2DLAM: Joint Delay-Doppler Estimation in UAV mmWave System via Large AI Model

Daixin Xie[1*], Chenxi Liu[1*], Wei Wang[2*], Fengxian Guo[1], Xiaoling Hu[1]

[1]State Key Laboratory of Networking and Switching Technology,
Beijing University of Posts and Telecommunications, Beijing, 100876, China
[2]Peng Cheng Laboratory, Shenzhen, 518000, China
{xdxxiedaixin, chenxi.liu, fengxianguo, xiaolinghu, }@bupt.edu.cn, wangw01@pcl.ac.cn

Abstract

Unmanned aerial vehicle (UAV) has been recognized as a promising platform for fulfilling high-rate data communication in the sixth-generation (6G) wireless networks, due to the benefits of strong line-of-sight (LoS) link probability, controlled mobility, and on-demand deployment. In UAV-enabled communication systems, channel estimation plays a crucial role and has been facing increasing challenges, especially in high frequency millimeter-wave (mmWave) band. This is because the channel coherence time in UAV mmWave systems will be significantly shortened by the severe delay and Doppler effect arised from high-speed movement of the UAV, making the traditional channel estimation approaches designed for low-speed scenarios less applicable. To address this issue, in this paper, we propose a novel channel estimation algorithm suitable for UAV-enabled communication systems, in which a terrestrial base station equipped with a large-scale array communicates with a UAV target of high mobility. In particular, the proposed algorithm leverages the large artificial intelligence model (LAM) to jointly estimate the time-varying parameters in the delay-Doppler domain within the shortened coherence time, thus improving the accuracy of channel estimation in high-speed scenarios. More specifically, we characterize the time-varying channel matrices as two-dimensional (2D) images, thus allowing us for incorporating the pre-trained imageGPT (iGPT) model to handle the joint delay and Doppler parameters estimation. Through comparisons with various benchmarks, we demonstrate how our proposed algorithm can accurately estimate the delay and Doppler parameters in the considered UAV-enabled communication systems with a relatively small training cost.

Introduction

Unmanned aerial vehicles (UAVs)-enabled communication has been regarded as the essential component of the sixth generation (6G) wireless networks. Utilizing the benefits of low-power consumption, robust line-of-sight (LoS) links, and portable deployment, UAV-enabled communication is applied to facilitate numerous emerging applications with growing data traffic, such as emergency assistance (Ye et al. 2023), live-streaming of sports events (Li and Wang 2023) and precision agriculture (Mukhamediev et al. 2023). In light of this context, there are ongoing research interests in UAV communication systems, particularly in the utilization of millimeter-wave (mmWave) frequency bands. This area of study is of particular interest due to the unique capabilities of high-mobility UAVs to establish strong LoS links with compact antenna arrays and their ability to dynamically adjust their location in the three-dimensional space, in order to support high data-rate communication services.

However, in practice, the high mobility and high carrier frequency of UAV mmWave communication systems over a relatively long distance introduce additional time delay and Doppler shifts. It is inevitably caused by high carrier frequency, fast movement and angular spread, resulting in delay effect and carrier frequency offset (Xiao et al. 2022). In this case, the channel coherence time is short and the distribution of UAV mmWave channel is complex, which makes it difficult to recover the channel state information (CSI) accurately. To address this challenge, researchers have introduced two types of estimators: the model-based channel estimator and the traditional artificial intelligence (TAI)-based channel estimation. For model-based estimators, compressive sensing theory has been proposed to recover the channel matrices (Ayach et al. 2014; Zhao et al. 2021). For the TAI-based estimators, the problem is solved by introducing classical machine learning approaches (Chen, Yan, and Han 2021; Xu, Feng, and Li 2023). However, these two kinds of techniques have their own limitations. The model-based estimators are limited by the performance of the models themselves and typically require extensive prior knowledge about the scenarios and the underlying theory of the algorithms. On the other hand, the estimators leveraging TAI frameworks are better at conducting sample analysis and classification tasks. In particular, these methods primarily concentrate on slow fading channels, neglecting the Doppler effect and round-trip delay, which is a critical factor in UAV millimeter-wave (mmWave) channels. Consequently, they are not suitable for application in such dynamic environments characterized by high-mobility UAVs.

Recently, generative artificial intelligence (GAI)-based algorithms have been proposed to enhance the learning capabilities with more diversified tasks in wireless communication. In particular, Transformer-based algorithms, which are capable of extracting long-range dependencies in the input sequence. This approach enables the processing of mul-

*These authors contributed equally.

Copyright © 2025, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

tiple communication signals by assigning equal weight to each, simplifying the process and enhancing efficiency. It also excels in processing parallel input sequences of varying numbers, thereby suitable for channel parameters estimation with multiple instances of communication signals. This provides a potentially powerful tool to learn the complex features of the UAV mmWave channel so as to solve the channel estimation problem (e.g. (Liu et al. 2023; Kim et al. 2023; Hu et al. 2023)). In (Liu et al. 2023), a channel estimator Vision Transformer (CE-ViT) based on self-attention mechanism is proposed for dynamic scenarios. Moreover, Transformer-based parametric Terahertz (THz) channel acquisition is proposed in (Kim et al. 2023) to depict the correlations between channel sparsity and received communication signal, which leads to less training overhead in the online training process. In (Hu et al. 2023), a Transformer-based channel estimator is proposed for reconfigurable intelligent surface (RIS)-assisted communication. However, although these works have performance gains in capturing the channel feature with the Transformer framework, the aforementioned works were unable to adapt to the channel parameter estimation in UAV mmWave systems. Firstly, the existing works were not typically adapted to the high-mobility scenarios for UAVs. They just solved the channel estimation process without paying close attention to the delay and Doppler variation, thus limiting the scalability in actual deployments. Secondly, these algorithms just introduce the GAI framework mechanically, which cannot activate the great potential of learning complex distribution from the wireless channel contained in the GAI-based network. Therefore, it is essential to design a novel channel parameter estimation scheme with better scalability and estimation capability.

