Enabling Coexistence of Indoor Millimeter-Wave Networking and Human Activity Sensing

Moh Sabbir Saadat; Sanjib Sur
Computer Science and Engineerng, University of South Carolina, Columbia, SC, USA
msaadat@email.sc.edu; sur@cse.sc.edu

Abstract*—We propose *mNetS*, a system that enables the coexistence of indoor millimeter-wave (mmWave) networking and sensing for at-home physical rehabilitation and continuous health diagnostics. Although mmWave signal has been shown to effectively capture human activities/posture, running networking and sensing applications simultaneously remains a challenge. Typically, sensing can only be performed opportunistically to minimize the impact on networking, resulting in low-rate data capture with irregular spacing that affects sensing accuracy. To overcome this challenge, *mNetS* leverages idle times within the data transfer process to perform sensing opportunistically. Next, using a customized Dynamic Graph Convolutional Neural Network (DGCNN), *mNetS* extracts relevant features from low-rate sensing samples that are adaptively combined and regressed to estimate high-rate sensing samples. To demonstrate the effectiveness of *mNetS*, we collect signal reflections from various human activities that involve diverse movements of different body parts. We train and test *mNetS* on these data samples, and our results show a significant improvement in estimating high-rate signal samples, qualitatively and quantitatively. Such high-rate signal samples estimated by *mNetS* improves the performance of a typical human activity sensing application such as the classification of activity markedly over reduced rate samples or high-rate samples estimated through a naive linear interpolation.

***Index Terms*—mmWave, Sensing, PCD, Graph Neural Networks**

I. Introduction

Continuous sensing of human activity at home has proved to be useful for numerous health applications, such as remote physical therapy, injury prevention, and detection of abnormal posture or gait. In-person appointments for physical therapy can be time-consuming and inconvenient, especially for patients with adverse physical health conditions. Moreover, fixed appointments may not be enough to diagnose certain health conditions or events. For example, a person's posture or gait may become increasingly abnormal over time. This could be an indicator for worsening physical fitness - increasing fatigue, reduced bone density, increasing obesity *etc.*- or, worse, an early onset of stroke. Such conditions may be hard to detect properly over a short diagnosis window. Sudden collapse at home is also a widely recurring health event, especially for senior citizens, which requires immediate detection and medical attention. Recovering patients with surgical wounds or injury may also require continuous monitoring due to the requirements of specific sleeping or sitting postures. Furthermore, the COVID-19 pandemic has highlighted the importance of systems which allow at home physical therapy or health diagnosis [1]–[7]. Camera-based systems can achieve high precision in such continuous human sensing, including estimation of 3D posture. However, there are two major problems with such systems. Firstly, cameras pose privacy issue due to capturing clear true-color images at home. Recently, there has been several reports of camera-based monitoring systems being hijacked by third-party adversaries [8]–[10] which makes such a camera-based approach highly unattractive for at home human sensing. Secondly, a camera-based system would require sufficient lighting condition which is impractical for at home continuous sensing.

Fortunately, next-generation wireless networking devices, such as 5G wireless routers [11], [12], operating at high-frequency millimeter-wave (mmWave), provide an opportunity to bring privacy non-invasive and low-cost human activity sensing systems to the masses which is also capable of working in any lighting condition. These mmWave networking devices are designed to offer multi-Gbps of throughput and sub-ms data transfer latency, enabling 5G-and-beyond applications. Also, existing research works have demonstrated the potential of mmWave wireless sensing for a range of applications, such as tracking and identifying human subjects [1], [2], [13] recognizing gait cycles [6], [7], estimating postures or silhouettes [14]–[16], or recognizing gestures [17]–[19]. The small wavelength and large bandwidth of the mmWave signals enable high-resolution monitoring of activities compared to the traditional Wi-Fi or LTE systems. Additionally, mmWave devices provide an advantage over camera-based systems at home, as wireless signals can work under dark environments, and preserves privacy. However, the design of mmWave sensing on networking devices presents two challenges.

First, although mmWave devices can serve as effective human activity sensors, simultaneously running sensing and networking applications is challenging. For instance, when a user moves in front of a mmWave streaming device, the Line-of-Sight (LOS) communication path may become disrupted, and redirecting the beam towards a Non-Line-of-Sight (NLOS) path can compromise both sensing accuracy and streaming quality. One solution to enable the coexistence of networking and sensing applications is to equip devices with specialized sensing hardware that operates on distinct mmWave spectrum segments or spatial regions to avoid interference. However, such an approach is not ideal for the widespread deployment of sensing applications on many existing, inexpensive mmWave devices currently in use or in development. *Second*, mmWave devices are vulnerable to more specular and variable reflectivity challenges (compared to Wi-Fi or LTE) due to their high-frequency operations. So, depending on the location, orientation, and absorption properties of objects and humans, some of the signals transmitted may not reach back to the device [20]. Consequently, the representation of signals in 3D can be challenging, leading to a loss of information about the target human shape.

To address these challenges, we propose mNetS, that enables coexistence of mmWave networking and at home human activity sensing. The key idea is to first translate the reflections of the mmWave signal into a 3D view with sparse cluster of points in a Point Cloud Data (PCD), and then use the PCD for human activity sensing. Prior research on mmWave sensing for human activity has relied on high sample acquisition rate of sensing radar. [6], [7], [14], [15], [17]. In contrast, *mNetS* explores sensing with reduced sample acqusition rate due to the need for time sharing sensing with networking. To address this challenge, *mNetS* leverages opportunistic idle times within the data transfer process for sensing. As a result, only partial temporal observations are available, necessitating the use of a deep learning model to recover the missing information over time. To this end, *mNetS* designs a set of feature extraction modules to extract relevant high-dimensional features from the PCD samples in the ground truth low-temporal rate PCD sequence; the high-dimensional features are then adaptively combined based on the relative temporal nearness of the measured PCD to estimate the missing PCDs at an intervening time step. Such a learning model works since many human activities comprise a temporal series of well-defined movements, so missing PCD can be learned from several existing data samples. To extract the relevant features, *mNetS* designs a customized Dynamic Graph Convolutional Neural Network (DGCNN) [21], which is effective for data with structural irregularity and order invariance, such as the PCD [21]–[23].

We implement and evaluate *mNetS* on a Commercial-off-the-shelf (COTS) mmWave testbed by collecting data samples for 7 distinct activities, such as walking, squatting, lunges, *etc.*, for about 1 hour for each activity, over a period of 2 months. The 7 activities capture wide variations in human posture sequences, enabling our system to be robust for most indoor activities. We collect this dataset at 40 ms sensing intervals, which serves as the ground truth high rate samples, and we use various undersampling with random intervals to emulate opportunistic sensing. In total, we have collected nearly 7,100 data samples (total size: 84 GB), with approximately equal distribution across the 7 activity classes - 800 to 1100 in each class; we use 5,700 samples for training, and the rest of the samples are used for testing and benchmarking *mNetS*. Our results show that *mNetS* estimates missing PCD with a median L1 Chamfer Distance (L1-ChD) of 31 cm and a median Earth Mover's Distance (EMD) of 7 cm, when the undersampling rates are varied from 3× to 8×, which indicate a good match. Such high-rate PCD also improves the performance of human activity sensing systems over PCD with missed frames or estimated with linear interpolation. For example, estimated PCD from *mNetS* can improve the performance of a typical human activity classification application over PCD estimated by linear interpolation from 49.5% to 71.7%.

In summary, *mNetS* enables human activity sensing and networking without requiring additional hardware in an indoor and privacy-preserving manner. By estimating high-rate samples from a low-rate sample sequence, *mNetS* overcomes missing sensing samples caused by co-existing networking applications. To achieve this, mNetS translates the sensing frames to PCD representation, and then, estimates the missing PCDs. Our approach is the first to estimate high-rate PCD from low-rate PCD in the mmWave signal domain. The estimated PCD generated by *mNetS* can improve various human activity sensing applications in scenarios where high-rate sampling of sensing is infeasible due to co-existing networking.

II. Background and Challenges

A. PCD from mmWave Signal Reflections

To sense the activities from mmWave devices, a wide bandwidth signal is transmitted from multiple antennas, and the reflections received from this transmission are combined to determine the intensity at different spatial locations (Figure 1[a]). The transmitted signal reflects off of various objects in the scene, including both the dynamic target and static reflectors. The reflections are received as a weighted sum of time-delayed transmitted signals from all reflecting points, weighted by the reflectivity of each point. However, the reflections from some static objects may be strong enough to overshadow the desired target points, making it necessary to remove the static background.

