User Identification: The Key Enabler for Multi-User Vision-Aided Wireless Communications

Gouranga Charan and Ahmed Alkhateeb
School of Electrical, Computer and Energy Engineering - Arizona State University
Emails: {gcharan, alkhateeb}@asu.edu

Abstract—Vision-aided wireless communication is attracting increasing interest and finding new use cases in various wireless communication applications. These vision-aided communication frameworks leverage the visual data (captured, for example, by cameras installed at the infrastructure or mobile devices) to construct some perception about the communication environment (geometry, users, scatterers, etc.). This is typically achieved through the use of deep learning and advances in computer vision and visual scene understanding. Prior work has investigated various problems such as vision-aided beam, blockage, and hand-off prediction in millimeter wave (mmWave) systems and vision-aided covariance prediction in massive MIMO systems. This prior work, however, has focused on scenarios with a single object (user) moving in front of the camera. To enable vision-aided wireless communication in practice, however, it is important for these systems to be able to operate in crowded scenarios with multiple objects in the visual scene.

In this paper, we define the user identification task as the key enabler for realistic vision-aided wireless communication systems that can operate in crowded scenarios and support multi-user applications. The objective of the user identification task is to identify the target communication user from the other candidate objects (distractors) in the visual scene. We develop machine learning models that process either one frame or a sequence of frames of visual and wireless data to efficiently identify the target user in the visual/communication environment. Using the large-scale multi-modal sense and communication dataset, DeepSense 6G, which is based on real-world measurements, we show that the developed approaches can successfully identify the target users with more than 97% accuracy in realistic settings. This paves the way for scaling the vision-aided wireless communication applications to real-world scenarios and practical deployments.

Index Terms—Millimeter-wave, user identification, sensing, camera, deep learning, computer vision.

I. INTRODUCTION

Communication over millimeter wave (mmWave) and sub-terahertz (sub-THz) bands plays a fundamental role in satisfying the high data rate requirements in 5G and beyond [1], [2]. These systems, however, need to deploy large antenna arrays and use narrow beams at the transmitters and receivers to guarantee sufficient receive signal power. Selecting the optimal beams for these large antenna arrays is typically associated with significant training overhead, which makes it difficult for these mmWave/THz communication systems to support highly-mobile wireless applications such as virtual/augmented reality and connected vehicles. Furthermore, these high-frequency signals rely on line-of-sight (LOS) links to achieve sufficient receive signal power. Blocking these LOS links by the moving objects in the environment may disconnect the communication session or cause sudden and significant degradation in the link quality. This is due to the high penetration loss of the mmWave/sub-terahertz signals resulting in a severe reduction of receive power for NLOS links.

Leveraging machine learning (ML) to address these challenges has gained increasing interest in the last few years [4]–[8]. The role of machine learning (and artificial intelligence in general) in tackling problems such as beam training overhead, the sensitivity of mmWave/sub-THz signals to blockages, and demands for low-latency communications has been first investigated using only wireless signals. These solutions, however, are limited in their ability to scale to complex/crowded, or realistic scenarios. This motivated the development of machine learning-based approaches that leverage side information to overcome the challenges associated with the mmWave/sub-THz communication systems. In order to predict blockages early enough, i.e., before they block the links, solutions based on vision, radar, and LiDAR sensory data were proposed for the first time in [9]–[12]. Similarly, for fast mmWave/sub-THz beam prediction, solutions based on vision, position, radar,
and LiDAR were proposed in [13]–[18]. These sensing-aided wireless communication solutions were developed, however, for single-candidate scenarios and might not scale to a real-world setting with multiple objects in the environment. An important question that arises then is how do we develop sensing-aided wireless communication solutions that can scale to real-world scenarios with multiple objects in the sensing scene?

To enable sensing-aided communication systems in real-world settings, we need to enable these systems to operate in multi-candidate and multi-user settings. To illustrate this, consider the example of a sensing-aided beam prediction task. In practice, and from the basestation perspective, there can be multiple relevant objects in the wireless environment. Any of these objects can be the object of interest (the user). Therefore, the machine learning models must demonstrate a deep understanding of the wireless environment to be able to predict the optimal beam indices correctly. In particular, it needs to identify the probable transmitting candidate among the different objects in the environment. One main approach to achieve that is by leveraging additional sensing information (attributes) for the user. This brings the following important question: How can machine learning models leverage additional sensing data such as position or wireless receive power to identify the target user in the sensing (e.g., visual) scene?

In this paper, we focus on visual sensing and attempt to answer this question. The main contributions of the paper can be summarized as follows:

- Formulating the user identification problem in vision-aided mmWave/THz wireless communication networks considering practical visual and communication models.
- Developing machine learning approaches that are capable of (i) detecting the objects of interest in the wireless environment and (ii) efficiently identifying the user in the visual scene among the different objects in the environment.
- Providing the first real-world evaluation of sensing-aided transmitter identification based on our large-scale dataset, DeepSense 6G [19], that consists of co-existing multimodal sensing and wireless communication data.

Based on the adopted real-world dataset, the developed solution achieves ≈ 97% and ≈ 99% transmitter identification accuracy for input sequence lengths of 1 and 5, respectively. This highlights the potential of leveraging machine learning and sensing data in addressing the critical task of identifying the transmitter in the scene. In particular, the ability to identify the transmitter in the scene enables the network to make proactive beam/basestation switching decisions and predict future line-of-sight link blockage, enhancing the overall network reliability and latency performance.