Large artificial intelligence models (LAMs) based on GAI framework have significantly propelled advancements in the field of computer vision (CV) and natural language processing (NLP). In particular, the LAM exhibits robust analytical and pattern recognition competence for deciphering the nuanced meanings and contexts within natural language, enabling it to produce comprehensive responses to targeted inquiries (Du et al. 2023). Conversely, by leveraging pre-trained knowledge, LAMs have demonstrated the capacity to adapt to new tasks with a few datasets. This efficiency could enhance scalability in scenarios where the availability of training data is constrained within the UAV mmWave communication systems (Jin et al. 2024). Therefore, this offers an alternating approach to mitigate the aforementioned limitations in current channel estimators. However, most of the existing works in the field of wireless communication mainly focused on the time series forecasting tasks in slow fading channels (Liu et al. 2024; Zhang et al. 2024). Consequently, the internal properties of the UAV mmWave channel in high-mobility scenarios are not considered in these works, leading to a degradation in network performance or even causing the network to malfunction.

Inspired by the above research, in this paper, we propose a novel channel delay-Doppler estimation algorithm for the UAV mmWave communication system by LAM. In this system, a base station (BS) with a multi-antenna array performs channel parameter estimation procedure based on the pilot signals within the coherence time. Our contributions can be outlined as: 1) We consider a UAV mmWave communication system where the geometry-based channel model is utilized. The UAV sends the uplink pilot signals to the BS. Leveraging the multiple received uplink signals within coherence time, we define the channel parameter estimation problem to minimize delay-Doppler estimation error. 2) To address the aforementioned problem, we propose a novel joint delay-Doppler estimation method by LAM. To the best of our knowledge, this paper marks the initial endeavor to employ a pre-trained LAM for channel parameters estimation. 3) We examine the effect of key system parameters on the estimation performance in the UAV mmWave system, demonstrating its better estimation performance over multiple benchmarks with fewer training parameters, especially when outputting both delay and Doppler estimates simultaneously, the estimation errors for both parameters are relatively low.

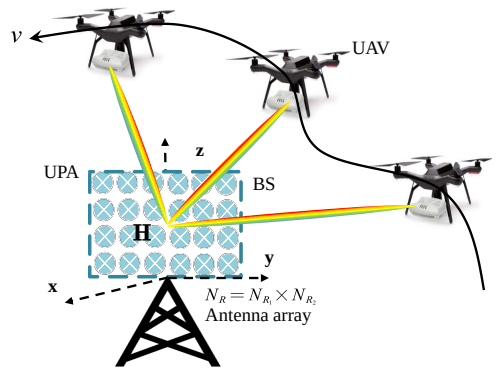


Figure 1: The UAV mmWave system with uniform planar antenna array.

System Model And Problem Formulation

In this section, we first present the UAV mmWave communication system and then formulate a problem of estimating delay-Doppler vectors through the uplink communication signals.

We consider a TDD UAV mmWave communication system, in which K single-antenna UAVs served by a single base station (denoted by the BS). We consider that the BS is equipped with a uniform planar array (UPA) of $N_R = N_{R_1} \times N_{R_2}$ receive antennas, as shown in Figure 1. Moreover, We assume the k-th UAV fly at different fixed altitudes with the speed of v, and transmits the uplink pilots to the BS. Suppose the BS only has limited prior information about the 3D UAV position from the UAV navigation information (i.e., it has no idea about the accurate angle, velocity and distance between BS and the UAVs) (Bertizzolo et al. 2019).

Communication Signal Model

In this subsection, we introduce the channel model employed in our paper. We first represent the position of the

BS as $\mathbf{u}_{BS} = [x_r, y_r, z_r]^T$. Then, the position of the i-th element antenna unit at the BS can be expressed as

$$\mathbf{u}_{BS,(i,j)} = \mathbf{u}_{BS} + \mathbf{a}_{BS,(i,j)}, \tag{1}$$

where

$$\mathbf{a}_{BS,(i,j)} = d[i-1, 0, j-1]^T \tag{2}$$

denotes the displacement of the (i,j)-th element antenna from the first element antenna at the BS. In (2), $d = \frac{\lambda_c}{2}$ is the antenna spacing with λ_c denoting the carrier wavelength, $i \in \{1, 2, \cdots, N_{R_1}\}$, and $j \in \{1, 2, \cdots, N_{R_2}\}$.

Afterwards, we specifically utilize a geometry-based model (Xie et al. 2024) to describe the mmWave channel between the BS and the k-th UAV. Note that this geometry-dependent model has shown to be effective in accurately capturing the channel responses of time-varying channels (as in (Wang and Zhang 2022; Chen et al. 2024)), thus is suitable for our considered dynamic UAV mmWave systems. Moreover, since the non-line-of-sight (NLoS) path is much weaker than the LoS path for UAV mmWave communications, we only consider the dominated line-of-sight (LoS) paths in the communication channel. As per the rules of the geometry-based model, we formulate the channel between the k-th UAV and the (i,j)-th element antenna on the BS as follows

$$h_{k,i,j} = \frac{\lambda_c}{4\pi d_{k,i,j}} e^{j\frac{-2\pi d_{k,i,j}}{\lambda_c}}, \tag{3}$$

where $d_{k,i,j}$ denotes the distance from the k-th UAV to the (i,j)-th element antenna unit at the BS, given by

$$d_{k,i,j} = \left\| \mathbf{u}_{BS,(i,j)} - \mathbf{u}_{U,k} \right\|_2. \tag{4}$$

In (4), $\mathbf{u}_{U,k}$ denotes the location of the k-th single-antenna UAV.