To remove the static background from mobile objects, we can use the Doppler information present in consecutive reflections. When the signal reflects off a dynamic object, the round trip delays of the consecutive reflections either increase or decrease, depending on whether the object is moving further away or getting closer. To identify the strong reflections corresponding to objects at distinct pairs of range and Doppler velocity, we can use a 2D Fast Fourier Transform (FFT) on the reflected signals and generate a range-Doppler heatmap (Figure 1[b]). We then use a Constant False Alarm Rate algorithm to identify reflectors as {range, Doppler} pairs. Each pair is then decomposed into its distinct Direction of Arrivals (DoA) in azimuth and elevation, $(\theta_{az}, \theta_{el})$, to generate all points, and a threshold is applied to remove the static points and construct a Point Cloud Data (PCD) for dynamic targets. Figure 1(c) shows examples of three depth images and the corresponding PCD for a human performing different activities around 2 to 3 meters away from a mmWave device.

B. Challenges with Joint Networking-Sensing

MmWave networking relies on directional beams between the Access Point (AP) and user, with continuous steering of the beam towards mobile users. In cases where sensing is integrated into the same system, the beam must also be steered towards sensing target, which can impact the networking

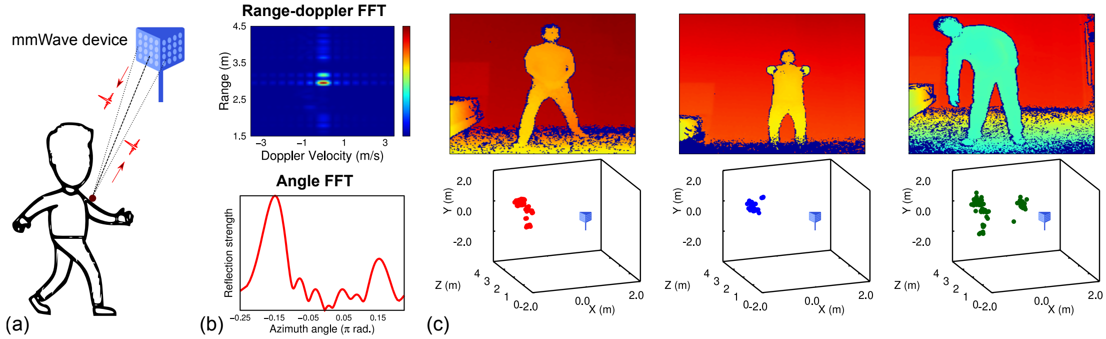

Fig. 1: (a) The transmitted signal reflects off the scene and captured by the mmWave device. (b) Range-doppler heatmap and angle information resolve the reflectors in distance, velocity, and direction. (c) Distinct human activity maps to distinct PCD.

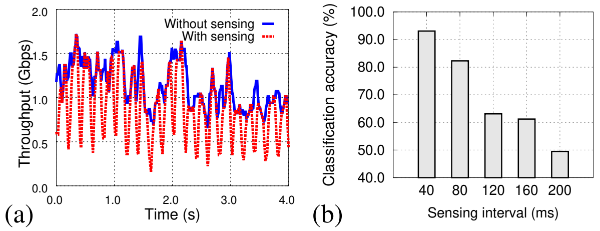

Fig. 2: (a) Introducing sensing reduces the networking throughput and increases the latency. Sensing interval: 200 ms. (b) Increasing sensing interval degrades sensing accuracy.

throughput. Besides, allocating different beams or spectrums for sensing and networking is infeasible due to the requirement of additional hardware and/or interference [13], [24]–[27]. Another approach could be time-multiplexing sensing and networking operations; but this leads to a tradeoff in performance between the two. Therefore, integrating sensing with networking in mmWave networks remains a challenging problem.

Impact on Throughput Performance. To understand the impact of allocating dedicated time slots for sensing on networking throughput performance, we conduct a simulation of the IEEE 802.11ad network [28] operating in an indoor environment. The details of our simulation is discussed in Section IV-B. A mmWave AP streams data to a mobile user ("user"), and aims to sense the activities of a dynamic target ("target") simultaneously. The user's mobility is simulated as a random walk within the span of the indoor setting, and the RSS at the user is obtained by the Ray-tracing method [29], [30]. The RSS is then translated to the downlink throughput by simulating a Single Carrier communication in IEEE 802.11ad [28]. To inject periodic sensing, the networking packets are switched off for a fixed period τ_s, at every interval of τ_p, where τ_s is the sensing duration. Figure 2(a) shows the frequent drops in throughput due to sampling the sensing frames at 200 ms intervals with sensing duration of 40 ms, which reduces the average throughput by 250 Mbps and increases the standard deviation to 350 Mbps. The performance is significantly affected due to frequent network disruption for sensing packets. Additionally, a minimum of 40 ms latency is introduced due to simultaneous sensing operation to the network packet transmission which could adversely affect real-time and critical applications [31].

Impact on Sensing Accuracy. While reducing the sensing interval can potentially improve the throughput, it can result in significant inaccuracies in classifying different activities of the target. Activity classification requires sensing signal over a window of time since an activity is composed of a set of well-defined body movements over time. At reduced sensing interval, an activity classifier has access to fewer samples of sensing signal, and thus, it becomes more difficult to infer the activity class. We use the reflected mmWave signals from the dynamic target in the above setup for 18 different dynamic activities such as squats, lunges, stretching, and static exercise postures such as arms up, standing on one feet, *etc.*, and use the mmWave signal classifier from [3] to predict the classes. Figure 2(b) shows that when the mmWave sensing samples is captured at 40 ms interval, the classification accuracy could be more than 90%. However, it quickly drops below 50% with 200 ms sensing interval. Therefore, the absence of high-rate sensing samples impacts the accuracy of activity recognition, which emphasizes the importance of reconstructing high-rate sensing samples to improve sensing accuracy without affecting networking performance.

III. *mNetS* DESIGN

A. Overview

mNetS aims to reconstruct high-rate, regularly spaced sensing samples from an integrated networking-sensing mmWave system. This system addresses the issue of missing temporal information caused by joint networking and sensing operations, making it valuable for pervasive, indoor human activity sensing applications. Additionally, repurposing existing networking infrastructure reduces installation overhead and costs. To this end, *mNetS* first identifies strong reflectors in the scene and filters out zero-doppler reflections to convert the acquired signals into a PCD representation. It then designs a customized deep learning framework that consumes a ground truth low-rate PCD sequence and produces a high-rate PCD sequence. The deep learning framework is composed of feature extraction modules that cast a pair of real, sampled PCD into a much higher-dimensional space than the input dimension. By leveraging a stack of well-trained feature modules, *mNetS*

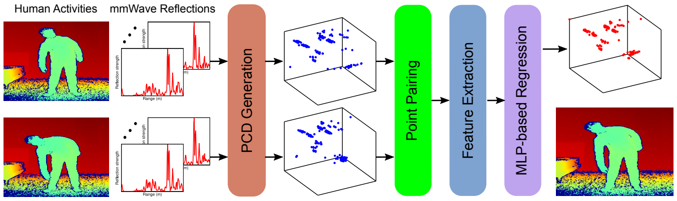

Fig. 3: System overview of *mNetS*.

learns a deeper mapping of the PCD data, enabling the system to estimate an intervening time step. The higher-dimensional feature maps are adaptively combined based on the relative temporal nearness of the two sampled PCD, and then regressed to the point space through a stack of multi-layer perceptrons. To effectively learn complex spatiotemporal features from real PCD, *mNetS* uses a Dynamic Graph Convolutional Neural Network (DGCNN) [21] and reconstruct high-quality, high-rate PCD for better sensing performance. Figure 3 shows an overview of *mNetS*, which takes two sensing samples at opportunistic networking idle slots, and predicts a missing PCD in between the two slots.