II. SENSING-AIDED TRANSMITTER IDENTIFICATION: SYSTEM MODEL AND PROBLEM FORMULATION

The utilization of additional sensing data has shown great potential for 5G and beyond wireless communication systems and can help overcome some of the significant challenges associated with them. However, some fundamental challenges still need to be investigated to develop and implement sensing-aided solutions in the real world. One such real challenge is the ability of the communication system to distinguish between objects transmitting radio signals in the wireless environment (hereafter referred to as transmitters) and non-transmitting objects (referred to as the distractors). This ability to identify the objects of interest or the transmitters in the wireless environment is referred to as the transmitter identification task. In this work, the transmitter identification task is poised and studied in a mmWave communication setting. In this section, we first present the adopted wireless communication system model in Section II-A and then formulate the sensing-aided transmitter identification problem in Section II-B.

A. System Model

This work considers a communication scenario where a mmWave basestation is serving a mobile user located in a busy environment with multiple moving objects such as other vehicles, pedestrians, etc. The adopted system model comprises of a basestation equipped with an $M$-element uniform linear array (ULA) and an RGB camera, operating at a mmWave frequency band. The mmWave basestation is serving a mobile user (transmitter) that is, for simplicity, considered to have a single antenna. The adopted communication system employs OFDM transmission with $K$ subcarriers and a cyclic prefix of length $D$. The basestation employs a pre-defined beamforming codebook $\mathcal{F} = \{ f_q \}_{q=1}^Q$, where $f_q \in \mathbb{C}^{M \times 1}$ and $Q$ is the total number of beamforming vectors. Let $h_k[t] \in \mathbb{C}^{M \times 1}$ denotes the channel between the basestation and the user at the $k$th subcarrier and time $t$. If the basestation uses the beamforming vector $f_q[t] \in \mathcal{F}$ to serve the user, the receive signal can be expressed as

$$y_k[t] = h_k^T[t] f_q[t] x + n_k[t], \quad (1)$$

where $n_k[t]$ is a noise sample drawn from a complex Gaussian distribution $\mathcal{N}_c(0, \sigma^2)$. The transmitted complex symbol $x \in \mathbb{C}$ need to satisfy the following constraint $\mathbb{E}[|x|^2] = P$, where $P$ is the average symbol power. The beamforming vector $f^*[t] \in \mathcal{F}$ at each time step $t$ is selected to maximize the average receive SNR and is defined as

$$f^*[t] = \underset{f_q[t] \in \mathcal{F}}{\operatorname{argmax}} \frac{1}{K} \sum_{k=1}^{K} \text{SNR}[h_k^T[t] f_q[t]]^2, \quad (2)$$

where SNR is the transmit signal-to-noise ratio. At any time instant $t$, the receive power vector of effective channel gain with codebook elements can therefore be expressed as $p[t] = [p_1[t], \ldots, p_Q[t]]$, where $p(t) \in \mathbb{R}^{Q \times 1}$ and $p_q[t]$ is defined as

$$p_q[t] = |h_k^T[t] f_q[t]|^2, \quad q \in \{1, \ldots, Q\}. \quad (3)$$

B. Problem Formulation

Given the system model in Section II-A we provide the formal definition of the sensing-aided transmitter identification task in this section. For this, a general description of the
Transmitter identification is a multimodal machine learning task with the primary objective of identifying the transmitter among the different objects present in the wireless environment. The inputs to the machine learning model are the available sensing and wireless data obtained from the environment. We propose to observe a sequence of RGB images of the wireless environment captured by the camera installed at the basestation and utilize the sensing data along with the mmWave receive power vectors to identify the transmitter in the scene. The wireless channel vector $h$ (as defined in Section II-A), in general, encodes more detailed information regarding the wireless environment, such as the different propagation paths between the transmitter and the receiver; making it a better alternative for the task as compared to the receive power vector. Nevertheless, in the mmWave communication system, it is challenging to obtain this channel information analytically.

The transmitter identification task can be formally defined as follow. Let $X[t] \in \mathbb{R}^{W \times H \times C}$ denote a single RGB image of the environment captured at the basestation at time instant $t$, where $W$, $H$, and $C$ are the width, height, and the number of color channels for the image. Further, let $p[t]$ denote the mmWave receive power vector at the basestation. At any time instant $\tau \in \mathbb{Z}$, the basestation captures a sequence of RGB images and the mmWave receive power vectors, $S[\tau]$, defined as

$$S[\tau] = \{X[t], p[t]\}_{t=\tau-r+1}^{\tau}, \quad (4)$$

where $r \in \mathbb{Z}$ is the length of the input sequence or the observation window to identify the transmitter. In particular, at any given time instant $\tau$, the goal in this work is for the basestation to observe the sequence of data samples $S[\tau]$ to predict the bounding-box vector $b_{Tx}[\tau] \in \mathbb{R}^2$ corresponding to the transmitter in the image samples. In order to identify the transmitter, we define a function $f_\Theta$ that maps the observed sequence of data samples, $S[\tau]$ to a prediction (estimate) of the bounding-box vector, $b_{Tx}[\tau]$. The function $f_\Theta$ can be formally expressed as