Owing to the relative motion of the k-th UAV and BS, a Doppler shift is observed in the communication channel. As such, the beamspace channel between transceivers at time t and frequency f is expressed as

$$\mathbf{H}(t,f) = e^{j2\pi(\nu_{k,n}t - f\tau_{k,n})} \begin{pmatrix} h_{1,1} & h_{1,2} & \cdots & h_{1,N_{R_2}} \\ h_{2,1} & h_{2,2} & \cdots & h_{2,N_{R_2}} \\ \vdots & \vdots & \ddots & \vdots \\ h_{N_{R_1},1} & h_{N_{R_1},2} & \cdots & h_{N_{R_1},N_{R_2}} \end{pmatrix}, \tag{5}$$

where $\nu_{k,n} = \frac{v\cos\theta_p}{\lambda_c}$ denotes the round-trip Doppler shift with θ_p denotes the angles of arrival of the path, and $\tau_{k,n}$ denotes the round-trip delay.

Based on (1)–(5), we express the signal received from the k-th UAV by the BS during t-th, $t \in \{1, 2, \cdots, \tau^{UL}\}$, time slot within the n-th, $n \in \{1, 2, \cdots, N\}$, epoch as follows

$$\mathbf{y}_{k,n}(t) = \sqrt{P_{k,n}}\mathbf{H}_{\mathbf{vec}} x_{k,n}(t) + \mathbf{n}_{k,c}(t), \tag{6}$$

where $P_{k,n}$ denotes the transmit power, $\mathbf{H}_{\mathbf{vec}}$ denotes the vectorized $\mathbf{H}$, $x_{k,n}(t)$ is the transmitted pilot at time slot t, and $\mathbf{n}_{k,c}(t)$ denotes the additive white Gaussian noise (AWGN) with zero mean and covariance matrix σ_c^2, i.e., $\mathbf{n}_{k,c}(t) \sim \mathcal{CN}(0, \sigma_c^2)$. Note that we assume the delay-Doppler pair remains nearly constant during the limited duration of single epoch and the transmitted pilot is known at the BS. For simplicity, we assume that $x_{k,n}(t) = 1$, the signal-to-noise ratio (SNR) from the UAV to the BS can be expressed as $\mathrm{SNR} = \frac{P_{k,n}|\varrho|^2}{\sigma_n^2}$, where $\varrho = \frac{\lambda\sqrt{N_R}}{4\pi\bar{d}}$ and $\bar{d}$ denotes the distance between UAV and BS.

Problem Formulation

Given that the UAV navigation system cannot provide a *priori* knowledge about UAV targets with high-accuracy (Jongsintawee et al. 2016; Zaliva and Franchetti 2014), in this study, we aim to estimate the CSI of moving UAV target. We formulate the uplink signal with τ^{UL} time steps at the n-th epoch as

$$\mathbf{y}_{k,n,\tau^{UL}}^{UL} = \left[\mathbf{y}_{k,n}(1), \mathbf{y}_{k,n}(2), \cdots, \mathbf{y}_{k,n}\left(\tau^{UL}\right)\right], \tag{7}$$

where $\mathbf{y}_{k,n,\tau^{UL}}^{UL}$ represents the sets of uplink signals at the n-th epoch. Therefore, the problem can be formulated as,

$$\begin{aligned} &\min_{\Psi} \ \mathbf{NMSE} = E\left\{\frac{\|\hat{\mathbf{p}}_{k,n} - \mathbf{p}_{k,n}\|^2}{\|\mathbf{p}_{k,n}\|^2}\right\} \\ &s.t. \ \hat{\mathbf{p}}_{k,n} = h_{\Psi}\left(\mathbf{y}_{k,n,\tau^{UL}}^{UL}\right), \end{aligned} \tag{8}$$

where $\hat{\mathbf{p}}_{k,n}$ denotes the estimated $(\hat{\tau}_{k,n}, \hat{\nu}_{k,n})$ and $\mathbf{p}_{k,n}$ denotes the actual delay-Doppler pair, $\mathbf{y}_{k,n}(t)$ denotes the received uplink pilots at the t-th time slot and $h_{\Psi}(\cdot)$ denotes the mapping function derived from the learning-based training process where Ψ represents the learnable network parameters.

Due to the sparsity of mmWave channel in the angular domain, the channel parameter estimation problem is previously combined with the channel estimation of slow fading massive multiple-input multiple-output (MIMO) channels with the conventional deep learning techniques (e.g., convolutional neural network (CNN) (Xu, Feng, and Li 2023; Chen, Yan, and Han 2021), long short term memory (LSTM) (Jiang and Schotten 2020)). However, it is difficult for CNNs to handle the data with series. The input sequence of LSTMs and other recurrent neural network (RNN)-derived algorithms is processed element-by-element, which means that its parallelization ability is relatively weak, which limits the training speed of the model. To resolve this issue, in this paper, we introduce a novel LAM-based delay-Doppler estimation framework to concurrently process the received signals over multiple time slots within the coherence time. Remarkably, since the sparse channel can be regarded as a 2D natural picture to handle, we introduce the pre-trained imageGPT (iGPT) model for improving the learning capability. The details of our proposed algorithm will be presented in the following section.

Proposed LAM-based delay-Doppler Estimation Network

In this section, we propose our novel LAM-based delay-Doppler estimation algorithm, which is designed based on

the ViT framework integrating the pre-trained iGPT network. Specifically, the iGPT network is employed and other critical parameters in the ViT framework are fine-tuned, merging the advantages of LAM-based feature extraction with those of data-driven deep learning methodologies.