B. Temporal Prediction to Generate High-Rate PCD for Sensing

While several past works in the computer vision domain have addressed the problem of recovering missing PCD based on trajectory based interpolation or upsampling [32]–[36], there has been limited research in the mmWave domain. Unlike vision PCD, mmWave PCD could be highly noisy and missing a lot of points due to specular and multi-path signal reflections. Thus, the point-to-point direct mapping between adjacent frames is infeasible, making the trajectory estimation method unsuitable for mmWave PCD sequences. Instead, we exploit the stability of high-dimensional feature maps to estimate an intervening frame, given opportunistically sampled frames, and then regressing the combined feature maps to lower three dimensions. The scaling of the adaptive combination is based on the relative nearness of the time step. Furthermore, PCD poses additional challenges due to their structural irregularity and order invariance. Inspired by the recent works on order-invariant and graph-based neural models [21]–[23], we propose to extract the feature maps of each mmWave PCD frame through a graph-based neural network. The feature maps of the sampled PCD are then adaptively combined and passed through an MLP-based decoder to generate a PCD in an intervening time step, thus increasing the sensing sample rates. Figure 4 shows the overall architecture of *mNetS*'s temporal prediction network.

1) Extracting Feature Maps of mmWave PCD: PCD poses a unique challenge in feature extraction due to their structural irregularity and order invariance, and traditional convolutional neural networks are not effective with such data structures. Since PCD from mmWave reflections is highly sparse and noisy, it is also important for our feature extractor to learn the geometric features with both local and global context. Inspired by the previous work [21], we use a DGCNN-based feature extractor with EdgeConv layers to extract the features from the mmWave PCD. The EdgeConv layers of DGCNN can generate comprehensive contextual information by capturing not only local and non-local features for every point but also multi-scale geometric features; it also allows the feature extractor to learn to be less susceptible to noisy points. At its core, EdgeConv layers operate by dynamically generating a local neighborhood graph, and then, applying convolution-like operation on the graph edges. The graph is generated using the k-nearest neighbors in the feature space and then, aggregating the resulting feature maps with an aggregation operation. With 80 to 100 points in each PCD, the default setting of $k = 20$ and an aggregation operation of "max-pooling" is used in our case.

Our feature extraction module takes two sampled PCD and passes them through a stack of shared EdgeConv layers to create the respective feature maps. The EdgeConv layers successively raise the dimensionality space of each PCD from 3 to 16, 64, 64, 128, 256, and 512, and finally, the outputs from each EdgeConv layer must undergo an activation function to incorporate non-linearity. Since ReLU activation has the tendency to make many neurons inactive (the dying neuron problem), we use a Leaky-ReLU, which clips any value below 0 to a value close to 0 with a small gradient. Table I shows the parameters of the EdgeConv layers.

2) Adaptive Combination of Feature Maps: mNetS adaptively combines two adjacent PCD to generate a missing PCD in between. Intuitively, while a set of points could change between successive PCD, especially when they are far apart in time, their feature maps will more likely remain immune to change. Moreover, feature maps of a PCD will have a greater resemblance with the feature maps of a nearer PCD than a further one. Let's assume the feature maps of two ground truth PCD, *i.e.*, actively sensed from the target scene, P_i and P_j at time steps t_i and t_j, are $F_{i,m}$ and $F_{j,m}$, respectively. Here, m is the dimensionality of the feature maps. To predict the PCD $\hat{P}_k$ at time step t_k (where $t_i < t_k < t_j$), we first predict its feature map, $\hat{F}_{k,m}$ as an adaptive combination of $F_{i,m}$ and $F_{j,m}$, and then deconvolve $\hat{F}_{k,m}$ to generate $\hat{P}_k$. So, the feature maps are adaptively combined through weighted summation:

$$\hat{F}_{k,m} = w_i \cdot F_{i,m} + w_j \cdot F_{j,m} \tag{1}$$

$$where, w_i = \frac{t_j - t_k}{t_j - t_i}; \quad w_j = \frac{t_k - t_i}{t_j - t_i} \tag{2}$$

where the weights for the feature maps are determined based on the time separation between the predicted and ground truth PCD. However, adding the two feature maps from raw PCD may not produce a coherently learnable feature space for two reasons.

First, the points in each real PCD are ordered arbitrarily, and one point in PCD, P_i could be far away from its most corresponding point in the same index in PCD, P_j. To mitigate this, we apply a point pairing operation to reorder the points in P_j such that the corresponding pairs of points in the same indices in P_i and P_j form the closest pairs. *Second*, the point

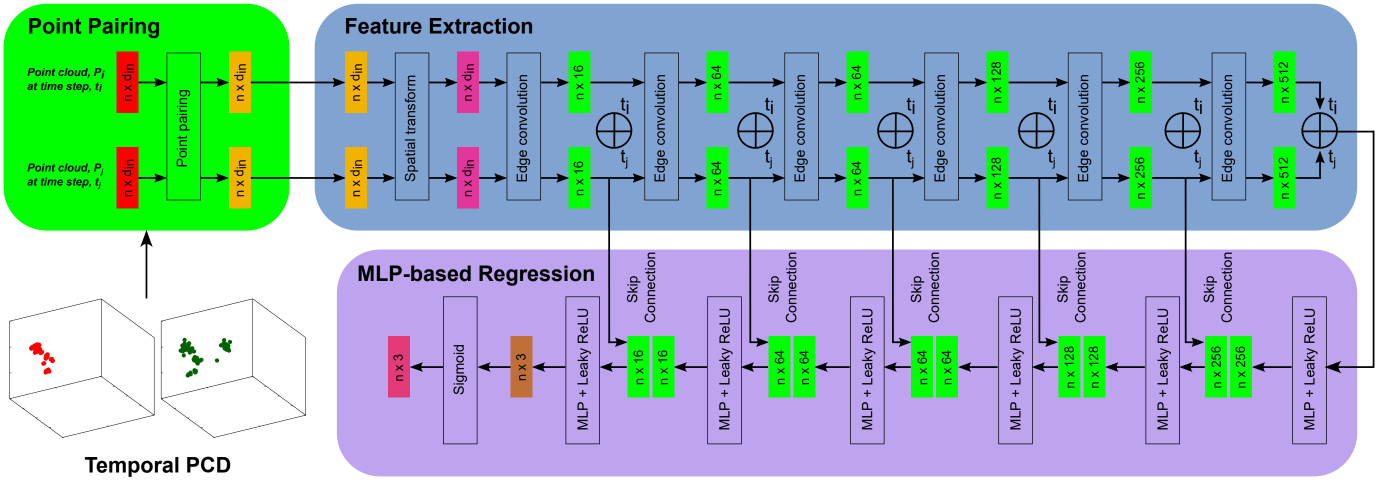

Fig. 4: Temporal prediction network in *mNetS*.

TABLE I: Temporal prediction network's parameters. ST: Spatial Transform; EC: Edge Convolution; MLP: Multi-Layer Perceptron; LR: Leaky-ReLU.

Encoder	ST	EC1	EC2	EC3	EC4	EC5	EC6
Number of points	N, N	N, N	N, N	N, N	N, N	N, N	N, N
Input channels	5	5	16	64	64	128	256
Output channels	5	16	64	64	128	256	512
Activation Function	Linear	LR	LR	LR	LR	LR	LR
Decoder	MLP1	MLP2	MLP3	MLP4	MLP5	MLP6	Output
Number of points	N	N	N	N	N	N	N
Input channels	512	256+256	128+128	64+64	64+64	16+16	3
Output channels	256	128	64	64	16	3	3
Activation Function	LR	LR	LR	LR	LR	LR	Sigmoid

coordinates in the input PCD may not be distributed in an uniform scale. As a result, contributions to the neuron from inputs spanning over a wider range in raw values tend to overshadow contributions from inputs over a smaller range. Therefore, we scale the input coordinates to scale (0,1) because each of the layer employ Leaky ReLU activation which clips any negative values close to 0. Thus, it is desirable that all our information in our input data is above 0, which helps to normalize the scale.