$$f_\Theta : S[\tau] \rightarrow b_{Tx}[\tau]. \quad (5)$$

In this work, we develop a machine learning model to learn this prediction function $f_\Theta$, that takes in the observed sequence of data samples $S[\tau]$ and predicts the bounding box of the transmitter $b_{Tx}[\tau]$. Let $D = \{(S, b_{Tx})\}_u^{U}$ represent the dataset of independent samples consisting of sensing data-bounding box vector pairs collected from the real wireless environment, where $U$ is the total number of samples in the dataset. The prediction function is parameterized by $\Theta$ representing the model parameters. The dataset $D$ of labeled samples is then utilized to optimize the prediction function $f_\Theta$ such that it maintains high fidelity for any samples drawn from this dataset. The objective of the optimization function is to maximize the number of correct predictions over all the samples in the dataset $D$. The optimization problem can be formally written as

$$f^*_\Theta = \arg\max_{f_\Theta(\cdot)} \prod_{u=1}^{U} \mathbb{P}(b_{Tx,u} = b_{Tx,u} | S_u), \quad (6)$$

where the joint probability distribution in (6) is due to the implicit assumption that the samples on $D$ are drawn from an independent and identical distribution (i.i.d.). In the next section, we present our proposed machine learning-based solution for the sensing-aided transmitter identification task.

III. SENSING-AIDED TRANSMITTER IDENTIFICATION: A DEEP LEARNING SOLUTION

In this section, we present an in-depth overview of the proposed sensing-aided transmitter identification solution. First, we present the key idea in Section III-A and then explain the details of our proposed solution in Section III-B and Section III-C

A. Key Idea

With their large bandwidth, the mmWave/sub-THz communication systems can satisfy the high data rate requirements of several current and future applications. However, communication in these bands is faced with several challenges. One major challenge arises from the high sensitivity of the mmWave/sub-THz signals to blockages. For this, high-frequency signals suffer from significant penetration loss and primarily rely on line-of-sight (LOS) communication. The high-frequency signals, further, suffer from severe path loss. To overcome this huge path loss, the mmWave/sub-THz communication systems must deploy large antenna arrays and use narrow directed beams to guarantee a sufficient receive signal-to-noise ratio (SNR). This dependence of the mmWave/sub-THz systems on LOS links and the usage of directive radiation patterns form the basic building block of our proposed transmitter identification solution.

The directivity of antenna arrays can be visualized as a way of concentrating the emitted radiation in a single direction. For ULAs, this directivity is achieved by the beamforming vectors in the pre-defined codebook $\mathcal{F}$. The beamforming vectors can be envisioned as slicing the scene (spatial dimension) into multiple (possibly overlapping) sectors, where each sector is associated with a particular beam value. This sectoring of the wireless environment by the beamforming vectors can be extended to a visual scene. Note that the RGB image is merely a projection of the 3D space onto a 2D image plane. The sectoring induced by the beamforming vectors can then be projected onto the 2D image plane, resulting in the form of image sectoring. Therefore, the knowledge of the optimal beamforming vector or the receive power vector, in general, can be translated to directional information in an image, i.e., the direction from which the current received signal arrived.

Furthermore, the recent advancements in machine learning and computer vision have enabled several new capabilities, such as object detection, multi-object tracking, and image segmentation, to name a few. Therefore, utilizing state-of-the-art object detection models makes it possible to identify
is one data sample enough for
question that arises now
the transmitters with just one data sample is possible. The
scene.
between the transmitting objects from the distractors in the
receive power vectors can enable us to differentiate
capabilities paired with the directional information obtained
in near real-time. The fast and accurate object detection
different objects in the wireless environment with high fidelity
in near real-time. The fast and accurate object detection
capabilities paired with the directional information obtained
from the receive power vectors can enable us to differentiate
between the transmitting objects from the distractors in the
scene.

Based on the idea proposed above, theoretically, detecting
the transmitters with just one data sample is possible. The
question that arises now is one data sample enough for
correctly identifying the transmitter? In order to answer
this question, we first need to understand the challenges
associated with this approach. There are primarily three main
challenges: (i) Object detection models are not perfect. There
is a possibility that several objects, including the transmitters,
might not be detected, which might result in false detection.
(ii) The transmitter can be partially or entirely occluded in a
particular instance. Relying on just that one sample to identify
the transmitter will result in a wrong prediction. (iii) One
of this solution’s key components is identifying the user’s
approximate location in the scene by utilizing the optimal
beamforming or receive power vectors. It is essential to point
out here that generating sharp and directive beams with no side
lobes are challenging due to the non-idealities and impairments
in the hardware. Such hardware limitations will result in
non-ideal sectoring and lead to a distractor being labeled
as a transmitter. One possible solution to overcome these
challenges is to observe a sequence of data samples (image
and wireless data) to determine the transmitter accurately. At
any given instant $\tau$, by observing a sequence of $r$ current and
previous samples, we inherently reduce the effect of non-ideal
sectoring and the probability of missed detection. Therefore,
in conclusion, although it might be possible to detect the
transmitter using just one pair of image-wireless data, it is
imperative to observe a sequence of data samples to increase
the probability of correctly identifying the transmitter in the
environment. In this work, we propose to utilize a sequence of
$\tau$ image and wireless data samples to predict the transmitter
in the scene with high fidelity.