Preliminaries on ViT Algorithm

In this subsection, we provide a brief review of the ViT algorithm (Dosovitskiy et al. 2021). Inspired by the effectiveness of the standard Transformer in the NLP tasks, a new algorithm, namely ViT, achieves prominent results compared to state-of-the-art convolutional networks and the original Transformer in the field of image recognition. The network

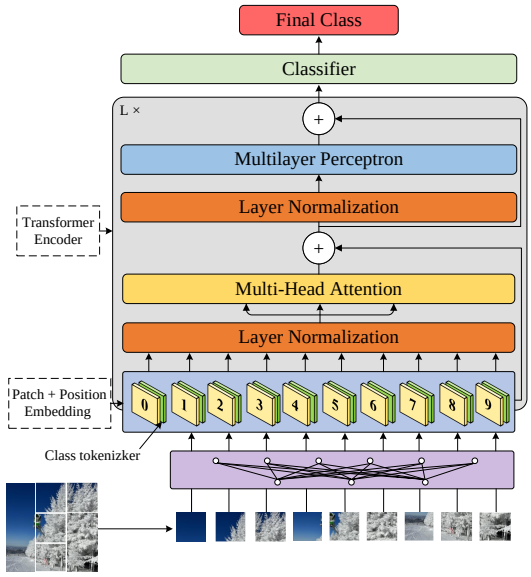


Figure 2: Illustration of the ViT structure.

structure is illustrated in Figure 2. The image is split into modules of the same size and then the patches are inputted concurrently. All the patches are linearly embedded, add position embeddings. Afterwards, they are fed into the traditional Transformer encoder simultaneously. The classifier is composed of two layers, incorporating a GELU activation function within its architecture. In this case, the overall network structure can be written as

$$y_0 = \left[v_{class}; v_p^1 E; v_p^2 E; \cdots; v_p^N E\right] + E_{pos} \tag{9a}$$

$$y'_k = \mathrm{MHA}\left(\mathrm{LN}\left(y_{k-1}\right)\right) + y_{k-1}, \quad k = 1, \cdots, K \tag{9b}$$

$$y_k = \mathrm{MLP}\left(\mathrm{LN}\left(y\prime_k\right)\right) + y\prime_k, \quad k = 1, \cdots, K \tag{9c}$$

$$r = \mathrm{LN}\left(y_K^0\right). \tag{9d}$$

where $v_p^i E$ denotes the patches as $i = 1, \cdots, N$, E denotes the patch embeddings and E_{pos} denotes the position embeddings, $\mathrm{LN}(\cdot)$ denotes the layer normalization, $\mathrm{MHA}(\cdot)$ denotes the multi-head attention layer and $\mathrm{MLP}(\cdot)$ denotes the Multilayer Perceptron. Specifically, v_{class} denotes the learnable embedding which plays a role as the image representation r at the K-th layer output of the Transformer encoder y_K^0. In practice, the final classifier can be replaced with other functional layers, i.e., calculate mean value, to acquire the specific numerical solutions as the problem that needs to be solved is the regression tasks (Zhou et al. 2021).

LAM-Based Framework For Delay-Doppler Estimation

Despite ViT being an image classification method of good performance, its algorithmic processing pattern cannot be directly applied to the channel parameter estimation problem from UAV mmWave communication systems. Furthermore, the absence of channel-specific design in the UAV mmWave communication system fails to capture the interplay between channel sparsity and the received communication signals. Meanwhile, due to the sparsity in our considered channel model, we can regard the channel matrices as a typical 2-D picture. As such, the delay-Doppler estimation can be regarded as a representative image recognition problem. For these aforementioned reasons, in this subsection, we present a delay-Doppler estimation algorithm by introducing LAM-based estimation module based on the ViT framework discussed above. For brevity, we refer to our proposed LAM-based algorithm as the 2DLAM algorithm. The network architecture is shown in Figure 3, which consists of a preprocessor module, an embedding module, a parameter estimation module and an output module. The preprocessor is to parallelize and refine the received signals as the manageable input data. The embedding module follows the ViT's fashion introduced in the preceding subsection, which provide an order of embeddings of these patches as an input to the parameter estimation module. For high-accuracy feature extraction, we introduce the pre-trained iGPT model as the parameter estimation module. Finally, the output module is to formulate the output sequence of delay-Doppler pair according to the pattern of solving regression problems in Transformer. Details of the four modules are given below.

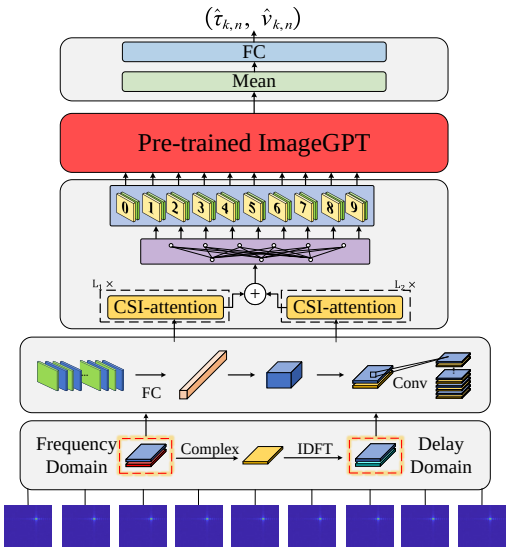


Figure 3: Illustration of the proposed 2DLAM structure.

Preprocessor Module During the n-th epoch, the previous τ^{UL} slots communication signals are received from the k-th UAV and then the network estimates the $(\tau_{k,n}, \nu_{k,n})$. Inspired by the splitting and flattening operations in the ViT, we parallelize the τ^{UL}-slot received signals and then concurrently feed them into the preprocessor unit. Note that the existing deep learning methods cannot directly handle the complex channel matrices, thereby we transform the complex-valued channel matrices into the two-layer real-valued matrices, whose layers contains the real and imaginary parts, i.e.,

$$\mathbf{Y}_{UL} = \left[\mathrm{Re}\left(\mathbf{y}_{k,n}(1), \cdots, \mathbf{y}_{k,n}\left(\tau^{UL}\right)\right); \mathrm{Im}\left(\mathbf{y}_{k,n}(1), \cdots, \mathbf{y}_{k,n}\left(\tau^{UL}\right)\right)\right]^T \in \mathbb{R}^{\tau^{UL} \times 2N_R} \tag{10}$$

Next, based on the utility function in (8), the optimization process can be divided into delay estimation and Doppler estimation. Due to the time-frequency characteristics of the mmWave channel in (5), for the delay estimation, the channel matrices should be converted into its corresponding representation in the delay domain via the inverse discrete Fourier transform (IDFT) (Yang et al. 2020; Liu et al. 2024), i.e.,

$$\mathbf{Y}_{\tau} = \mathbf{F}^H \mathbf{Y}_{UL}, \tag{11}$$

where $\mathbf{F}$ denotes the DFT matrix. The Doppler estimation utilizes the time-frequency domain matrices directly without any conversion.