3) Deconvolution of Feature Maps to Point Space: Once the feature maps are combined, *mNetS* aims to deconvolve it to generate the PCD in real point space. A decoder network is tasked to deconvolve the high-dimensional feature maps into a 3-dimensional PCD in the polar coordinate/Euclidean space and then converted to Cartesian/XYZ coordinates. The decoder network employs a stack of MLP layers with decreasing output dimensions: 256, 128, 64, 64, 16, and 3 to regress the feature maps to the Euclidean space with LeakyReLU as the activation function for all the layers. We use MLP layers in the decoder since the EdgeConv layers in the encoder already learn the essential complex features which map the input PCD, P_i and P_j, to the corresponding feature maps of the output PCD, $\hat{P}_k$. Thus, a vanilla MLP model can regress to the Euclidean space from the adaptively combined feature maps, $\hat{F}_{k,m}$, after successively reducing the dimensionality. During training, however, such a deep network with 6 EdgeConv layers in the encoder and 6 MLP layers in the decoder suffer from the *vanishing gradient* problem, where the training procedure of a neural network adjusts the weights in each layer successively in proportion to the partial derivative of the loss function with respect to the weight through *backpropagation*. The problem arises when the partial derivatives in the earlier layers become too small that the corresponding weights become immune to any change. Thus, the loss function asymptotically approaches a higher minimum value than what is expected. To overcome this challenge, we employ *long skip connections* between the encoder and decoder [37]. Here, the lower dimensional features from the encoder are combined through an aggregation operation to the higher-dimensional outputs in the mirroring decoder layer (see Figure 4). For PCD feature maps, "concatenation" is the suitable option, which keeps all the feature channels from both the encoder and the decoder, and enables the gradients to have significant influence from the lower dimensional features. Table I shows the parameters for the MLP layers.

4) Loss Function: For accurate estimation of missing PCD, we must tune the EdgeConv and MLP layers through training. To this end, we use loss functions, which measure the geometrical difference between the estimated PCD, $\hat{P}_k$, and the corresponding ground truth PCD, P_k. Two widely used metrics for quantitatively measuring the difference between two point sets are the Chamfer Distance (ChD) and the Earth

Mover's Distance (EMD). ChD involves comparing each point in one set with the nearest point in the other set, and then computing the average squared L2-norm distance between the corresponding pairs, and is calculated as [38]:

$$L_{ChD} = \sum_{p_1 \in S_1} \min_{p_2 \in S_2} ||p_1 - p_2||_2^2 + \sum_{p_2 \in S_2} \min_{p_1 \in S_1} ||p_2 - p_1||_2^2 \tag{3}$$

where S_1 and S_2 are two point sets, and N_1 and N_2 are the number of points in them. However, training the network based only on a point to point comparison does not suffice, especially for mmWave PCD. This is because mmWave PCD may have a lot of noisy points, and we observe that when the network is trained only with ChD as the loss function, it struggles to learn the overall geometric structure. So, we also add EMD to the loss function, which determines the minimum cost of transforming one point set into the other, where the cost is defined as the distance between each pair of points [39]:

$$L_{EMD}(S_1, S_2) = \frac{1}{N_1} \min_{\phi : S_1 \to S_2} \sum_{p_1 \in S_1} ||p_1 - \phi(p_1)||_2 \tag{4}$$

where S_1 and S_2 are two point sets, N_1 is the number of points in S_1 and $\phi : S_1 \to S_2$ is a one-to-one mathematical correspondence function that maps each point in S_1 to exactly one point in S_2. Since EMD is computationally expensive, we resort to an approximate EMD based on Sinkhorn loss [39]. We combine the two losses by weighting them with hyper-parameters (λ_C, λ_E) as:

$$L = \lambda_C \cdot L_{ChD} + \lambda_E \cdot L_{EMD} \tag{5}$$

We discuss the tuning of these parameters in Section IV-A3. By using this combined loss during training, the network can accurately learn to estimate missing PCD in a robust and efficient manner.

IV. Implementation

A. Hardware, Data, and Training

1) Acquisition of PCD and Ground Truth: Due to the lack of open-source mmWave datasets, we design a custom hardware setup (Figure 5) by integrating mmWave transceivers to capture reflection signals, and a Kinect v2 depth camera to collect ground truth samples. Unfortunately, existing COTS mmWave networking devices which operate at 60 GHz as per IEEE 802.11ad standard [28] do not allow user access to raw reflection signal. Thus, we build our system from two 76-81 GHz mmWave transceivers, TI IWR1443BOOST [40], to resolve the reflections in both azimuth and elevation directions. [14] has shown that the 77 GHz signal band exhibits similar reflection strength as 60 GHz after accounting for greater Friis path loss. The two devices are attached in their place by rails with a fixed horizontal and vertical separation of 11.1 cm and 5.5 cm, respectively. Each device consists of 3 transmit and 4 receive antennas that are positioned in two distinct rows with 8 and 4 linear channels, respectively, providing a best resolution of 14.3° in both azimuth and elevation directions. The small number of transmit and receive antennas and their positioning emulates a real networking device with constrained hardware resources. The resolution in the depth dimension is given by the bandwidth of the signal, and with 3.07 GHz bandwidth in *mNetS*, the system achieves a depth resolution of 4.89 cm. To capture data in real time, we also attach a data capture module, TI DCA1000EVM [41], to each mmWave transceiver. We follow Section II-A to translate the mmWave reflections, and then merge two PCD from two transceivers by translating one *w.r.t.* another following their X and Y-axes separations. The setup is configured to generate PCD at 25 fps, which serves as the ground truth to our temporal prediction network. We resample the PCD in time to simulate lower sensing rates, which serve as the input to the same network. The Kinect v2 depth camera captures RGB and depth images at 30 fps within the same FoV as the mmWave transceivers, and provides visual references.

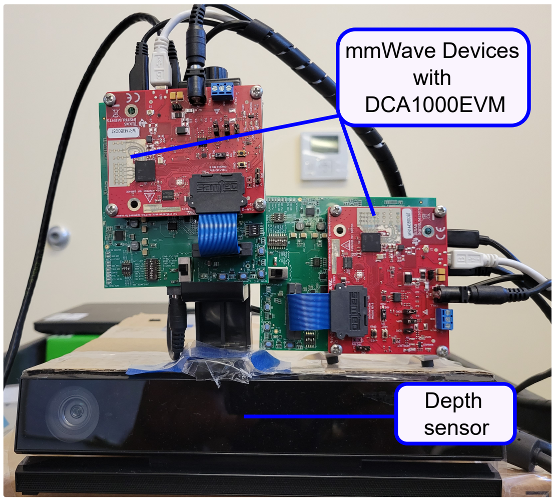

Fig. 5: Experimental platform with two mmWave transceivers [40] and a co-located Kinect V2 [42] for ground truth samples.

2) Data Synchronization: Due to the lack of hardware synchronization between the mmWave transceivers and depth camera, we employ a software synchronization. To this end, we first configure the mmWave devices and program to send a command to the Kinect to begin capturing frames, and the timestamp of this command is retained as the start time of the Kinect. Next, the transceivers synchronize themselves by exchanging TCP packets, and we store the corresponding global timestamps to start acquiring the mmWave reflections. During pre-processing, the timestamps enable us to find the common start time in which we have corresponding frames from all three devices. Frames outside this time are discarded. Since the transceivers and Kinect have 25 and 30 fps, respectively, we employ frame interpolation to fill in the gap.

3) Network Training: We train the models on a GPU server with 2 NVIDIA RTX A6000 cores by implementing the network on Tensorflow using Python 3.7 and use the "Adam" optimizer for fine-tuning. The entire dataset is split into non-overlapping training, validation, and testing sets. The proposed network is trained for undersampling rates from 3× to 8× which emulate low rate sampling under varying network traffic: heavy traffic forces the sensing module to sample at very low rate (high undersampling) and vice versa. For each epoch, training is achieved by outputting an average training loss, and then the validation dataset is passed through the trained model to generate the average loss on the validation set. We allow the networks to train until convergence, which is when the average validation loss does not decrease more than 10 times in the last 20 epochs. The training is repeated with several choices of hyper-parameters, λ_C and λ_E, and we find that a choice of $(\lambda_C, \lambda_E) = (0.3, 0.7)$ gives us the best result in terms of structural output. There are two possible reasons: *First,* the ChD component has a higher range in our data, so this choice of hyper-parameters brings the contribution from ChD down. *Second,* the network gives more importance to reconstructing the global shape rather than details at the point level. Across all our datasets, the network converged in 95 epochs, taking approximately 100 minutes to complete, with an initial learning rate of 1×10^{-3} and an exponential decay at a rate of 0.7. Once the model has been trained, the network takes less than 2 ms to generate a PCD, indicating that *mNetS* could infer activities in near real-time.