B. A Single Sample-based Approach

This subsection presents the proposed solution for transmit-
ter identification in a real-wireless environment with multiple
candidates. It proposes a novel approach that utilizes bimodal
visual and wireless data in $D$ to identify the transmitter in
the scene. We will first present the solution for identifying
the transmitter using one data sample and then extend the
proposed solution for a sequence-based approach. A three-
step architecture is proposed for the single data sample-
based transmitter identification task. The first step of the
proposed architecture relies on DNNs to produce bounding
boxes enclosing relevant objects in the scene. It is performed to
detect all the probable transmitting objects in the environment.
In the second step, the DNN uses wireless data to predict the
probable bounding-box centers of the transmitting candidate.
The last step involves filtering detected candidates that are not
the radio transmitter. An in-depth overview of the three-step
DNN architecture is provided below. The detailed solution for
single data sample-based transmitter identification is presented
in Fig. 2.

(i) Bounding box detection: In order to detect the trans-
mittting candidate in real-wireless settings, the first step is to
identify all the relevant objects in the scene (scene analysis).
A pre-trained object detector is adopted for this purpose.
The object detector is modified to detect two classes of
objects in the scene, labeled as “Tx (transmitter)” and “No
Tx (Distractors)”. The former label encompasses all objects
that are of relevance to the wireless system in the scene. For
example, in a scene depicting a city street, relevant objects
include, but are not limited to, cars, trucks, buses, pedestrians,
and cyclists. The other label includes those cases where no
relevant objects are present in the scene. The modified object
detector is fine-tuned in a supervised fashion using a subset
of the manually labeled dataset described in Section IV-B.

A YOLOv3 architecture is selected for the bounding box

Fig. 2. The figure presents the proposed single sample-based transmitter identification model that leverages both visual and wireless data to predict the
transmitter in the scene.
During inference, the fine-tuned YOLOv3 model generates bounding boxes for the detected candidates in the scene and their confidence scores. By using those output bounding boxes, the relevant-object matrix $B \in \mathbb{R}^{N \times 2}$ is constructed such that each row has only the normalized coordinates of the center of a bounding box, where $N$ is the number of relevant objects in the scene.

(ii) **Bounding box center prediction:** In this step, both relevant-object matrix $B$ and wireless receive power vector are utilized to predict the bounding box center coordinates of the transmitter. This step involves learning a prediction function that estimates the bounding box center of a transmitter using the receive power vector. The primary objective is to encode the relation between the receive power vector and object location in the image. The function is learned using a 2-layered feed-forward neural network

$$f_{\Theta_2} : r[t] \rightarrow \hat{b}_{Tx}[t],$$

where $\hat{b}_{Tx}[t] \in \mathbb{R}^{2 \times 1}$ is a vector with an initial prediction of the centers of the transmitting candidate and the $r[t] \in \mathbb{R}^{Q \times 1}$ is the mmWave receive power vector at any time instant $t$. Let $\mathcal{D}_2 = \{ (r, b_{Tx}) \}_{u=1}^{U}$ and $\mathcal{D}_2 \subset \mathcal{D}$ be a dataset comprising of the mmWave receive power vectors and the ground-truth bounding box center coordinates of the transmitter. The prediction function $f_{\Theta_2}$ is parameterized by a set $\Theta_2$ representing the model parameters and learned from the dataset $\mathcal{D}_2$ of the labeled data samples. Since $\hat{b}_{Tx}$ is an initial estimate that solely relies on position data, it is not expected to be a final prediction but merely an estimate. The idea here is to use the initial estimate in conjunction with the relevant-object matrix $B$ to identify (or select) the object responsible for the radio signal.

(iii) **Bounding box selection:** The previous two steps, i.e., bounding box detection and the bounding box center prediction, provide two vital pieces of information: (i) The first stage helps identify all the objects in the environment. More specifically, it outputs the relevant-object matrix $B$ comprising the bounding box coordinates of all the objects of interest (probable transmitters) in the wireless environment. (ii) In step two, using the additional modality, i.e., the wireless receive power vector, we can estimate the approximate center coordinates of the transmitter in the scene. We can then utilize these two pieces of information to identify the transmitter in the scene accurately. This is done using the nearest neighbor algorithm with a Euclidean distance metric. We first calculate the Euclidean distance between the predicted center coordinates and all the objects in $B$. The element of $B$ with the shortest distance to $\hat{b}_{Tx}$ is picked as the nearest neighbor and, hence, the predicted transmitter object. The assumption here is that a well-trained prediction function $f_{\Theta_2}$ can predict the center coordinates close to the actual values, and hence the Euclidean distance-based metric can help in accurately detecting the transmitter.

