.2$ ns with a resolution of $\Delta\tau = 0.01$ ns for the reduced dictionary in the delay domain. The channel tracking algorithm produces $N_{\text{est}} = 5$ estimated paths for each channel, and *PathNet* [3], [4] is used to determine the path orders. The paths classified as LOS or first order reflections (which matter for localization) are matched against their nearest true paths in the channels. The matching results are shown in Fig. 3 as an evaluation of the channel tracking performance. The errors of the estimated DoD and DoA can be limited to a maximum of $3°$ and $8°$. The accuracy is higher for DoD estimations because a larger array is employed at the RSU. The TDoA estimation errors are limited to ≤ 7 ns.

The estimated channels combined with the initial location estimates are used for both training and testing V-ChATNet, so that the network can capture the noise features when using the proposed methods. The input sequence has a length of $C = 16$, and the sample interval is set to $\mathcal{I} = 25$ to capture the channel temporal variations. V-ChATNet is trained with the batch size set to 64, the learning rate set to 1×10^{-4}, and the number of training epochs set to 1000 with the
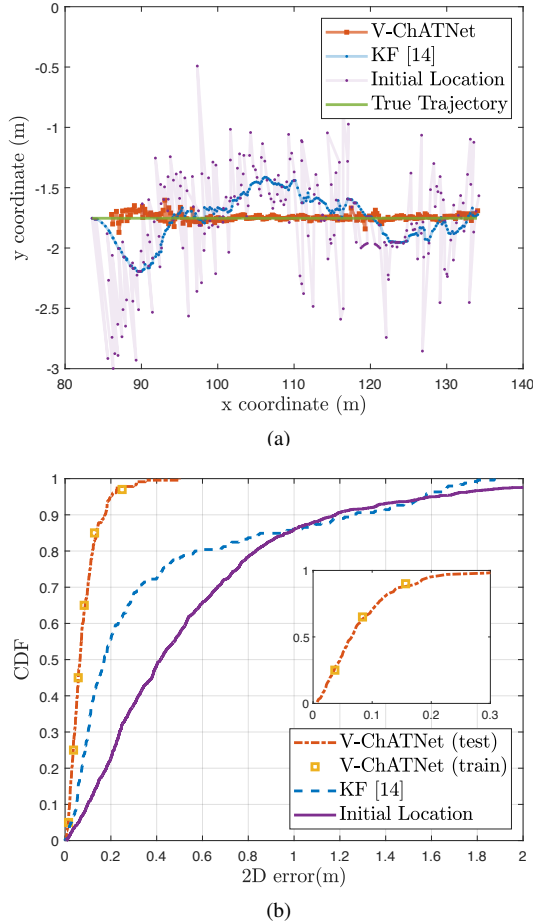
V. CONCLUSION

We developed a system for joint location and channel tracking in realistic urban environments. The low complexity MOMP algorithm with reduced dictionaries enabled accurate channel tracking results, leading to an initial position estimate based on the solution of a system of geometric equations with an error below ~ 1.61 m 95% of the time. Then, we developed V-ChATNet, an attention network that takes the sequences of the historical channel estimates and the initial location estimates, and extracts both the temporal and spatial features, providing a correction to the current location estimate. The experimental results show that our method achieves an accuracy of 0.2 m for 95% of the time for a vehicle driving at 60 km/h in a realistic environment simulated by ray tracing.

REFERENCES

[1] 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.

[2] 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. Wireless Commun.*, vol. 17, no. 3, pp. 1822–1835, 2018.

[3] 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 *IEEE12th Sensor Array and Multichannel Signal Processing Workshop (SAM)*, 2022, pp. 355–359.

[4] 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 positioning," *arXiv preprint*, 2023.

[5] S.-H. Wu and G.-Y. Lu, "Compressive beam and channel tracking with reconfigurable hybrid beamforming in mmWave MIMO OFDM systems," *IEEE Transactions on Wireless Communications*, 2022.

[6] L. Zhu, D. He, B. Ai, K. Guan, S. Dang, J. Kim, H. Chung, and Z. Zhong, "A ray tracing and joint spectrum based clustering and tracking algorithm for internet of intelligent vehicles," *Journal of Communications and Information Networks*, vol. 5, no. 3, pp. 265–281, 2020.

[7] Y. Chen, L. Yan, C. Han, and M. Tao, "Millidegree-level direction-of-arrival estimation and tracking for terahertz ultra-massive MIMO systems," *IEEE Transactions on Wireless Communications*, vol. 21, no. 2, pp. 869–883, 2021.

[8] 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, 2021.

[9] X. Chu, Z. Lu, D. Gesbert, L. Wang, X. Wen, M. Wu, and M. Li, "Joint vehicular localization and reflective mapping based on team channel-SLAM," *IEEE Transactions on Wireless Communications*, vol. 21, no. 10, pp. 7957–7974, 2022.

[10] H. Kim, K. Granström, L. Svensson, S. Kim, and H. Wymeersch, "PMBM-based SLAM filters in 5G mmWave vehicular networks," *IEEE Transactions on Vehicular Technology*, vol. 71, no. 8, pp. 8646–8661, 2022.

[11] 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.

[12] ——, "Low complexity joint position and channel estimation at millimeter wave based on multidimensional orthogonal matching pursuit," in *2022 30th European Signal Processing Conference (EUSIPCO)*. IEEE, 2022, pp. 1002–1006.

[13] S. Chaudhari, V. Mithal, G. Polatkan, and R. Ramanath, "An attentive survey of attention models," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 12, no. 5, pp. 1–32, 2021.

[14] M. B. Khalkhali, A. Vahedian, and H. S. Yazdi, "Vehicle tracking with kalman filter using online situation assessment," *Robotics and Autonomous Systems*, vol. 131, p. 103596, 2020.

Fig. 4: Vehicle position tracking performance based on the testing set. (a) An example of tracking results from sky view; (b) Vehicle position tracking accuracy using V-ChATNet, comparing with that using KF [14] and initial localization based on channel geometric transformations.

early stopping strategy to avoid overfitting. Fig. 4a shows an example of the tracking results of a trajectory from the testing set, where the initial location estimate using the channel geometric transformation is denoted as "Initial Location", and a Kalman filter (KF) [14] is also simulated for comparison. Results show that the positions corrected by V-ChATNet are the most accurately aligned with the true trajectory. The cumulative distribution function (CDF) of the localization accuracy can be observed in Fig. 4b. The initial localization based on the geometric system of equations yields the 5, 50, 80, and 95-th percentile 2D errors of 0.058, 0.418, 0.833, and 1.611 m. These values decrease to 0.027, 0.171, 0.575, and 1.574 m when using the KF, while V-ChATNet offers a significant improvement in the accuracy, reducing the errors up to 1.4 m, with the percentile values of 0.014, 0.065, 0.120, and 0.197 m. Moreover, the network exhibits consistent performance on both the training and testing sets, indicating that it possesses reliable generalization capabilities.