B. IEEE 802.11ad Network Simulation

Since our custom-made system does not support a real-time evaluation of the joint mmWave networking and sensing operations, we evaluate the effectiveness of *mNetS* by simulating the IEEE 802.11ad protocol based on an open-source, realistic Ray-Tracing method [29]. The Ray-Tracing method is more accurate than conventional simulations that rely on Friis path loss models because it takes into account the environmental layout and provides a more precise channel estimation. We estimate the indoor channels and modified the IEEE 802.11ad MAC layer to enable data scheduling to the user and sensing tasks for human activity recognition at opportunistic time slots. This allowed us to accurately quantify the data throughput and sensing performance of *mNetS* in various indoor scenarios.

1) Channel Estimation and Throughput Calculation: To estimate the channel in a realistic indoor environment, we first capture the visual PCD from an AR-capable smartphone [43]. Then, we apply a Ray-tracing method by varying the position of the AP and mobile users within the PCD. Specifically, for each combination of AP and user location and beam direction from the AP, we compute the expected signal reflection profile for a downlink channel using the approach proposed in [29]. To evaluate the performance of the channel, we compute the signal-to-noise ratio (SNR) based on the IEEE 802.11ad receiver sensitivity parameters, and simulate data transmission via the Single Carrier PHY payloads with the MCS varying from 2 to 13 [28]. Each MCS supports a pre-defined maximum bitrate, and we select the optimal MCS to minimize the packet error rate. To this end, we process the received packets through synchronization in time, demodulation, and frequency offset correction, and compare the recovered bits with the transmitted bits to classify each packet as either 'erroneous' or 'correct'. Based on this classification, we determine the MCS expected to achieve the lowest packet error rate at the estimated SNR.

To translate the SNR and MCS to effective throughput, we adjust the PHY throughput using the MAC layer efficiency. The efficiency is computed as the ratio between the MAC payload and the PHY packet length, which includes various fields such as the preamble, channel estimation, beam training, and interframe spacing [28]. In our simulation, we model the mobile users as performing a random walk at each time step, with a step interval of 10 ms. We evaluate our approach across three different environments, considering both 'mobile' and 'static' scenarios. For the 'mobile' scenario, we assume the user moves at a typical walking speed between 1.2 to 1.4 m/s. At each time step, we compute the channel between the AP and the user, and then estimate the effective throughput.

2) Injecting Sensing Packets into Networking Protocols: To simulate joint networking and sensing, we introduce special packets dedicated to sensing, which temporarily suspend data transfer during their duration. At each time interval, the AP executes one of three possible actions: beam search, sensing packet transmission, or data packet transmission. Beam search is performed periodically at intervals of 100 ms, introducing a beam search latency of approximately 2 ms for a single user and a 16-antenna AP, following the IEEE 802.11ad standard beam searching protocol [28], [44]. Sensing frame acquisition is performed at a specified frame rate, introducing a sensing delay at each interval, during which the system switches the beam from networking users to the sensing target to acquire a frame of sensing data. The remaining time is used for data transfer. This process allows us to compute the effective throughput when the sensing and networking applications run simultaneously.

V. Performance Evaluation

A. Microbenchmark results

1) Quantitative Results: We first evaluate *mNetS*'s ability to predict the PCD when the sensing samples are captured at a lower rate, and compare the quality of PCD when they are captured at a higher rate. To this end, the ground truth reflection signals are captured at 40 ms intervals, and we downsample them by different factors in time, and use them as the input to the temporal prediction framework. For each trial, we have a sequence of PCD, and we create samples for different undersampling rates, N, by selecting a sequence of $N+1$ back to back PCDs such that the first and the last PCD are used as inputs to the temporal prediction network to predict the $2^{nd}, 3^{rd}, 4^{th}, \cdots (N-1)^{th}$ PCD at the corresponding time steps. This undersampling rates vary from 3× to 8×. Note that our model only takes two captured PCDs rather than a sequence of captured PCDs to estimate PCDs in the intervening time steps. Thus, the model can still *estimate high temporal rate PCD sequence from an irregularly sampled PCD sequence* as long as the model is trained to estimate from a sequence with smallest temporal rate possible. We collect nearly 7,100 mmWave reflection samples from a human performing 7 different human activities following [3], demonstrating large

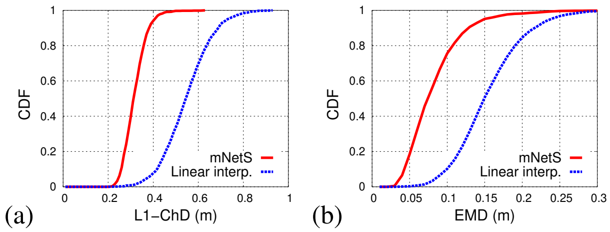

Fig. 6: Performance of temporal prediction compared with a trajectory-based linear interpolation: (a) L1-ChD; (b) EMD.

variations in human movements, and our model is trained and validated with nearly 5,700 samples and tested with additional 1,400 samples. In addition, we implement a linear trajectory interpolation method that estimates the missing PCD by finding the closest pairs of points in the two sampled PCD, and then, interpolating based on linear trajectory at the required time step.

Figure 6 presents the Cumulative Distribution Function (CDF) of the ChD and EMD metrics, which are used to evaluate the quality of the predicted PCD generated by the linear interpolation and *mNetS*, in comparison to the ground truth. The results indicate that the *mNetS* outperforms the linear interpolation method, with a median improvement of 24 cm or 43.6% in L1-ChD metric. Also, the 90^{th} percentile L1-ChD value is only 38 cm, suggesting that the *mNetS* produces PCD that closely resembles the ground truth. Additionally, the median EMD value is improved by 8 cm or 53.3% with *mNetS*, indicating that the model enhances the overall geometric structure of the predicted PCD. This shows *mNetS* is able to generate the missing information in time and improve predictions over trajectory estimation through linear interpolation. The resulting estimated PCD sequence can further assist in better human activity sensing without compromising the networking performance.

2) Qualitative Results: mNetS's approach to predict an intervening frame from a pair of real frames is by learning a set of features from the real frames which are adaptively combined. This allows *mNetS* to predict the salient shape information which is passed from one frame to the next. The approach first learns to identify clusters of points, and then, predicts whether an identified cluster is a valid, real cluster, or it is due to noisy points. Noisy points are ignored by *mNetS* since they do not pass information from one frame to the next coherently, and only the global structural information is retained. Figure 7 shows a set of examples of predicted PCD from *mNetS*, and their corresponding ground truths. *mNetS* predicts the major cluster(s) in the ground truths, which are large numbers of dense points that are unlikely to be noisy, while any minor cluster(s) are predicted only if there is sufficient common information in the real frames. Figure 8 shows an example of PCD prediction in a sequence of frames in 2 seconds. The global shape information passes successively through the frames by means of the major cluster, and the smaller cluster begins to fade away as we move to the last frame from the beginning.

3) Effect of Different Activities: To understand the impact of different activities on *mNetS*, we evaluate its relative performance across 7 different human activities that involve varying body movements, including ‘Lunges’, ‘Squat’, ‘Walking’, among others. Figures 9(a–b) show the ChD and EMD results, as bar plots with median and standard deviations. Overall, *mNetS* performs similarly across most of the activities, with errors falling within an acceptable range. However, we observe that activities involving faster body movements or more extensive coverage with arms or legs lead to increased errors. This is particularly noticeable for the ‘Squat’ activity where the likelihood of specular reflections is higher due to faster movement of the entire body. Despite this, across different activity types, the median L1-ChD and EMD ranges from 28 to 36 cm and 8 and 11 cm, respectively, indicating that *mNetS* can still accurately predict the PCD for a range of activities.

4) Performance of Joint Networking and Sensing: We now evaluate the ability of *mNetS* to enable joint networking and sensing operations by estimating the effect of networking throughput at different sensing overheads and comparing PCD prediction performance. Our networking simulation is carried out in an indoor environment, following Section IV-B, where a single AP serves a mobile networking user and senses activities of another human from sensing frames. The sensing frame spans 40 ms, and the sensing overhead is gradually increased by introducing sensing delay at intervals of 200 ms, 100 ms, 60 ms, and 50 ms, resulting in sensing overheads of 20%, 40%, 60%, and 80%. We carry out 7 trials at each sensing overhead, across three different indoor environments, including one corridor and two office rooms. Figure 10 shows the networking throughput and the performance of *mNetS* in terms of estimating missing PCD at each sensing overhead. The results show a tradeoff between sensing and networking performance, and *mNetS* attempts to improve this tradeoff. By reducing sensing overhead by 5×, *mNetS* allows sensing at the same frame rate as 100%, predicting the missed frames with L1-ChD that is approximately 45% lower than the linear interpolation. The reduction of networking throughput will be minimized to approximately 250 Mbps, keeping the throughput above 1 Gbps. In Section V-C, we will also evaluate the effect of this better PCD prediction in terms of human activity classification.