### C. A Sequence-based Approach

The three-step solution proposed in Section III-B can help identify the transmitter from one data sample. The question that now arises is how do we extend this solution to determine the probable transmitters from a sequence of data samples? It is essential to note here that, in this work, we do not consider the no-transmitter scenario; in other words, data collected at every time step $t$ will have a transmitter present in the wireless environment. The underlying principle of the sequence-based transmitter identification task is as follows: Instead of relying on just one data sample to identify the transmitter, the proposed approach observes a sequence of $r$ data samples. The object identified as a transmitter the maximum number of times in these $r$ consecutive samples is tagged as the transmitting candidate. However,
moving from a single-sample-based solution to a sequence-based approach has its own challenges. This primarily arises from the difficulty in ensuring that the objects detected as transmitters in two consecutive time steps are the same. Given a sequence of \( r \) image samples, the transmitter should be present in every image. Furthermore, as mentioned in Section [II-A], we consider a mobile transmitter in this work. Therefore, the location of the transmitter and other objects in the wireless environment and the RGB image is not fixed across these \( r \) consecutive samples. Therefore, to truly perform sequence-based transmitter identification, we also need to track all the relevant objects through time, in addition to identifying which of these objects is the transmitter in the scene. The proposed sequence-based transmitter identification solution comprises three steps: (i) Object association-based tracking, (ii) transmitter identification, and (iii) maximum probability-based identification.

(i) Object association-based tracking: Multiple object tracking (MOT) is an active research area, and several state-of-the-art algorithms have been proposed. This work focuses on vehicle-to-infrastructure communication, with the primary object of interest being mobile vehicles. Therefore, in this work, we develop a simple distance-based tracking algorithm instead of utilizing state-of-the-art MOT algorithms. We adopt a Euclidean distance-based measurement technique, similar to the bounding box selection step described in Section [III-B]. The first stage of the proposed object association-based tracking algorithm is to detect the different objects of interest across the \( r \) image samples in the sequence and extract the bounding box center coordinates. Let us assume that there are \( N_1 \) and \( N_2 \) detected objects in the first and second image of the sequence with different objects labeled from 1, \ldots, \( N_1 \) for the first image and labeled 1, \ldots, \( N_2 \) for the second image. Now there are two possibilities: (i) Same number of objects in two consecutive image sample, i.e., \( N_1 = N_2 \) and (ii) different number of objects, i.e., either \( N_1 > N_2 \) or \( N_1 < N_2 \). In the second stage of the algorithm, we calculate the Euclidean distance between each detected object in the first image and the objects in the second image. The objects in the second image are then re-numbered based on this calculated distance. For example, if the 3rd object in the first image has the shortest Euclidean distance with the 1st detected object in the second image, then this object is re-numbered as 3. The principle behind this algorithm is that for two consecutive image samples, the distance between the bounding box center coordinates will be the least for the same object compared to other objects in the scene. Therefore, this algorithm helps identify and track the different objects across the \( r \) image samples.

(ii) Transmitter identification: This step is similar to the single-sample-based transmitter identification described in Section [II-B] and is performed for all the \( r \) data samples in the sequence. In combination with step one, i.e., object association-based tracking, for each data sample in the sequence, the output of this step is \( \hat{N} \in \mathbb{R}^1 \), denoting the index of the transmitter in the scene.
### Table I
**Number of Data Sequences in the Development Dataset**

| Number of Objects | Scenario 3 Training | Scenario 3 Validation | Scenario 4 Training | Scenario 4 Validation |
|-------------------|---------------------|-----------------------|---------------------|-----------------------|
| 1                 | 376                 | 140                   | 417                 | 187                   |
| 2                 | 291                 | 86                    | 325                 | 140                   |
| 3                 | 140                 | 46                    | 182                 | 79                    |
| 4                 | 61                  | 28                    | 83                  | 30                    |
| 5                 | 32                  | 12                    | 27                  | 12                    |
| 6                 | 6                   | 7                     | 6                   | 9                     |
| 7                 | 0                   | 3                     | 7                   | 0                     |

(iii) **Maximum probability-based identification** The final step of the proposed sequence-based solution is to detect the transmitter based on the observed sequence data accurately. The input to this step is the vector of \( r \) indices, i.e., \( \{N_1, \ldots, N_r\} \) that are obtained by performing object association-based tracking and transmitter identification in step one and two, respectively. The object that has been identified as the transmitter most often across the \( r \) data samples is finally identified as the transmitting candidate in the scene.

### IV. Testbed Description and Development Dataset

In order to evaluate the performance of the proposed sensing-aided user identification solution, we utilize the **DeepSense 6G** [19] dataset. DeepSense 6G is a real-world multi-modal dataset for sensing-aided wireless communication applications. It contains co-existing multi-modal data such as vision, mmWave wireless communication, GPS data, LiDAR, and radar collected in a real-wireless environment. This section presents a brief overview of the scenario adopted from the DeepSense 6G dataset, followed by the analysis of the final development dataset utilized for evaluating the proposed solution.

#### A. DeepSense 6G: Testbed 1

This study adopts different scenarios of the DeepSense 6G dataset specifically designed to study high-frequency wireless communication applications in a multi-candidate setting. The hardware testbed and the locations for collecting these data are shown in Fig. 4. The DeepSense testbed 1 is utilized for this data collection consisting of (i) a stationary unit (acting as the base station) and (ii) a mobile transmitter (vehicle). The stationary unit \{unit1 (RX)\} is equipped with a standard-resolution RGB camera and mmWave Phased array. The stationary unit adopts a 16-element \((M = 16)\) 60GHz-band phased array and it receives the transmitted signal using an over-sampled codebook of 64 pre-defined beams \((Q = 64)\). The mmWave phased array and the RGB camera are placed such that their fields of view are aligned. In this data collection scenario, the mobile unit \{unit2 (TX)\} is a vehicle equipped with a mmWave transmitter and GPS antenna/receiver. The transmitter consists of a quasi-omni antenna constantly transmitting (omnidirectional) at the 60 GHz band. Each data sample consists of an RGB image of the wireless environment and a 64-element mmWave receive power vector. For more information regarding the data collection setup and testbed, please refer to [19].

