AIGC for RF-based Human Activity Sensing

Ziqi Wang, *Student Member*, IEEE, Chao Yang, *Member*, IEEE, and Shiwen Mao, *Fellow, IEEE*

***Abstract*—Radio Frequency (RF) sensing has been considered as an effective approach to human perception of non-intrusive and high-privacy scenarios. However, the existing wireless sensing techniques mostly rely on extensive labeled RF sensing data for offline training, while wireless sensory data collection is highly time-consuming and costly. To ridge this gap, we investigate the problem of generalized dataset augmentation with an Artificial intelligence (AI) Generated Content (AIGC) approach, termed RF-AIGC, for wireless sensing, which can not only purposefully generate new RF sensing data but reduce the data collection cost by augmenting a limited training dataset with synthesized RF data. We propose a conditional Recurrent Generative Adversarial Network (termed RF-CRGAN) to generate labeled synthetic RF data for specified human activities for multiple wireless sensing platforms, such as WiFi, Radio-Frequency Identification (RFID), and millimeter wave (mmWave) radar. We also propose a holistic quantitative method to help evaluate and explain the effects of the synthesized data. The experimental results demonstrate that the proposed approach can effectively enhance the diversity of training data and achieve similar performance as real data.**

***Index Terms*—3D human pose estimation, Artificial intelligence generated content (AIGC), Generative Adversarial Network (GAN), Radio frequency (RF) sensing, Data augmentation.**

I. Introduction

Wireless communication technologies are extensively employed in in our daily lives, and RF signals (e.g., WiFi channel state information (CSI), millimeter wave (mmWave) radar, and Radio-Frequency Identification (RFID)) are becoming the frontier of intelligent human perception. These signals are refracted, scattered, or reflected by nearby objects or human body, while the frequency, phase, and attenuation variations can be exploited for sensing, especially for Human Activity Recognition (HAR). Compared to the use of computer vision (CV) or wearable devices, RF-based sensing offers the desirable device-free, low-cost, and non-intrusive advantages. In the past decade, considerable progress has been made in Machine Learning (ML)-powered wireless communications and networking [1], and among various ML algorithms, Deep Neural Networks (DNN) have been used as an essential technique for HAR tasks [2]. To achieve a satisfactory HAR performance, a huge amount of training data often needs to be collected, processed, and labeled. However, unlike text or image data, RF data is significantly harder to collect and has distinct randomness properties. First, the open-space propagation environment greatly affects the sensitivity of RF data; any change in the transceivers' locations or the surroundings could result in a very different data domain. Second, measured RF data is highly susceptible to various factors, such as transceiver devices, waveforms, frequency ranges, and protocols. For example, a 77 GHz mmWave channel is substantially different from a 900 MHz RFID channel, even in the same propagation environment. Third, there are significant time-varying fluctuations in the wireless channel according to the time of day, day of the week, and months. Owing to these intermingled spatial, spectral, and temporal features, collecting RF datasets is a very expensive undertaking, and what's worse, the usefulness of a collected RF dataset may be severely restricted for a different setting.

Recently, artificial intelligence-generated content, or AIGC, has started a prominent revolution in the ML world. Unparalleled products such as Sora, ChatGPT, Gemini, and MidJourney are paving the way towards Artificial General Intelligence (AGI). Generative adversarial networks (GANs), Transformers, and diffusion models are used to create new data that closely resemble existing data, and are the driving force behind these AI products, while they are mainly developed in the context of text-to-image generation or text-prompted AI agents. A natural question is raised: *Can we exploit the potentials of AIGC to address wireless communication problems, especially, to generating RF data that can be useful in different settings?* Data augmentation has been explored for RF sensing data, but has focused heavily on pre-defined transformation on a single RF sensing data source of WiFi CSI [3]. Generative Adversarial Networks (GANs), a representative AIGC technique, can provide cutting-edge realism and adaptability. Pan et al. [4] devised a GAN-based system to control the manipulation of images, achieving flexibility, precision, and generality. GANs have also been deployed in the RF sensing domain. However, they are only employed as a performance booster via fine-tuning or augmentation [5], [6]. It is important to mention that AIGC demands creatively generating and manipulating data of quality and diversity. The simplistically synthesized data would have limited usefulness for RF sensing applications, including HAR and 3D human pose estimation. In conclusion, the currently available approaches are not generalizable enough to support the synthesis of various human activity-related RF sensing data.

