Object Trajectory Estimation with Multi-Band Wi-Fi Neural Dynamic Fusion

Kato, Sorachi; Wang, Pu; Koike-Akino, Toshiaki; Fujihashi, Takuya; Mansour, Hassan; Boufounos, Petros T.
TR2024-019 March 16, 2024

Abstract

In contrast to existing multi-band Wi-Fi fusion in a frame-to-frame basis for simple classification, this paper considers asynchronous sequence-to-sequence fusion between sub-7 GHz channel state information (CSI) and 60 GHz beam SNR for more challenging downstream tasks such as continuous regression. To handle the timing disparity between the two channel measurements, we extend our recently proposed dual-decoder neural dynamic (DDND) framework with latent ordinary differential equations (ODEs), align the distinct latent dynamic states at the same time instances, and introduce a post-ODE fusion framework. The resulting neural dynamic fusion (NDF) framework is trained in an end-to-end fashion with a modified variational autoencoder loss function. Evaluation over a newly collected in-house multi-band Wi-Fi dataset shows the advantage of the proposed NDF method over frame-based and DDND methods.

*IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*
2024

© 2024 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works.
OBJECT TRAJECTORY ESTIMATION WITH
MULTI-BAND WI-FI NEURAL DYNAMIC FUSION

Sorachi Kato¹², Pu Wang¹, Toshiaki Koike-Akino¹, Takuya Fujihashi², Hassan Mansour¹, Petros Boufounos¹

¹Mitsubishi Electric Research Laboratories (MERL), Cambridge, MA 02139, USA
²Graduate School of Information Science and Technology, Osaka University, Suita, Osaka, Japan

ABSTRACT
In contrast to existing multi-band Wi-Fi fusion in a frame-to-frame basis for simple classification, this paper considers asynchronous sequence-to-sequence fusion between sub-7 GHz channel state information (CSI) and 60 GHz beam SNR for more challenging downstream tasks such as continuous regression. To handle the timing disparity between the two channel measurements, we extend our recently proposed dual-decoder neural dynamic (DDND) framework with latent ordinary differential equations (ODEs), align the distinct latent dynamic states at the same time instances, and introduce a post-ODE fusion framework. The resulting neural dynamic fusion (NDF) framework is trained in an end-to-end fashion with a modified variational autoencoder loss function. Evaluation over a newly collected in-house multi-band Wi-Fi dataset shows the advantage of the proposed NDF method over frame-based and DDND methods.

Index Terms— WLAN sensing, 802.11bf, Wi-Fi sensing, ISAC, localization, multi-band fusion, and dynamic learning.

1. INTRODUCTION
Wi-Fi sensing, e.g., device localization and device-free human sensing, has received much attention in the past decade from both academia and industry. This trend has been manifested by the establishment of a new task group (TG) - 802.11bf WLAN Sensing – in September 2020, to go beyond data transmission and meet industry demands for robust and reliable wireless sensing.

Existing studies are primarily based on the coarse-grained receiver signal strength indicator (RSSI) and the fine-grained channel state information (CSI) in terms of channel frequency response over OFDM subcarriers [1–5]. At a high frame rate, CSI reflects intrinsic channel features from frequency subcarriers (delay) and multiple transmitter-receiver pairs (angle) but may experience channel instability to even small-scale environment changes. These features can be extracted from frame-based or sequence-based frameworks [6–9]. On the other hand, mid-grained mmWave beam training measurements at 60 GHz, e.g., beam SNR, have shown better channel stability over time [10–18]. These beam SNR measurements originate from sector-level directional beam training, a mandatory step for mmWave Wi-Fi to compensate for large path loss and establish the link between the AP and the user. However, they suffer from low frame rate and irregular sample intervals due to the beam training overhead and follow-up association steps. To deal with such intermittent sampling issues over multiple frames, [19] proposed a dual-decoder neural dynamic learning (DDND) framework that learns the underlying latent dynamics in a continuous-time fashion by exploiting the neural ODE framework [20–23].