#### B. DeepSense 6G: Development Dataset

This work utilizes scenarios 1, 3, and 4 of the DeepSense 6G dataset. The adopted DeepSense scenarios include diverse data collected at different locations and during different times of the day (day and night). In particular, scenarios 3 and 4 are collected at the same location (Rural Rd., Tempe) but at different times of the day. Scenario 1 is collected at a different location (McAllister Ave., Tempe) and primarily consists of data collected during the day time. In Figure 5, we present sample dataset images from scenarios 1, 3, and 4, which highlights the diversity in these scenarios. At any given time instant, \( t \), the multi-modal scenario dataset comprises the following: An RGB image, \( X_t \), the corresponding receive power vector \( r_t \) and the user position. We further generate the ground-truth bounding box center coordinates of the transmitter in the scene, \( b_{TX,t} \) (manually labeled). To form the development dataset of the user identification prediction task described in Section II-B, the offered DeepSense data is further processed using a sliding window to generate a time-series dataset consisting of 1, 3, and 5 input data images \((r = 1, 3, \text{and} 5)\) and the corresponding mmWave receive power. The final step in the processing pipeline is dividing the dataset into training and test sets following a 70 – 30% split. In Table 2, we present the details of the development datasets for the sensing-aided transmitter identification task.

The main objective of this work is to develop a multi-modal user-identification solution. To evaluate the efficacy of the proposed sensing-aided user-identification solution, we first utilize the development datasets of scenarios 3 and 4. The proposed solution is trained and tested on the development dataset of scenarios 3 and 4. Given that these scenarios are collected at the same location, but at different time of the day, it also helps to investigate the proposed solution’s ability to generalize across different input data distribution. For this, we train the model with the labeled dataset of one of the scenarios and test on the dataset of the other scenario. Next, to analyze the model’s ability to adapt to unseen dataset, we utilize the scenario 1 development dataset. It involves training the

### Table II
**Design and Training Hyper-parameters**

| Parameters                      | MLP |
|---------------------------------|-----|
| Batch Size                      | 32  |
| Learning Rate                   | \( \times 10^{-3} \) |
| Learning Rate Decay             | epochs 80 and 120 |
| Learning Rate Reduction Factor  | 0.1 |
| Dropout                         | 0.3 |
| Total Training Epochs           | 150 |
Fig. 5. In this figure, we present the different image sample of scenarios 1, 3, and 5. Figures (a), (b), and (c) are from scenarios 3 and 4. It shows the different lighting conditions (day, dusk, and night) in which the dataset was collected, highlighting the diversity in these scenarios. In figures (d), (e), and (f), we present the image samples from scenario 1. It highlights the difference between scenarios 3, 4 and scenario 1.

Fig. 6. This figure presents the user identification accuracy versus the input sequence length of 1, 3 and 5 for both scenarios 3 and 4. It is observed that observing a sequence of samples help in improving the identification accuracy.

V. EXPERIMENTAL SETUP:

In this section, we first discuss the neural network training parameters and the adopted evaluation metrics. As described in Section III, the proposed single sample-based transmitter identification solution comprises three steps. As part of the second step, i.e., bounding box center prediction, the mmWave receive power vectors are provided as input to the feed-forward neural network to predict the approximate bounding box center coordinates of the transmitter. The 2-layered feed-forward neural network is trained using the labeled development dataset discussed in Section IV-B using a cross-entropy loss function. All the simulations were performed on a single NVIDIA Quadro 6000 GPU using the PyTorch deep learning framework. The detailed design and training hyper-parameters are presented in Table II. We utilize the top-1 accuracy metric as the primary method of evaluating the proposed solution. The top-1 accuracy is defined as follows:

$$
Acc_{top-1} = \frac{1}{U} \sum_{u=1}^{U} \mathbb{I}\{\hat{s}_{u}[\tau] = s_{u}[\tau]\},
$$

where $\hat{b}_{TX,u}[\tau]$ and $b_{TX,u}[\tau]$ are the predicted and ground-truth bounding box center coordinates, respectively. $U$ is the total number of samples present in the validation/test set. $\mathbb{I}\{\}$ is the indicator function.

VI. PERFORMANCE EVALUATION

This section presents the detailed evaluation of the proposed sensing-aided transmitter identification solution.

A. Can visual and wireless data be utilized for transmitter identification?

To answer this question, we evaluate the proposed single sample-based transmitter identification solution on the development dataset of both scenarios 3 and 4 as described in
Section IV-B In particular, the proposed solution is trained and tested individually in both scenarios. In Fig. 6, we plot the achieved user identification accuracy for both the scenarios. It is observed that for the input sequence length of 1, the proposed solution achieved an accuracy of 98.43% and 97.16% for scenarios 3 and 4, respectively. The high accuracy of the proposed approach with just one observed sample highlights that sensing-aided solutions can enable user identification in a multi-candidate scenario. However, relying on just one sample to identify the user has its challenges. A key component of the proposed solution relies on the accurate detection of the objects of interest in the wireless environment. However, the state-of-the-art image-based object detection models have mean average precision (mAP) of 60–70%, which highlights that these models are imperfect. The inherent non-idealities of the real-world data further make it more challenging to detect objects accurately. This might result in the user itself not being detected, leading to errors in user identification. Another challenge arises when the user is blocked by other stationary and dynamic objects in the environment. By just relying on a single sample, it is not possible to detect the blocked user. A promising solution to overcome these limitations is to observe a sequence of image samples. It helps in increasing the probability of object detection and, in general, improving user identification accuracy. In the next subsection, we present the performance of the proposed user-identification solution on the sequence data.