Embedding Module Given the preprocessed data $\mathbf{Y}_{\tau}$, we apply embedding module to accomplish patch embedding and add class tokenizker. Firstly, we encode each patch into a token embedding. Inspired by (Liu et al. 2024), we utilize the CSI-attention module to achieve the patch embedding. This step aims to make the essential feature extraction more effective in the subsequent steps. As per the rules in (Woo et al. 2018), the overall process in the CSI-attention module can be expressed as

$$\mathcal{B}_C(\mathbf{Y}') = Sigmoid\left(\mathcal{F}_2\left(\mathcal{F}_1\left(\mathbf{Y}'_{avg}\right)\right) + \mathcal{F}_2\left(\mathcal{F}_1\left(\mathbf{Y}'_{\max}\right)\right)\right) \tag{12}$$

where $\mathbf{Y}'_{avg}$ and $\mathbf{Y}'_{\max}$ denote the average-pooled and max-pooled features, $\mathbf{Y}' \in \mathbb{R}^{\tau' \times L \times 2N_R}$ is the rearranged input tensors, $Sigmoid(\cdot)$ denotes the sigmoid function, The convolutional blocks $\mathcal{F}_1(\cdot)$ and $\mathcal{F}_2(\cdot)$ are shared among inputs, with a ReLU activation function succeeding $\mathcal{F}_1(\cdot)$.

We formulate the output of the CSI-attention module as

$$\mathbf{Y}_{\mathcal{B}_C} = \mathcal{B}_C(\mathbf{Y}_{UL}) + \mathcal{B}_C(\mathbf{Y}_{\tau}), \tag{13}$$

where $\mathcal{B}_C(\cdot)$ denotes the CSI attention module with iterating specific number of times.

For position embedding, similar to (Vaswani et al. 2023), we apply sine and cosine functions with different frequencies, which is given by

$$E_{pos}(2i) = \sin\left(\mathrm{pos}/10000^{2i/d_{\mathrm{model}}}\right), \tag{14a}$$

$$E_{pos}(2i+1) = \cos\left(\mathrm{pos}/10000^{2i/d_{\mathrm{model}}}\right), \tag{14b}$$

where d_{model} denotes the feature dimension of the pre-trained LAM module. Consequently, the final embeddings can be given as,

$$y_0 = \mathbf{Y}_{\mathcal{B}_C} + E_{pos}. \tag{15}$$

Parameter Estimation Module In this step, we adopt the pre-trained iGPT as the feature extractor. Transformer models such as BERT and GPT-2(Radford et al. 2019) are domain agnostic. This means that they can be applied directly to one-dimensional sequences of any form. However, these models are not designed specifically for the characteristics of 2-D images, such as sparsity and object appearance. Recent works report the adaptability and effectiveness of the iGPT (Chen et al. 2020) in the realm of image classification. Therefore, we adjust the channel matrices into the 1-D sequence and obtain the embeddings of the received signals. We formulate the processing of iGPT as

$$y'_k = \mathrm{iGPT}(y_0), \tag{16}$$

where $\mathrm{iGPT}(\cdot)$ denotes the backbone language model which is pre-trained via ImageNet-21k dataset and it is image-specific model.

Inspired by the progress of the text processing in GPT-2 network and image-oriented sequence Transformer in iGPT, we apply the pre-trained ImageGPT-large network as the backbone for the parameter estimation module. The structure of iGPT is almost the same as the GPT-2, while the different activation layer function is utilized and the layer normalization does not mean center the inputs (Chen et al. 2020). Notably, during the off-line training procedure, the learnable parameters in the iGPT network are frozen. This operation also reduces the training overhead during the off-line training process.

Output Module To obtain the delay-Doppler estimation, in this step, we introduce the output module for regression problems. We first utilize a mean value calculator to obtain the mean in the dimension of the time slot. Afterwards, we propose a fully convolutional layer to achieve the final output $(\hat{\tau}_{k,n}, \hat{\nu}_{k,n})$. The procedure in the output module can be expressed as,

$$r = \mathcal{B}_O(y'_k), \tag{17}$$

in which $\mathcal{B}_O(\cdot)$ denotes the output function consists of a mean value calculator and two fully connected layers.

Model Training

Following the rules of self-supervised learning, the 2DLAM network mainly works in two stages: offline training and online estimation. In the offline training stage, as the BS already knows the approximate location range from the UAV navigation system, we use the prior knowledge to adjust the label of training sets into a similar order of magnitude, i.e., multiply the delay by 10^7 and Doppler by 10^{-3}. Then we establish the self-supervised learning procedure to optimize all the learnable parameters. In the online estimation stage, after accepting the testing set data, the trained 2DLAM network can output the estimated delay-Doppler pairs.

Complexity Analysis

In this subsection, we present the complexity and normalized mean square error (NMSE) comparisons between our proposed 2DLAM algorithm and other comparative methods as $\mathrm{SNR} = -1$ dB, including the matched-filtering (MF) scheme (Roberts et al. 2010), CNN scheme (Heng,

Mo, and Andrews 2022), LSTM scheme (Jiang and Schotten 2020) and ViT scheme (Dosovitskiy et al. 2021). Note

Model	No. Network Parameters	Training Time	NMSE
MF	0	0 ms	-5.90 dB
CNN	2.16M/2.16M	1.45 ms	-19.09 dB
LSTM	1.12M/1.12M	19.97ms	-30.10 dB
ViT	1.85M/1.85M	14.38 ms	-30.01 dB
2DLAM	1.82M/172.56M	38.25 ms	-35.51 dB

Table 1: Complexity (training parameters/total parameters and the training cost per batch) and delay NMSE comparisons.

that since the size of channel matrices is different from the three-channel natural image, we modify some key configurations in the ViT network, i.e., hidden size, MLP size, number of layers and number of heads, in order to maximize the effectiveness of the ViT framework on the problem we are solving. Similar modifications are also applied on the comparative CNN algorithm. The Doppler NMSE is defined as $\frac{\|\hat{\nu}_{k,n}-\nu_{k,n}\|^2}{\|\nu_{k,n}\|^2}$, and the delay NMSE is defined as $\frac{\|\hat{\tau}_{k,n}-\tau_{k,n}\|^2}{\|\tau_{k,n}\|^2}$.