Fusion-based approaches have been considered in the literature for robustness and better accuracy. Heterogeneous sensor fusion was studied between Wi-Fi and other modalities, e.g., Bluetooth and acoustics [24–26]. Within Wi-Fi channel measurements, CSI and RSSI can be simply concatenated for a joint feature extraction [27]. [28] proposed to fuse the phase and amplitude of the fine-grained CSI for localization. When multiple access points (APs) are deployed in the scene, multi-AP fusion was proposed in [29, 30] by exploring the generalized interview and intraview discriminant correlation analysis and, respectively, the maximum mean discrepancy (MMD) criterion. To the best of our knowledge, our previous work in [16] is the only effort considering multi-band Wi-Fi fusion between CSI and beam SNR. However, it is limited to simple classification tasks, e.g., pose classification (over 8 stationary poses), seat occupancy sensing (8 static patterns), and fixed-grid localization. Despite being sampled at different time instances, the two channel measurements can be simply combined on a frame-to-frame basis as these asynchronous samples correspond to the same stationary label (e.g., pose, occupancy, location) and their respective sampling time becomes irrelevant for the fusion; see the top plot of Fig. 1.

In this paper, we substantially advance the multi-band Wi-Fi fusion framework, evolving from the static frame-to-frame basis of [16] to a dynamic asynchronous sequence-to-sequence basis (see the bottom plot of Fig. 1.), thus supporting more challenging downstream tasks, e.g., continuous regression and continuous-time object trajectory estimation, as opposed to simple classification problems. Referring to the neural dynamic fusion (NDF) framework, it is achieved by extending the beam SNR-only DDND framework of [19] to two separate neural ODE decoders defined on a shared time axis, forcing the two decoders to generate virtual latent dynamic states in the same time instances and combining these synchronized
post-ODE latent dynamic states via a fusion block for continuous trajectory estimation. To further align the two distinct latent spaces over time, we encode both the CSI and the beam SNR sequences to estimate their respective initial conditions at a common starting time \( t_0 \), which can precede the first sample from either channel measurement. To train the proposed multi-decoder neural dynamic fusion network, we consider a loss function that is a weighted sum of waveform reconstruction losses at asynchronous time instances and coordinate estimation errors at these synchronized time instances. With a newly collected in-house multi-band Wi-Fi dataset for robot trajectory estimation, comprehensive performance evaluation confirms the effectiveness of the proposed neural dynamic fusion over a list of baseline methods.

### 2. PROBLEM FORMULATION

We formulate trajectory estimation as a regression with asynchronous CSI and beam SNR sequences. At time \( t_n \), we collect a set of \( M \) beam SNR values \( \mathbf{b}_n = [b_1, b_2, \cdots, b_M]^\top \in \mathbb{R}^{M_b \times 1} \), each corresponding to one beam training pattern. For CSI, at time \( t_n \), we collect a channel frequency response matrix \( \mathbf{C}_n \in \mathbb{C}^{N_f \times N_t \times N_s} \) over \( N_s \) OFDM subcarriers, \( N_t \) transmitting antennas, and \( N_f \) receiving antennas. For a time window of length \( \Delta T_w \), we collect \( N_b \) beam SNR samples and \( N_c \) CSI samples with which we aim to estimate the trajectory at \( N_p \) time instances \( t_{n}^{b}, n = 1, \cdots, N_p \). Note that \( N_b \) and \( N_c \) may vary from one time window to another.

The problem of interest is to estimate the coordinate of a trajectory at any time point \( t_{n}^{b} \) within the time window with the beam SNR and CSI input sequences along with their respective time stamps,

\[
\{b_n, t_n\}_{n=0}^{N_b}, \{C_n, t_n\}_{n=0}^{N_c} \rightarrow \{p_n\}_{n=0}^{N_p}, \tag{1}
\]

where \( p_n = [x_n, y_n]^\top \) consists of two-dimensional coordinates at \( t_{n}^{p} \).

### 3. MULTI-BAND NEURAL DYNAMIC FUSION

The proposed multi-band sequence-to-sequence Wi-Fi fusion framework is shown in Fig. 2. From left to right, we have two ODE-RNN encoders for both beam SNR and CSI, two separate latent dynamic ODE blocks, a post-ODE fusion block, and three decoders for waveform reconstruction and coordinate estimation. In the following, we introduce each block in more details\(^1\).