B. Does observing a sequence of data samples help?

In order to evaluate the effect of sequence data, we constructed a time-series dataset with a window length of 3 and 5 for both the scenarios 3 and 4. Furthermore, we extended the single sample-based solution to the sequence-based transmitter identification solution as proposed in Section III-C. In Fig. 6 we present the transmitter identification accuracy versus the input sequence length for both scenarios. It is observed that increasing the input sequence length for both scenarios helped achieve better identification performance. In order to further investigate the impact of observing a sequence of data samples, we plot the number of objects versus the prediction accuracy for both scenarios in Fig. 7. It is important to highlight here that we have only plotted the performance up to 4 objects. This is primarily because the number of sequences with more than 4 objects is extremely small, as shown in Table I. Therefore, it is difficult to draw any meaningful conclusion from the performance of those sequences. In Fig. 7 we observe an interesting trend; as the number of objects increases, the sequence-based approach achieves better performance. The improved performance can be attributed to the fact that as the number of objects in the scene increases, the chances of missed object detection, etc., also increase. Therefore, relying on just one sample will lead to reduced identification accuracy. The results, therefore, validate our initial intuition that observing a sequence of input data samples should perform better than just observing one data sample.

C. Does variation in data distribution impact the model’s performance?

In Sections VI-A and VI-B we presented the user identification performance of the sensing-aided solution for single-sample and sequence-based approaches, respectively. It is observed that for these approaches, the proposed solution can identify the user with high fidelity. However, the proposed solution was trained and evaluated on the same scenario dataset in these experiments. Although such an experimental
D. Can the proposed solution adapt to unseen scenarios?

In Section VI-C, we investigate the proposed solution’s ability to generalize across different data distributions. Another challenge towards real-world deployment is the fast adaptation to unseen scenarios. Given that the 5G and beyond basestations will be deployed across different locations developing site-specific machine learning models, i.e., different models trained in a supervised fashion for each location, is not feasible. This is primarily due to the unavailability of such labeled datasets for each location. Overcoming this challenge necessitates the development of efficient solutions that can adapt quickly to an unseen location with few or no labeled data samples. To evaluate such adaptation capability of the proposed solution, we utilize scenarios 1 and 4 of the DeepSense 6G dataset. These two scenarios were collected at different locations and at different times of the day. As shown in Figure 9, scenario 1 consists of a 2-lane street, whereas scenario 4 is a 6-lane street. Further, the distance of the basestation from the street differs for these two locations. Such differences result in variations in the distribution of the mmWave receive power. All these make it highly challenging for any solution to adapt to an unseen scenario. In Figure 10, we present the achieved user identification accuracy for two cases: (i) The proposed ML model is trained and tested on one scenario dataset alone, and (ii) The model has trained on scenario 4 training dataset and evaluated on the test set of scenario 1. We observe a 7 – 9% drop in accuracy between the two cases, signifying how challenging this adaptation task is. The key takeaway is that even with no training data from scenario 1, the proposed solution can identify the users with ≈ 90% accuracy. Such a performance highlights the ability of the proposed solution to adapt to unseen scenarios.

E. Does user-speed impact the transmitter identification performance?

Given the dynamic nature of the dataset, the speed of each user (vehicle) varies with time. Therefore, it is important to consider the impact of vehicle speed on the transmitter identification accuracy. In order to calculate the user speed, we utilize the position of the user (available as part of the DeepSense 6G dataset). In particular, we estimate a user speed by considering the difference between the initial and final position in each sequence with 5 samples and divide it by 5. Further, we calculate the speed mean \( \tilde{\mu} \) and standard-deviation
Using $\tilde{\mu}$ and $\tilde{\sigma}$, we divide those users into three buckets: (i) slow-moving user with speeds less than or equal to $\tilde{\mu} - \tilde{\sigma}/2$; (ii) fast-moving user with speeds greater than or equal to $\tilde{\mu} + \tilde{\sigma}/2$; and (iii) average-speed user with speeds between those of slow- and fast moving users. In Fig. 11, the user identification accuracy versus the vehicle speed is presented. It is observed that for both scenarios 3 and 4, the slower moving users result in better user identification accuracy. However, in most of the samples, the difference in accuracy between the slow and fast moving user is very small. This further highlights the model’s ability to identify the user even for fast-moving vehicles with very high confidence.

**VII. Conclusion**

This paper explores the potential of leveraging visual and mmWave wireless data for identifying the probable transmitter in the wireless environment. It takes an essential step toward addressing the concern about the practicality of the sensing-aided wireless communication solution in real multi-object communication settings. It does so by (i) defining the novel transmitter identification task, (ii) proposing a deep learning-based solution, and (iii) extending the solution from a single sample-based approach to a sequence-based solution. The key takeaways of evaluating our proposed transmitter identification solution based on the large-scale real-world dataset, DeepSense 6G, can be summarized as follows: (i) Even with a single data sample, the proposed solution achieves $\approx 97\%$ transmitter identification accuracy. (ii) For data samples with more objects in the wireless environment, observing a sequence of previous data samples helps achieve better prediction accuracy than just the current data sample. These results highlight the potential gains of leveraging visual and wireless data in identifying probable transmitters in the wireless environment and place more emphasis on designing better algorithms to tap into the wealth of information in the input sensing data.