Notably, the MF scheme is the traditional method using the weighted least-squares scheme. Thus, the number of network parameters and training time per batch are ignored in the MF scheme. Given the perspective of space complexity, in Table 2, we can see that the proposed 2DLAM algorithm can achieve significantly better performance with the cost of a larger number of total parameters. However, the 2DLAM algorithm requires much fewer training parameters than the comparison schemes because the parameters of the pre-trained iGPT are frozen. This indicates the effectiveness of the pre-trained LAM module. Furthermore, given the perspective of time complexity, we can observe that the superior performance from the proposed 2DLAM algorithm is obtained with the cost of high training time. This can be attributed to the utilization of large LAM model compared with the existing schemes, which results in high computational overhead.

Simulation Results

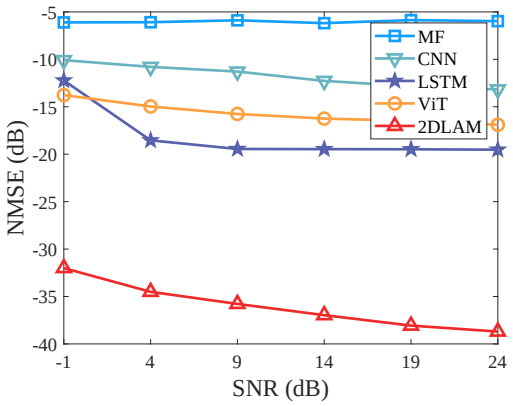
In this section, we first present the simulation setup of this work. Then we provide the simulation results to certify the effectiveness of our proposed algorithm.

Simulation Setup Unless otherwise specified, we use the simulation setup in the whole simulation process as follows. We adopt the antenna as $N_R = 64$ with $N_{R_1} = N_{R_2} = 8$ at the BS. The carrier frequency is 28 GHz. The UAV target is assumed to move with 30 $\sim$ 60 m/s while the round-trip Doppler shift approximately ranges from $f_d \in [3, 6]$ KHz. The relative position between the UAV and BS is assumed to be random variables, i.e., $[x_{k,n}, y_{k,n}] = [\dot{x}_{k,n} + \triangle x, \dot{y}_{k,n} + \triangle y]$ where $\triangle x \sim \mathcal{N}(0, 1)$ and $\triangle y \sim \mathcal{N}(0, 1)$. As such, we set $\tau^{UL} = 8$ as the number of time steps from the received signals and the length of each step is $\Delta T = 2$ms.

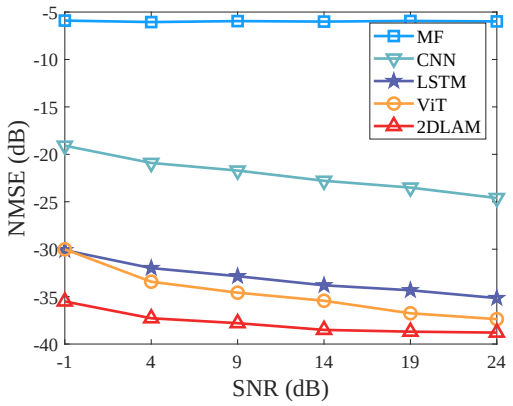
For the self-supervised learning, we adopt $L_1 = L_2 = 4$ as the number of CSI-attention module. Following the structure of iGPT-large, we set the feature dimension as 512 and the first $K = 4$ layers are introduced. We utilize the PyTorch deep learning framework to accomplish the overall proposed network. The training and the validating datasets consist of 20000 and 2560 samples, respectively, as the UAV target uniformly distributed with $\theta_p \in [60°, 120°]$, the mean inaccurate UAV distance between the transceivers is assumed as 90 m and and the variance is assumed as 1 m. The transmitted power $P_{k,n}$ is ranged from 5 dBm to 35 dBm. Considering the distance and transmitted power, we investigate the simulation results from $\mathrm{SNR} \in [-1, 24]$ dB. The testing dataset consists of 2560 samples with the mean inaccurate UAV position from the GPS system and barometer is $[0, 90, 10]^T$ m and the variance is assumed as 1 m. The network is trained with Adam optimizer as $\mathrm{betas} = (0.9, 0.999)$. The optimizer also introduces the regularization term and its decay rate of the weight is set as $\varepsilon = 0.001$. The learning rate is adjusted automatically, where the maximum learning rate is $\mathrm{lr_max} = 0.001$. The network is trained with 400 epoches.

Results Firstly, in Figure 4(a) and Figure 4(b), we study the Doppler NMSE performance under different SNRs for various schemes. For comparative analysis, the Doppler NMSE results of various benchmark methods are included. From the figures, the deep learning-based algorithms outperform the MF algorithm due to the superiority of neural network and the deep learning-based algorithms can obtain better NMSE performance with increasing SNR. This is because the deep learning framework is able to effectively extract features from the large number of training data for each UAV target. This fact exhibits the superior estimation performance compared to the other two deep learning-based algorithms. Moreover, the performance gap between our proposed 2DLAM and other comparative schemes are particularly evident at higher SNR. This is due to the basic ViT-based framework can handle the channel matrices concurrently with the same weight. Furthermore, our proposed 2DLAM algorithm can concurrently achieve significantly better NMSE performance, while the ViT algorithm can only achieve better performance than CNN and MF in the delay domain. It can be attributed to the usage of pre-trained iGPT in our proposed algorithm and the CSI-attention module, indicating it is an effective design for channel matrices feature extraction. In conclusion, these results verify that the proposed 2DLAM algorithm is a powerful algorithm that can effectively achieve better NMSE performance compared to other existing schemes and can effectively estimate the delay-Doppler properties at the same time. This improvement can be attributed to the suitable channel estimation framework of ViT, as well as the utilization of an image-specific iGPT backbone network. In addition, this gain can greatly improve the sensing performance of the radar system.