#### 3.1. Embedding Layers

For the fine-grained CSI matrix \( \mathbf{C}_n \), we follow standard calibration steps to remove sampling time offset (STO) between transmitter and receiver due to sampling frequency offsets and packet detection errors and carrier frequency offset (CFO) among receiver RF chains [31, 32]. The calibrated complex-valued CSI matrix at \( t_n^{c} \) is then flattened and mapped to an embedding space via a pre-trained one-dimensional convolution (Conv1D) network. Specifically, we have \( \mathcal{E}(\mathbf{C}_n) = c_n \in \mathbb{R}^{M_c \times 1} \) as the embedding vector of the CSI. For the mid-grained beam SNR \( b_n \), we directly take the raw beam SNR as the input to the following ODE-RNN encoders. To maintain the balance between the two input dimensions, we have approximately \( M_b \approx M_c \).

#### 3.2. ODE-RNN Encoders

For the two separate encoders for beam SNR and CSI, respectively, we follow the ODE-RNN encoder architecture [19, 21] to take the reversed input sequences \( t_{N}, t_{N-1}, \cdots, t_{1} \) to estimate the initial latent conditions at the common starting time \( t_0 \).

Let us start from beam SNR encoding. With an ODE-RNN encoder, each recurrent unit updates its hidden vector \( \mathbf{h}_n \in \mathbb{R}^{L_h \times 1} \) with an auxiliary vector \( \mathbf{h}_{n+1}^{b,c} \) and \( \mathbf{b}_n \).

\[
\mathbf{h}_n = \mathcal{G}_{\theta_b}(\mathbf{h}_n^{b,c}, \mathbf{b}_n), \tag{2}
\]

where \( \mathcal{G}_{\theta_b} \) can be either GRU or LSTM unit with learnable parameters \( \theta_b \). Then the ODE-RNN encoder utilizes an ODE function \( \mathcal{O}_{\phi_b} \) to describe the propagation of the latent vector in a continuous-time fashion,

\[
\frac{d\mathbf{h}(t)}{dt} = \mathcal{O}_{\phi_b}(\mathbf{h}(t), t), \tag{3}
\]

\(^1\)Note that the mathematical ornamental characters represent neural networks, and the \( \theta \) subscript at the bottom right denote learnable parameters.

\(^2\)Since the ODE-RNN encoder applies to both beam SNR and CSI, we ignore the superindex, e.g., \( \mathbf{h}_n^{b,c} \) or \( t_n^{b,c} \) for simplicity.
where the ODE function is parameterized by a multi-layer perceptron (MLP) network with learnable parameters $\theta_c^c$. Utilizing a numerical ODE solver $S$, e.g., Euler and Runge-Kutta solvers, one can then propagate the hidden vector $h_{n+1}$ at time $t_{n+1}$ to the auxiliary vector $\tilde{h}_n$ at the current time $t_n$ (recall that the time is in a reversed order for estimating the initial condition):

$$h_{n+1} = S(O_{\theta_c^c}(h_{n+1}, (t_n, t_{n+1})))$$

$$= h_{n+1} + \int_{t_n}^{t_{n+1}} O_{\theta_c^c}(h(\tau), \tau) d\tau. \quad (4)$$

By iterating between (2) and (4), we can propagate the latent encoding vector from $t_N$ to $t_0$, and the same procedure can be adopted for CSI encoding with another set of parameters $\theta_c^c$.

### 3.3. Latent Dynamic Learning

Once the hidden state $h_0$ at $t_0$ is obtained, $h_0$ is used to generate $z_0$, the initial condition in the latent space for latent dynamic learning. Following the VAE framework [33], the posterior distribution of $z_0$ is approximated as

$$q_b(z_0|h_0) = \mathcal{N}(\mu_{z_0}, \sigma_{z_0})$$

where the mean and standard deviation are mapped from $h_0$ by

$$\mu_{z_0}, \sigma_{z_0} = \mathcal{M}_b(h_0) \quad (5)$$

with $\mathcal{M}$ denoting an MLP. We sample $z_0^b \in \mathbb{R}^{L_z^b}$ and $z_0^c \in \mathbb{R}^{L_z^c}$ from their respective posterior mean and standard deviations as:

$$z_0^b = \mu_{z_0}^b + \sigma_{z_0}^b \odot \epsilon_1, \epsilon_1 \sim \mathcal{N}(0, I_{L_z^b}),$$

$$z_0^c = \mu_{z_0}^c + \sigma_{z_0}^c \odot \epsilon_2, \epsilon_2 \sim \mathcal{N}(0, I_{L_z^c}), \quad (6)$$

where $\odot$ represents Hadamard product.