To this end, we take one step further and propose the RF-AIGC system to precisely generate activity-enriched RF data that is generalized across an array of RF platforms [7], [8]. We investigate the similarities between vision-based 3D human pose data and RF sensing data. They both demonstrate distinctive features with different activity movement variations, subject skeletons, viewpoints, and locations. We find that by augmenting new 3D human pose data, RF data can be generated corresponding to the augmentations, hence achieving

Z. Wang and S. Mao are with the Department of Electrical and Computer Engineering, Auburn University, Auburn, AL 36849-5201 USA. C. Yang is with the Hangzhou Institute of Technology, Xidian university, Hangzhou, China 311231. This work was conducted when C. Yang was pursuing a PhD degree at Auburn University. E-mail: {zzw0104, czy0017}@auburn.edu, smao@ieee.org.

This work is supported in part by the NSF under Grants CNS-2107190, CNS-2148382, and CNS-2319342. This work was presented in part at IEEE VTC-Fall 2022 and IEEE RFID-TA 2022

AIGC on both labels and RF data. However, simple supervised learning with a one-to-one ratio between human pose data and RF data does not make the system adaptive to various augmented human pose data. To address this problem, we first pretrain an offline generator using collected ground truth data. We then fine-tune this base generator with augmented pose data via a weakly-supervised learning approach to create an online RF sensing data generator, termed RF-CRGAN. Note that only the offline training involves Kinect data, which is not needed in the online mode. Hence our framework is non-intrusive and privacy preserving. We also devise a pose augmentation method that allows users to precisely control the augmentation process. Extensive experiments are conducted to validate the high data generation ability and diversity of the proposed RF-AIGC system. An overview of the proposed system is presented in Fig. 1.

The main contributions of this study are summarized in the following.

- To the best of our knowledge, this is the first work to design an end-to-end AIGC framework conditioned on 3D human pose data for generating a diverse set of RF sensing data with fine-grained specifications. To achieve this goal, we design an online augmentation module that allows users to easily augment 3D human pose, which can then be mapped into the corresponding RF sensing data. The cost of training dataset collection can be significantly reduced.
- We investigate a generalization solution to data augmentation for multiple RF platforms. The proposed RF-AIGC system is able to synthesize high-quality RF sensing data across four different RF platforms.
- We prototype the proposed system and qualitatively and quantitatively evaluate the performance of our synthesized data through various metrics, including the Structural Similarity (SSIM) Index, fidelity, diversity, and multimodality. Via different downstream tasks and cross-dataset testing, the high utility of our synthesized data is demonstrated and measured.

The remainder of this paper is structured as follows. We review related work in Section II and discuss the challenges in RF data collection in Section III. The system design is elaborated in Section IV. Section V presents our experimental study and Section VI concludes this paper.