Moreover, in Figure 5, we plot the Doppler NMSE perfor-



(a) Doppler NMSE performance.



(b) Delay NMSE performance.

Figure 4: NMSE performance versus different SNRs for various schemes.

mance versus the number of steps τ^{UL} for different schemes with SNR = 4 dB. To this end, we include the NMSE performance of the comparative ViT algorithm under different τ^{UL}. We can see that, the Doppler NMSE performance of proposed 2DLAM scheme improves with the increasing time steps. Conversely, the ViT and LSTM network fails to yield performance improvements with an increasing number of time steps. We also observe that, when $\tau^{UL} = 1$, the difference of performance in results is negligible. As the number of time step increases, there is a noticeable growth of the performance gap between proposed 2DLAM approach and comparative methods, which verifies the effectiveness of the application of the pre-trained LAM method.

Finally, we derive the comparison of different number of iGPT layers on the delay NMSE performance when $\mathrm{SNR} = -1$ dB and their training costs (training parameters/total parameters and the training time per batch). We can see that as the number of iGPT layers increases, the training costs also increases accordingly. However, as the number of iGPT layers equals to 6, the delay NMSE performance can achieve better performance, which means that we can just deploy limited number of LAM to achieve better performance based on the actual training situation.

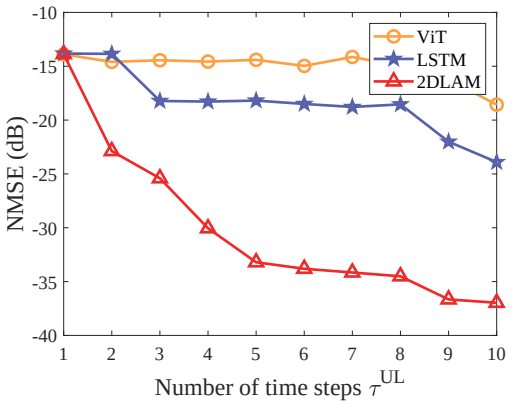


Figure 5: Doppler NMSE performance versus the number of steps τ^{UL} for different schemes with SNR = 4 dB.

No. LAM layers	No. Network Parameters	Training Time	NMSE
iGPT(2)	1.82M/59.27M	30.57 ms	-33.97 dB
iGPT(4)	1.82M/115.93M	36.33 ms	-32.08 dB
iGPT(6)	1.82M/172.56M	38.25 ms	-35.51 dB
iGPT(8)	1.82M/229.24M	39.03 ms	-35.44 dB

Table 2: The delay NMSE performance, number of network parameters versus different number of iGPT layers.

Conclusion

In this paper, we propose a novel LAM-based delay-Doppler estimation algorithm for UAV mmWave communication systems. The BS is equipped with a UPA and receives communication signals within the coherence block from the high-mobility UAV targets to accomplish channel parameter estimation. In our proposed algorithm, we introduce the Transformer-based ViT framework to dispose of the input signal sequence, thus improving the performance of parallel signal processing of sequences. Based on this, we apply the pre-trained iGPT model to improve the scalability and estimation capability in the UAV mmWave system. Simulation results show that our proposed algorithm can achieve better NMSE performance compared with other benchmark schemes, especially when the delay-Doppler properties are output at the same time. Furthermore, we demonstrate that our proposed algorithm can achieve state-of-the-art performance by leveraging the benefits of the pre-trained iGPT network with fewer training costs.

References

Ayach, O. E.; Rajagopal, S.; Abu-Surra, S.; Pi, Z.; and Heath, R. W. 2014. Spatially Sparse Precoding in Millimeter Wave MIMO Systems. *IEEE Transactions on Wireless Communications*, 13(3): 1499–1513.

Bertizzolo, L.; Polese, M.; Bonati, L.; Gosain, A.; Zorzi, M.; and Melodia, T. 2019. mmBAC: Location-aided mmWave Backhaul Management for UAV-based Aerial Cells. In *Proceedings of the 3rd ACM Workshop on Millimeter-Wave Networks and Sensing Systems*, 7–12. New York, NY, USA: Association for Computing Machinery.

Chen, M.; Radford, A.; Wu, J.; Jun, H.; Dhariwal, P.; Luan, D.; and Sutskever, I. 2020. Generative Pretraining From Pixels. https://api.semanticscholar.org/CorpusID:219781060. Accessed: 2024-11-02.

Chen, W.; Liu, C.; Wang, W.; Peng, M.; and Zhang, W. 2024. Adaptive Hybrid Beamforming for UAV mmWave Communications Against Asymmetric Jitter. *IEEE Transactions on Wireless Communications*, 23(8): 9432–9445.

Chen, Y.; Yan, L.; and Han, C. 2021. Hybrid Spherical- and Planar-Wave Modeling and DCNN-Powered Estimation of Terahertz Ultra-Massive MIMO Channels. *IEEE Transactions on Communications*, 69(10): 7063–7076.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; and Houlsby, N. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. arXiv:2010.11929.

Du, H.; Zhang, R.; Niyato, D.; Kang, J.; Xiong, Z.; Cui, S.; Shen, X.; and Kim, D. I. 2023. User-Centric Interactive AI for Distributed Diffusion Model-based AI-Generated Content. arXiv:2311.11094.

Heng, Y.; Mo, J.; and Andrews, J. G. 2022. Learning Site-Specific Probing Beams for Fast mmWave Beam Alignment. *IEEE Transactions on Wireless Communications*, 21(8): 5785–5800.