For the latent dynamic learning, we further enforce that the above initial latent conditions $z_0^b$ and $z_0^c$ are aligned at the common starting time $t_0^b = t_0^c = t_0$. In this way, the latent dynamic learning block takes the sampled initial latent condition at $t_0$ and propagates to the latent dynamic state at any query time $t_n$ achieved by using another continuous-time ODE function $O_d$ modeled by a neural network with parameters $\theta_d^{b/c}$.

We first query the beam SNR-related and CSI-related dynamic learning blocks with their respective sampling time $t_n^{b/c}$.

$$z_n^b = z_n^b = z_0^b + \int_{t_0}^{t_n} O_{\theta_d^b}(z(t), t) dt = S(O_{\theta_d^b}, z_{b}(t_0, t_n),)$$

$$z_n^c = z_n^c = z_0^c + \int_{t_0}^{t_n} O_{\theta_d^c}(z(t), t) dt = S(O_{\theta_d^c}, z_{c}(t_0, t_n).) \quad (8)$$

These latent states for beam SNR and CSI are then fed to the waveform decoders (see Section 3.5) for waveform/feature reconstruction.

Due to the continuous-time dynamic modeling capability, we then query the beam SNR-related and CSI-related dynamic learning blocks at a set of unseen but shared time instances $t_n^{b/c}$ for supervised training and downstream tasks. Specifically, we have

$$z_n^{b/c} = z_n^{b/c} = z_0^{b/c} + \int_{t_0}^{t_n} O_{\theta_d^{b/c}}(z(t), t) dt = S(O_{\theta_d^{b/c}}, z_{b/c}(t_0, t_n).) \quad (9)$$

where the above ODE associated parameters $\theta_d^{b/c}$ are the same as the ones in (8).

### 3.4. Post-ODE Latent Fusion

It is seen from (9) that the latent dynamic states $z_n^{b/c}$ for the beam SNR are synchronized with their corresponding CSI latent dynamic state $z_{n+1}^{b/c}$ at $t_n^{b/c}$. Such a latent dynamic alignment motivates us to combine the two post-ODE dynamic pathways into a fused dynamic pathway that might be more essential for downstream tasks such as the object trajectory estimation.

To this end, we first project the aligned post-ODE states into higher dimensions

$$\tilde{z}_n^{b/c} = \mathcal{M}_{\theta_d^{b/c}}(z_n^{b/c}), \quad n = 1, \cdots, N_p \quad (10)$$

using an MLP with learnable weights $\theta_d^{b/c}$, then concatenate these projected latent states, and finally compress it into a fused latent state $\tilde{z}_n$ at $t_n^{b/c}$.

$$\tilde{z}_n = \mathcal{M}_{\theta_f}(\tilde{z}_n^{b/c}, \tilde{z}_n^{c/c}), \quad n = 1, \cdots, N_p \quad (11)$$

where $\theta_f$ consists of the weight matrices and bias terms for the fusion MLP. It is expected that the post-ODE latent fusion also allows for a combination of the dynamics of distinct Wi-Fi propagation characteristics at different frequency bands in a complementary way and enhances the ability to represent a unified latent space for downstream tasks.

### 3.5. Multi-Head Decoders

From the latent dynamic learning and fusion blocks, we have the latent dynamic states in three distinct sets of time instances:

- $z_n^b, n = 1, \cdots, N_b, $ at the beam SNR sampling time $t_n^b$;
- $z_n^c, n = 1, \cdots, N_c, $ at the CSI sampling time $t_n^c$;
- $\tilde{z}_n, n = 1, \cdots, N_p, $ at shared time instances $t_n^{b/c}$.