B. Ablation Study

1) Effect of EMD Loss: Next, we evaluate the effect of adding EMD loss during training. Recall that while ChD makes a point-to-point comparison, EMD measures the global shape distortion, and could improve the network performance. We use the same training and testing samples as before, and re-train our model with and without the EMD loss component (Section III-B4). Figures 11(a-b) show that there is a slight improvement in PCD prediction in terms of both L1-ChD and EMD due to adding the EMD loss component during training. The improvement is relatively small since our feature extraction from the mmWave PCD uses DGCNN with Edge-

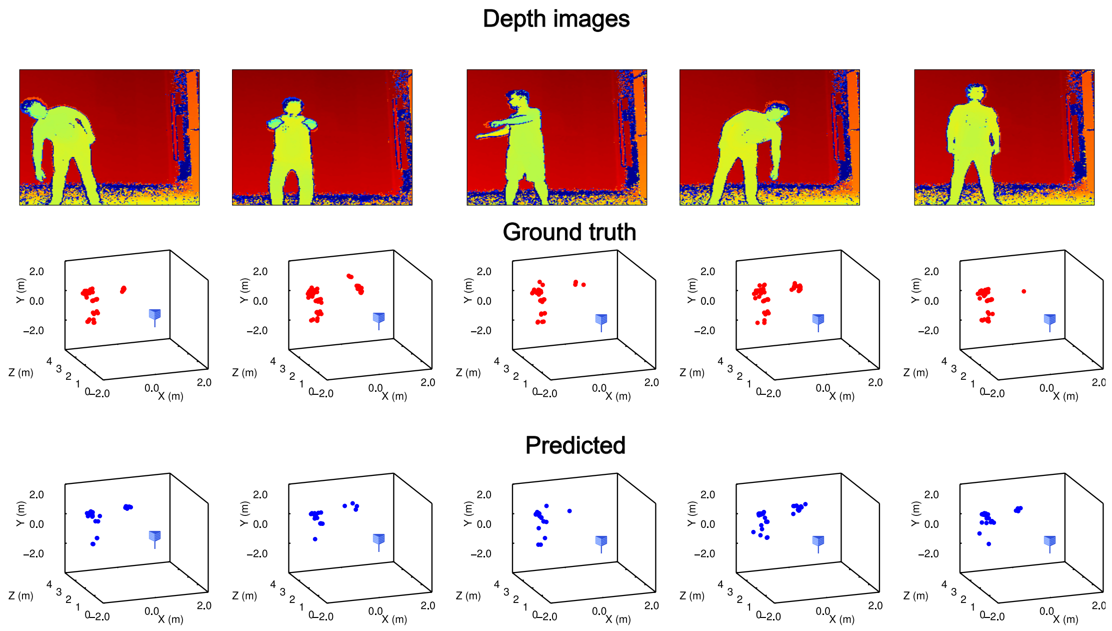

Fig. 7: Examples of PCD predicted by *mNetS* for different human activities.

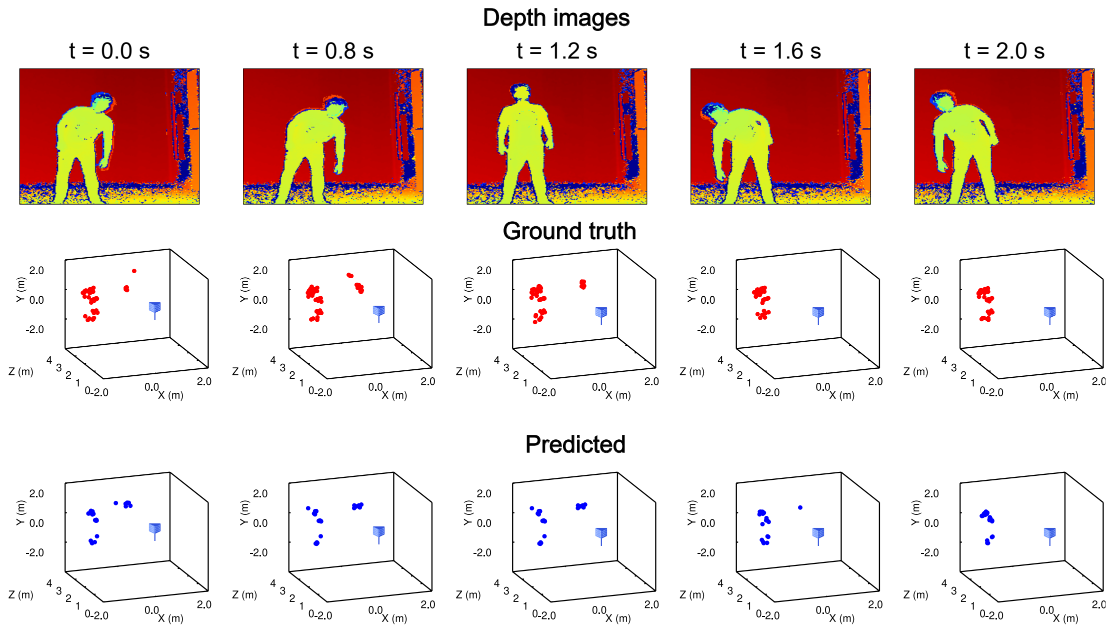

Fig. 8: Examples of PCD predicted by *mNetS* for human activity in time for 2 seconds.

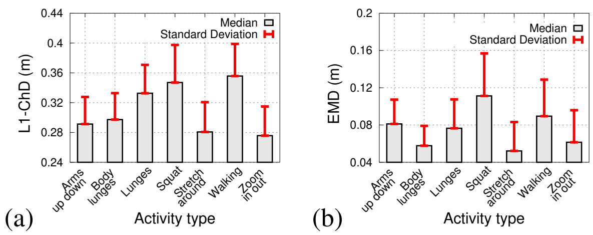

Fig. 9: Performance of temporal prediction across different activities. Faster activities increase the prediction errors, but within a tolerable range. (a) L1-ChD; (b) EMD.

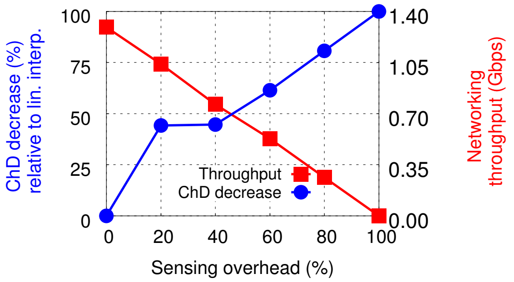

Fig. 10: *mNetS* achieves a good PCD prediction performance with less than 20% of sensing overhead, which translates to less than 25% of throughput drop.

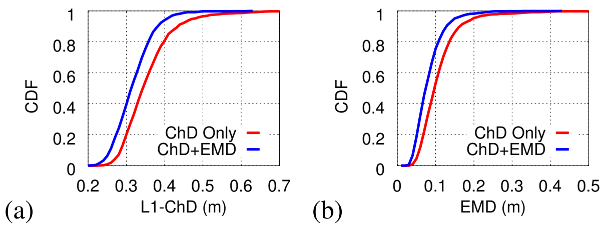

Fig. 11: Effect of training *mNetS* with and without EMD loss component in terms of (a) L1-ChD, and (b) EMD.

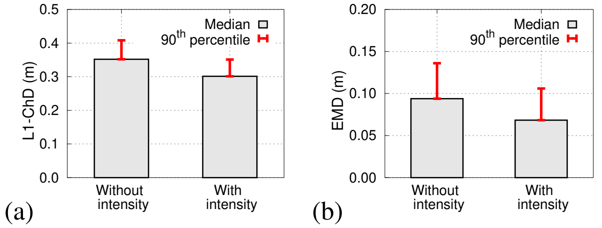

Fig. 12: Effect of intensity channel on *mNetS*'s prediction performance in terms of (a) L1-ChD, and (b) EMD.