**REFERENCES**

[1] T. S. Rappaport, Y. Xing, O. Kanhere, S. Ju, A. Madmavake, S. Mandal, A. Alkhateeb, and G. C. Trichopoulous, “Wireless communications and applications above 100 GHz: Opportunities and challenges for 6G and beyond,” IEEE Access, vol. 7, pp. 78 729–78 757, 2019.

[2] R. W. Heath, N. Gonzalez-Prelcic, S. Rangan, W. Roh, and A. M. Sayeed, “An overview of signal processing techniques for millimeter wave MIMO systems,” IEEE Journal of Selected Topics in Signal Processing, vol. 10, no. 3, pp. 436–453, April 2016.

[3] J. G. Andrews, T. Bai, M. N. Kulkarni, A. Alkhateeb, A. K. Gupta, and R. W. Heath, “Modeling and analyzing millimeter wave cellular systems,” IEEE Transactions on Communications, vol. 65, no. 1, pp. 403–430, 2017.

[4] A. Alkhateeb, S. Alex, P. Varkey, Y. Li, Q. Qu, and D. Tujkovic, “Deep learning coordinated beamforming for highly-mobile millimeter wave systems,” IEEE Access, vol. 6, pp. 37 328–37 348, 2018.

[5] A. Alkhateeb, I. Beltagy, and S. Alex, “Machine learning for reliable mmWave systems: Blockage prediction and proactive handoff,” in *Proc. of IEEE GlobalSIP*, Nov 2018, pp. 1055–1059.

[6] S. H. Lim, S. Kim, B. Shim, and J. W. Choi, “Deep learning-based beam tracking for millimeter-wave communications under mobility,” IEEE Transactions on Communications, vol. 69, no. 9, pp. 7458–7469, 2021.

[7] M. Alrabeiah and A. Alkhateeb, “Deep learning for mmwave beam and blockage prediction using sub-6GHz channels,” IEEE Transactions on Communications, pp. 1–1, 2020.

[8] S. Wu, M. Alrabeiah, C. Chakrabarti, and A. Alkhateeb, “Blockage prediction using wireless signatures: Deep learning enables real-world demonstration,” IEEE Open Journal of the Communications Society, vol. 3, pp. 776–796, 2022.

[9] G. Charan and A. Alkhateeb, “Computer vision aided blockage prediction in real-world millimeter wave deployments,” 2022. [Online]. Available: https://arxiv.org/abs/2203.01907

[10] U. Demirhan and A. Alkhateeb, “Radar aided proactive blockage prediction in real-world millimeter wave systems,” in *Proc. of IEEE ICC*, arXiv preprint, arXiv:2111.14805, 2021.

[11] ———, “Integrated sensing and communication for 6G: Ten key machine learning roles,” 2022. [Online]. Available: https://arxiv.org/abs/2208.02157

[12] S. Wu, C. Chakrabarti, and A. Alkhateeb, “LiDAR-aided mixed blockage prediction in real-world millimeter wave systems,” in *Proc. of IEEE WCNC*, arXiv preprint, arXiv:2111.09581, 2021.

[13] G. Charan, T. Osman, A. Hredzak, N. Thawdar, and A. Alkhateeb, “Vision-position multi-modal beam prediction using real millimeter wave datasets,” in 2022 IEEE Wireless Communications and Networking Conference (WCNC), 2022, pp. 2727–2731.

[14] M. Alrabeiah, A. Hredzak, and A. Alkhateeb, “Millimeter wave base stations with cameras: Vision-aided beam and blockage prediction,” in 2020 IEEE 91st Veh. Technol. Conf. (VTC2020-Spring), 2020, pp. 1–5.

[15] M. Arvinte, M. Tavares, and D. Samardzija, “Beam management in 5G NR using geolocation side information,” in 2019 53rd Annual Conference on Information Sciences and Systems (CISS), 2019, pp. 1–6.

[16] Y. Wang, M. Narasimha, and R. W. Heath, “Towards robustness: Machine learning for mmWave V2X with situational awareness,” in 2018 52nd Asilomar Conference on Signals, Systems, and Computers, 2018, pp. 1577–1581.

[17] U. Demirhan and A. Alkhateeb, “Radar aided 6G beam prediction: Deep learning algorithms and real-world demonstration,” 2021. [Online]. Available: https://arxiv.org/abs/2111.09676

[18] S. Jiang, G. Charan, and A. Alkhateeb, “LiDAR aided future beam prediction in real-world millimeter wave V2I communications,” IEEE Wireless Communications Letters, Oct. 2022. [Online]. Available: https://arxiv.org/abs/2203.05548

[19] A. Alkhateeb, G. Charan, T. Osman, A. Hredzak, and N. Srinivas, “DeepSense 6G: A large-scale real-world multi-modal sensing and communication dataset,” available on arXiv, 2022. [Online]. Available: https://www.DeepSense6G.net