For the first two sets of latent dynamic states, we use two separate MLP heads to decode them back to the beam SNR waveform or CSI embedding feature spaces as

$$\tilde{b}_n = \mathcal{M}_{\theta_b}(z_n^b), \quad \tilde{c}_n = \mathcal{M}_{\theta_c}(z_n^c). \quad (12)$$

For the fused latent states, we use another MLP head to project them into a coordinate trajectory as

$$\tilde{p}_n = \mathcal{M}_{\theta_p}(\tilde{z}_n), \quad n = 1, \cdots, N_p. \quad (13)$$

All MLP heads are shared over time steps.

### 3.6. Loss Function

In the following, we adopt a loss function modified from the evidence lower bound (ELBO) of the standard VAE [33] for the proposed NDF framework without providing the detailed derivation

$$L = \sum_{n=0}^{N_p} \| \tilde{p}_n - p_n \|^2$$

$$+ \lambda_1 \sum_{n=0}^{N_e} \| \tilde{c}_n - c_n \|_2 - \lambda_2 \sum_{t=0}^{L_p^c} \left( 1 + \log(\sigma_t^c)^2 - (\mu_t^c)^2 - (\sigma_t^c)^2 \right)$$

$$+ \lambda_3 \sum_{n=0}^{N_b} \| \tilde{b}_n - b_n \|_2 - \lambda_4 \sum_{t=0}^{L_b^b} \left( 1 + \log(\sigma_t^b)^2 - (\mu_t^b)^2 - (\sigma_t^b)^2 \right),$$

where the hyperparameters $\lambda_1/2/3/4$ play the trade-off roles between the supervised trajectory estimation errors to the waveform (beam SNR) and embedding feature (CSI) reconstruction errors.
Fig. 3: A TurtleBot testbed and trajectory configuration for data collection. The map on the right was created by a LiDAR on TurtleBot.

![Fig. 3](image)

Fig. 4: CDF of localization errors.

4. PERFORMANCE EVALUATION

4.1. In-House Testbed and Data Collection

We built a testbed to automatically collect CSIs and beam SNRs from a moving device, shown in Fig. 3. We used ASUS RT-AC86U for IEEE 802.11ac CSI collection, and TP-Link Talon AD7200 for IEEE 802.11ad beam SNR collection. Two routers were installed for IEEE 802.11ac CSI collection, and TP-Link Talon AD7200 for IEEE 802.11ad beam SNR collection. Two routers were installed for IEEE 802.11ac CSI collection, and TP-Link Talon AD7200 for IEEE 802.11ad beam SNR collection.

We used ASUS RT-AC86U to grasp driving the TurtleBot along a predefined route shown as red scatter points in Fig. 3. The TurtleBot was equipped with a LiDAR to grasp its position while moving, and recorded the coordinate information used as labels for training. We had $M_b = 36$ for the beam SNR and $(N_{Rx}, N_{Tx}, N_b) = (4, 2, 234)$ for the CSI.

4.2. Implementation

We set $\Delta T_b = 5$ seconds to group the corresponding CSI, beam SNR, and coordinate labels. We divided all 5 s sequences into training and test sets with a ratio of 8:2. The beam SNRs $\{b_n\}_{n=0}^{N_b}$ were normalized to $[0, 1]$, and so were the timestamps $\{t_n\}_{n=0}^{N_b}$ and $\{t_{n,b}\}_{n=0}^{N_b, b}$ to be compatible with the ODE layers. $M_b$ for the embedded CSI feature vectors $\{c_n\}_{n=0}^{N_b}$ was set to 36, and the features were also normalized to $[0, 1]$. The hidden status dimension of GRU $L_h$ was set to 20, and the latent dimension $L_z = L_{t,z} = 20$. For the loss function, we empirically set $\lambda_1 = \lambda_3 = 0.1$ for the waveform reconstruction and $\lambda_2 = \lambda_4 = 0.01$ for the KL divergence loss. Neural networks were implemented using PyTorch 2.0.1 on Python 3.11 and trained on a GPU with CUDA 12.1 enabled.

In this paper, we proposed an asynchronous multi-band Wi-Fi fusion framework using latent ODE learning. Specifically, the framework projects CSI and beam SNR onto their own latent space, utilizes a post-ODE neural dynamics fusion to align these measurements in the latent space, and estimates object trajectory from the aligned and fused latent variables. Real-world experiments validate the proposed neural dynamic fusion framework.
6. REFERENCES