Conv layers that already learns a global view of the features at each point. Thus, when *mNetS* is trained with only ChD, the feature set at each point still retains global shape. EMD still makes a small but significant improvement in predictive performance of *mNetS*; the median L1-ChD of 31 cm vs. 34 cm and median EMD of 7 cm vs. 10 cm with and without EMD loss, respectively.

2) Effect of Intensity Channel: In the prediction of mmWave PCD, noisy points from specular or multi-path reflections pose a significant challenge, as their appearance is random, and their features do not map coherently onto the next frame. To address this issue, we can exploit the fact that valid reflections generally have higher signal strength than random noisy points in mmWave PCD. Therefore, we add an additional channel to our input PCD to carry the reflection intensity information associated with each point, so that the network can learn to ignore those points. Figures 12(a-b) show the effect of training with and without the intensity information. The results show an improvement in predicting the PCD due to the inclusion of the intensity channel. Specifically, we observe a median reduction of 5 cm (14.3%) in L1-ChD and a median reduction of 3.5 cm (36.8%) in EMD.

3) Performance at Different Time Steps: mNetS must predict missing frames at different time steps relative to the real sampled frames. For instance, if sensing frames are unavailable for the last 320 ms, to generate sensing frames at 40 ms intervals, *mNetS* must reconstruct frames at 40 ms, 80 ms, . . ., 280 ms time steps to achieve the desired rate. Intuitively, predicted frames closer to the ground truth should exhibit better performance. Figure 13 illustrates this phenomenon. The best performance occurs at time steps 40 ms and 280 ms, which are the closest to the ground truth at 0 ms and 320 ms, respectively. Still, for the other time steps, *mNetS*'s performance does not degrade significantly, and L1-ChD *w.r.t.* ground truth still remains within 40 cm.

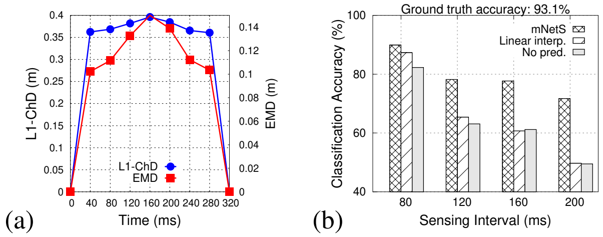

Fig. 13: (a) *mNetS*'s prediction performance at different time steps. (b) *mNetS* improves identifying activity types at reduced framerate.

C. Activity Classification Results

Finally, we evaluate *mNetS*'s ability to improve the performance of human activity classification. We create a classification network by adding an LSTM block after the MLP regression in our temporal prediction network and a set of dense layers to output class probabilities for 7 distinct classes. The network has 4 EdgeConv layers with 16, 64, 128, and 256 output channels, 5 MLP layers with 256, 128, 64, 16, and 1 output channels, an LSTM layer, and finally, 4 dense layers with 64, 32, 16, and 7 outputs. It takes a sequence of frames and extracts spatio-temporal features to predict class probabilities. We train the network on our data with a sequence of input frames spanning a 2-second interval and a 40 ms sensing rate. After training, we evaluate the classification performance on our test dataset and achieve 93.1% accuracy. The accuracy for each of the 7 classes is close to 90% with the lowest being 87% and the highest being 98%. This represents the ground truth performance when high rate sensing samples are available.

Then, we undersample the input data at different rates and replace the missing data frames with frames predicted by *mNetS* and trajectory-based linear interpolation. We also train the network to take inputs at the undersampled sensing intervals and evaluate the classification performance. Figure 13(b) shows the results. We observe that undersampled sensing rates can affect the classification performance significantly, dropping the accuracy from 93.1% at 40 ms sensing interval to 49.5% at 200 ms interval. The trajectory interpolation also shows a similar degradation in performance. In contrast, *mNetS* show a marked improvement over both the approaches, clearly sustaining the classification accuracy above 72%, even for 200 ms sensing interval. *In summary, the high rate predicted PCD from* mNetS *directly helps to improve the human activity classification performance.*

VI. Related Works

Sensing with RF Signals: Traditional contactless approaches on human activity sensing involve the use of vision or depth cameras, such as Kinect [45], which have privacy concerns as they generate a clear shape of the human body, and they are dependent on proper lighting conditions. To address these challenges, wireless signals have been extensively adopted for activity sensing. Existing works have been able to extract RF signatures from humans in an indoor environment, even in the presence of clutter, obstacles, and multi-person scenarios, and identify activities [4], [46]. Next-generation wireless infrastructure is expected to incorporate much higher frequency signals in the mmWave bands, promising a better sensing performance due to higher bandwidth, smaller wavelength, and larger number of antennas [6], [7], [17]. But most existing works on mmWave sensing do not work simultaneously with networking without affecting the performance. In contrast to the previous works, *mNetS* aims to enable the coexistence of networking and sensing for mmWave indoor networks.

Joint Networking and Sensing: Sensing human activity using RF signals requires leveraging the Channel State Information (CSI) between the networking device and the environment, which includes networking users and sensing targets. The existing body of works in integrating sensing on networking systems still present some limitations in applicability and tradeoff between networking throughput and sensing accuracy [13], [24]–[26], [47]. Some earlier works [25], [26] have proposed multi-armed beams for simultaneous networking and sensing. However, the challenge in creating multiple beams is to limit interference energy from one beam to the other, and this requires more expensive, sophisticated phased-array antennas. The interference issue in using multi-armed beams is more significant when the sensing beam must be sufficiently wide to capture reflections from all the points in the target. [13] proposes using the TRN fields of IEEE 802.11ay packets to estimate the CSI for multi-path propagation of signals between the networking system and the sensing targets. However, the temporal resolution of acquired micro-doppler signatures of human activities is limited by the beam training period, and thus, to achieve more fine-grained sensing, beam training frequency is increased at the cost of reduced throughput in networking. [24], [47] explored reusing networking packets for sensing. To address the bursty nature of networking packets, [24], [47] have explored compressed sensing techniques to exploit the inherent sparsity in mmWave reflections to reconstruct a full sequence of high temporal rate signal from low rate signal. However, this requires that the sensing can be invoked at specific time slots to optimize the performance of sparse recovery, which may be infeasible in a system where sensing frames are only captured opportunistically. In contrast to these existing works, *mNetS* is designed to execute mmWave sensing in a networking environment by repurposing the same hardware. Instead of imposing additional overhead or interference, *mNetS* opportunistically senses the target and then fills in missing information in time with deep learning.

VII. Conclusion

In this work, we present *mNetS*, an enabling technology for the coexistence of human activity sensing on networking systems. Such human activity sensing brings valuable applications in remote physical therapy and continuous health diagnostics at the users' home - *without any modification to indoor infrastructure* . *mNetS* achieves this by overcoming missing information in sensing samples resulting from concurrent networking. *mNetS* employs the feature extraction capability of a dynamic graph convolutional network to adaptively combine features from real sensing samples and estimate missing samples. Our experimental evaluation shows that *mNetS* effectively overcomes the challenges of mmWave signals and shows significant improvement in sample estimation, which improves the performance of sensing without affecting networking significantly.

Acknowledgments

This work is partially supported by the NSF under grants CNS-1910853, CAREER-2144505, and MRI-2018966.

References

[1] T. Wei and X. Zhang, "Mtrack: High-precision passive tracking using millimeter wave radios," in *Proceedings of the 21st Annual International Conference on Mobile Computing and Networking*, ser. MobiCom '15. New York, NY, USA: Association for Computing Machinery, 2015, p. 117–129. [Online]. Available: https://doi.org/10.1145/2789168.2790113

[2] P. Zhao, C. X. Lu, J. Wang, C. Chen, W. Wang, N. Trigoni, and A. Markham, "mid: Tracking and identifying people with millimeter wave radar," in *2019 15th International Conference on Distributed Computing in Sensor Systems (DCOSS)*, 2019, pp. 33–40.

[3] Edward M Sitar, et al., "A Millimeter-Wave Wireless Sensing Approach for at-Home Exercise Recognition," in *ACM MobiSys*, 2022.

[4] Lijie Fan, et al., "Learning Longterm Representations for Person Re-Identification Using Radio Signals," in *IEEE/CVF CVPR*, 2020.