Hu, R.; Hao, C.; Zhang, Y.; Yoo, T.; Namgoong, J.; and Xu, H. 2023. Deep Learning-Based Channel Estimation with Low-Density Pilot in MIMO-OFDM Systems. In *ICC 2023 - IEEE International Conference on Communications*, 2619–2624. Rome, Italy: IEEE.

Jiang, W.; and Schotten, H. D. 2020. Deep Learning for Fading Channel Prediction. *IEEE Open Journal of the Communications Society*, 1: 320–332.

Jin, M.; Wang, S.; Ma, L.; Chu, Z.; Zhang, J. Y.; Shi, X.; Chen, P.-Y.; Liang, Y.; Li, Y.-F.; Pan, S.; and Wen, Q. 2024. Time-LLM: Time Series Forecasting by Reprogramming Large Language Models. arXiv:2310.01728.

Jongsintawee, S.; Runraengwajiake, S.; Supnithi, P.; and Panachart, C. 2016. Improvement of GPS positioning accuracy when utilizing Klobuchar model with ionospheric conditions in Thailand. In *2016 13th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON)*, 1–5. Chiang Mai, Thailand: IEEE.

Kim, S.; Lee, A.; Ju, H.; Ngo, K. A.; Moon, J.; and Shim, B. 2023. Transformer-Based Channel Parameter Acquisition for Terahertz Ultra-Massive MIMO Systems. *IEEE Transactions on Vehicular Technology*, 72(11): 15127–15132.

Li, Q.; and Wang, T. 2023. Application of Computer Controlled UAV HD Image Technology in Sports Broadcast. In *2023 IEEE 3rd International Conference on Data Science and Computer Application (ICDSCA)*, 1586–1590. Dalian, China: IEEE.

Liu, B.; Liu, X.; Gao, S.; Cheng, X.; and Yang, L. 2024. LLM4CP: Adapting Large Language Models for Channel Prediction. *Journal of Communications and Information Networks*, 9(2): 113–125.

Liu, F.; Zhang, J.; Jiang, P.; Wen, C.-K.; and Jin, S. 2023. CE-ViT: A Robust Channel Estimator Based on Vision Transformer for OFDM Systems. In *GLOBECOM 2023 - 2023 IEEE Global Communications Conference*, 4798–4803. Kuala Lumpur, Malaysia: IEEE.

Mukhamediev, R. I.; Yakunin, K.; Aubakirov, M.; Assanov, I.; Kuchin, Y.; Symagulov, A.; Levashenko, V.; Zaitseva, E.; Sokolov, D.; and Amirgaliyev, Y. 2023. Coverage Path Planning Optimization of Heterogeneous UAVs Group for Precision Agriculture. *IEEE Access*, 11: 5789–5803.

Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; and Sutskever, I. 2019. Language Models are Unsupervised Multitask Learners. https://api.semanticscholar.org/CorpusID:160025533. Accessed: 2024-11-02.

Roberts, W.; Stoica, P.; Li, J.; Yardibi, T.; and Sadjadi, F. A. 2010. Iterative Adaptive Approaches to MIMO Radar Imaging. *IEEE Journal of Selected Topics in Signal Processing*, 4(1): 5–20.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2023. Attention Is All You Need. arXiv:1706.03762.

Wang, W.; and Zhang, W. 2022. Jittering Effects Analysis and Beam Training Design for UAV Millimeter Wave Communications. *IEEE Transactions on Wireless Communications*, 21(5): 3131–3146.

Woo, S.; Park, J.; Lee, J.-Y.; and Kweon, I. S. 2018. CBAM: Convolutional Block Attention Module. arXiv:1807.06521.

Xiao, Z.; Zhu, L.; Liu, Y.; Yi, P.; Zhang, R.; Xia, X.-G.; and Schober, R. 2022. A Survey on Millimeter-Wave Beamforming Enabled UAV Communications and Networking. *IEEE Communications Surveys Tutorials*, 24(1): 557–610.

Xie, D.; Liu, C.; Wang, W.; Hu, X.; and Peng, M. 2024. Deep Residual Learning for Channel Estimation in UAV mmWave Systems: A Model-Driven Approach. In *2024 IEEE/CIC International Conference on Communications in China (ICCC)*, 1009–1014.

Xu, L.; Feng, L.; and Li, W. 2023. CTGAN-assisted CNN for high-resolution wireless channel delay estimation. In *2023 IEEE 24th International Conference on High Performance Switching and Routing (HPSR)*, 1–8. Albuquerque, NM, USA: IEEE.

Yang, Y.; Sun, J.; Li, H.; and Xu, Z. 2020. ADMM-CSNet: A Deep Learning Approach for Image Compressive Sensing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(3): 521–538.

Ye, Z.; Wang, K.; Chen, Y.; Jiang, X.; and Song, G. 2023. Multi-UAV Navigation for Partially Observable Communication Coverage by Graph Reinforcement Learning. *IEEE Transactions on Mobile Computing*, 22(7): 4056–4069.

Zaliva, V.; and Franchetti, F. 2014. Barometric and GPS altitude sensor fusion. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 7525–7529. Florence, Italy: IEEE.

Zhang, R.; Du, H.; Liu, Y.; Niyato, D.; Kang, J.; Xiong, Z.; Jamalipour, A.; and Kim, D. I. 2024. Generative AI Agents with Large Language Model for Satellite Networks via a Mixture of Experts Transmission. *IEEE Journal on Selected Areas in Communications*, 1–1.

Zhao, J.; Liu, J.; Gao, F.; Jia, W.; and Zhang, W. 2021. Gridless Compressed Sensing Based Channel Estimation for UAV Wideband Communications With Beam Squint. *IEEE Transactions on Vehicular Technology*, 70(10): 10265–10277.

Zhou, D.; Kang, B.; Jin, X.; Yang, L.; Lian, X.; Jiang, Z.; Hou, Q.; and Feng, J. 2021. DeepViT: Towards Deeper Vision Transformer. arXiv:2103.11886.