[5] W. Wang, A. X. Liu, M. Shahzad, K. Ling, and S. Lu, "Understanding and modeling of wifi signal based human activity recognition," in *Proceedings of the 21st Annual International Conference on Mobile Computing and Networking*, ser. MobiCom '15. New York, NY, USA: Association for Computing Machinery, 2015, p. 65–76. [Online]. Available: https://doi.org/10.1145/2789168.2790093

[6] Zhen Meng, et al., "Gait Recognition for Co-Existing Multiple People Using Millimeter Wave Sensing," *AAAI*, vol. 34, 2020.

[7] Tao Li, et al., "MTPGait: Multi-Person Gait Recognition with Spatio-temporal Information via Millimeter Wave Radar," in *IEEE ICPADS*, 2021.

[8] Z. Li, Z. Xiao, Y. Zhu, I. Pattarachanyakul, B. Y. Zhao, and H. Zheng, "Adversarial localization against wireless cameras," ser. HotMobile '18. New York, NY, USA: Association for Computing Machinery, 2018, p. 87–92. [Online]. Available: https://doi.org/10.1145/3177102.3177106

[9] J. Pierce, "Smart home security cameras and shifting lines of creepiness: A design-led inquiry," in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, ser. CHI '19. New York, NY, USA: Association for Computing Machinery, 2019, p. 1–14. [Online]. Available: https://doi.org/10.1145/3290605.3300275

[10] J. Gong, X. Zhang, J. Ren, and Y. Zhang, "The invisible shadow: How security cameras leak private activities," in *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*, ser. CCS '21. New York, NY, USA: Association for Computing Machinery, 2021, p. 2780–2793. [Online]. Available: https://doi.org/10.1145/3460120.3484741

[11] Wired, "Review: Netgear Nighthawk M5 5G Router," 2022. [Online]. Available: https://www.wired.com/review/netgear-nighthawk-m5-5g-router/

[12] ZTE, "5G Indoor CPE Product," 2023. [Online]. Available: https://ztedevices.com/en-gl/mc801a/

[13] Jacopo Pegoraro, et al., "RAPID: Retrofitting IEEE 802.11ay Access Points for Indoor Human Detection and Sensing," 2022. [Online]. Available: https://arxiv.org/abs/2109.04819

[14] Aakriti Adhikari, et al., "MiShape: Accurate Human Silhouettes and Body Joints from Commodity Millimeter-Wave Devices," *ACM IMWUT*, vol. 6, no. 3, 2022.

[15] A. Sengupta, F. Jin, R. Zhang, and S. Cao, "mm-pose: Real-time human skeletal posture estimation using mmwave radars and cnns," *IEEE Sensors Journal*, vol. 20, no. 17, pp. 10 032–10 044, 2020.

[16] S. An and U. Y. Ogras, "Fast and scalable human pose estimation using mmwave point cloud," in *Proceedings of the 59th ACM/IEEE Design Automation Conference*, ser. DAC '22. New York, NY, USA: Association for Computing Machinery, 2022, p. 889–894. [Online]. Available: https://doi.org/10.1145/3489517.3530522

[17] Jaime Lien, et al., "Soli: Ubiquitous Gesture Sensing with Millimeter Wave Radar," *ACM Trans. Graph.*, vol. 35, no. 4, 2016.

[18] H. Liu, Y. Wang, A. Zhou, H. He, W. Wang, K. Wang, P. Pan, Y. Lu, L. Liu, and H. Ma, "Real-time arm gesture recognition in smart home scenarios via millimeter wave sensing," vol. 4, no. 4, dec 2020. [Online]. Available: https://doi.org/10.1145/3432235

[19] C. Liu, Y. Li, D. Ao, and H. Tian, "Spectrum-based hand gesture recognition using millimeter-wave radar parameter measurements," *IEEE Access*, vol. 7, pp. 79 147–79 158, 2019.

[20] Moraitis, Nektarios and Constantinou, Philip, "Indoor Channel Modeling at 60 GHz for Wireless LAN Applications," in *IEEE PIMRC*, 2002.

[21] Yue Wang, et al., "Dynamic Graph CNN for Learning on Point Clouds," *ACM Transactions on Graphics (TOG)*, 2019.

[22] e. a. Charles R. Qi, "PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation," in *IEEE CVPR*, 2017.

[23] Charles R. Qi, et al., "PointNet++: Deep Hierarchical Feature Learning on Point Sets in a Metric Space," in *NIPS*, 2017.

[24] Jacopo Pegoraro, et al., "SPARCS: A Sparse Recovery Approach for Integrated Communication and Human Sensing in mmWave Systems," in *IEEE IPSN*, 2022.

[25] Fan Liu, et al., "Toward Dual-functional Radar-Communication Systems: Optimal Waveform Design," *IEEE Transactions on Signal Processing*, vol. 66, no. 16, 2018.

[26] Haocheng Hua, et al., "Optimal Transmit Beamforming for Integrated Sensing and Communication," *arXiv preprint arXiv:2104.11871*, 2021.

[27] Carlos Baquero Barneto, et al., "Multibeam Design for Joint Communication and Sensing in 5G New Radio Networks," in *IEEE ICC*, 2020.

[28] IEEE Standards Association, "IEEE Standards 802.11ad-2012, Amendment 3: Enhancements for Very High Throughput in the 60 GHz Band," http://standards.ieee.org/findstds/standard/802.11ad-2012.html, 2012.

[29] Hem Regmi, et al., "Argus: Predictable Millimeter-Wave Picocells with Vision and Learning Augmentation," *ACM POMACS/SIGMETRICS*, vol. 6, no. 1, 2022.

[30] Shihao Ju, et al., "Scattering Mechanisms and Modeling for Terahertz Wireless Communications," in *IEEE ICC*, 2019.

[31] Lina Xu, "Context Aware Traffic Identification Kit (TriCK) for Network Selection in Future HetNets/5G Networks," in *ISNCC)*, 2017.

[32] Xingyu Liu, et al., "FlowNet3D: Learning Scene Flow in 3D Point Clouds," in *IEEE/CVF CVPR*, 2019.

[33] Yiming Zeng, et al., "IDEA-Net: Dynamic 3D Point Cloud Interpolation via Deep Embedding Alignment," in *IEEE/CVF CVPR*, 2022.

[34] Himangi Mittal, et al., "Just Go With the Flow: Self-Supervised Scene Flow Estimation," in *IEEE/CVF CVPR*, 2020.

[35] Fan Lu, et al., "PointINet: Point Cloud Frame Interpolation Network," in *AAAI*, 2021.

[36] Lukas Prantl, "Tranquil Clouds: Neural Networks for Learning Temporally Coherent Features in Point Clouds," in *ICLR*, 2020.

[37] Kaiming He, et al., "Deep Residual Learning for Image Recognition," in *IEEE/CVF CVPR*, 2016.

[38] Haoqiang Fan, et al., "A Point Set Generation Network for 3D Object Reconstruction from a Single Image," in *IEEE/CVF CVPR*, 2017.

[39] Ashish Shrivastava, et al., "Learning from Simulated and Unsupervised Images through Adversarial Training," in *IEEE/CVF CVPR*, 2016.

[40] Texas Instruments, "IWR1443BOOST," 2023. [Online]. Available: https://www.ti.com/product/IWR1443

[41] ——, "DCA1000EVM," 2023. [Online]. Available: https://www.ti.com/tool/DCA1000EVM

[42] Microsoft, "Kinect for Windows," 2023. [Online]. Available: https://learn.microsoft.com/en-us/windows/apps/design/devices/kinect-for-windows

[43] ASUSTek Computer Inc., "Zenfone AR: Go Beyond Reality," 2021. [Online]. Available: https://www.asus.com/us/Phone/ZenFone-AR-ZS571KL/

[44] Haitham Hassanieh, et al., "Fast Millimeter Wave Beam Alignment," in *ACM SIGCOMM*, 2018.

[45] Wenbing Zhao, et al., "A Kinect-based Rehabilitation Exercise Monitoring and Guidance System," in *IEEE International Conference on Software Engineering and Service Science*, 2014.

[46] Mingmin Zhao, et al., "Through-Wall Human Mesh Recovery Using Radio Signals," in *IEEE/CVF ICCV*, 2019.

[47] Ervin Sejdic, et al., "Compressive Sensing Meets Time-Frequency: An Overview of Recent Advances in Time-Frequency Processing of Sparse Signals," *Digital Signal Processing*, vol. 77, 2017.