II. Related Works

A. RF sensing of human activity

Various wireless technologies have been applied for HAR, such as RFID, WiFi, and mmWave radar. RFID is a near-field communication technology consisting of a reader and low-cost tags. Because tags can be classified due to their stored Electronic Product Code (EPC), they can be attached to different human body parts as wearable sensors. This sensing technology has a higher tolerance against environmental interference compared to other RF methods. Commodity RFID readers, e.g., the Impinj R420 reader used in this work, can extract useful information from the signals reflected from tags, including Received Signal Strength Indicator (RSSI), Doppler frequency, and phase angles. RFID has been applied for HAR ranging from simulating virtual touch screens in the air [9], human object interaction detection system [10], to free-weight exercise monitoring [11]. Similarly, WiFi CSI has become one of the most dominant RF sensing technologies in recent years for its ability to reveal fine-grained information at subcarrier level. CSI of the orthogonal frequency division multiplexing (OFDM) channel can be collected by leveraging a few open-source toolkits, such as amplitude, phase, and Doppler shift, which play a huge role in various motion sensing tasks [12]–[16]. Recently, FMCW radar has also been applied for HAR tasks, typically by detecting the direction and velocity of human body movements. Commodity devices such as Texas Instruments' (TI) IWR1443BOOST have gained increasing popularity for their lower cost and ease of deployment. Chirp frequency differences (also beat frequencies) can be utilized to derive distance, velocity, Doppler frequency, and angle information of the target object or human body. More activity-sensitive features such as micro-Doppler signatures can be obtained through short-time Fourier transform (STFT) [17]. These useful features from FMCW radar have also been explored for tasks involving tracking of the human body for fall detection [18], vital sign monitoring [19], and dynamic hand-gesture sensing [20].

This paper an AIGC-empowered method that is centered on large-scale HAR (instead of small-scale HAR such as vital signs and hand gesture recognition), which is essential in our daily lives because of its ability to deduce high-level knowledge about complex human behaviors [21]. All the aforementioned RF devices are suitable for large-scale HAR tasks, specifically 3D human pose estimation and daily activity classification. 3D human pose estimation typically requires a recurrent DNN to map the RF features into joint rotations for smooth 3D human movements [22]–[24], while classifier-oriented DNNs are mainly used for human activity recognition [25]–[27]. The main challenge is that these DNNs require a large amount of high quality training data, which takes huge amounts of time, labor, and computing power to collect and process. Moreover, the scarce labeled RF data impedes the development of DNNs that can generalize to unseen scenarios.

B. Data augmentation for RF sensing

Data augmentation is an important technique for camera- and sensor-based HAR to address the challenges mentioned above. In [28], inertial measurement unit (IMU) sensor data was generated and the generated data helped boost the HAR performance. However, this method has not been proven effective for RF sensing technologies due to the random characteristics and dynamic environments. A majority of studies that explore data augmentation in the RF domain leverage data transformation to modify existing data. For instance, the authors in [29] applied three operations to generate mmWave point cloud samples with varying distances, angles, and human motion velocities. Additionally, the work [30] synthesized different activity data through various transformations on CSI spectrograms. In spite of boosting the performance of the

models by adding generated data, these methods are limited to specific RF data formats of specific devices and do not convey complex movement characteristics.

Recently, data augmentation for mmWave Doppler radar data-based HAR received significant attention due to its unique Doppler-range and radial velocity features that can capture complex human activities. In [31], mmWave Doppler data were software-synthesized by deploying a virtual radar system within the 3D space that was constructed through vision-based data. Nevertheless, this application does not generalize to unseen activities, while still requiring extensive knowledge of the complex mmWave signal. Generative Adversarial Networks (GANs) [32] were applied to synthesize mmWave radar micro-doppler spectrums for enhanced HAR accuracy. Our previous work [33] leveraged a conditional GAN to synthesize RF samples of various types of modulation and different signal-to-noise (SNR) levels for automatic modulation classification. In [7], we utilized a GAN-based framework to generate synthesized RFID-based human pose data for augmentation purposes. However, these prior works lack the ability to generalize to unseen scenarios including different subjects, activities, downstream tasks, and RF platforms, which will be addressed in this work.

C. AIGC for RF-based human activity sensing

AIGC brings unparalleled freedom for users to create various forms of contents by simply providing a prompt. The most current backbones of AIGC are transformers and diffusion models. The Diffusion model [34] is an advanced method in the field of image generation with its high speed of generation and ease of engagement. Coupled with transformer-based architectures such as BERT [35] and CLIP [36], powerful text-to-image applications like stable diffusion [37] and DALL-E [38] are used by customers around the world. Recently, Wang et al. proposed a unified weighted conditional diffusion model (UN-CDM) to generate realistic human flow data for enhanced wireless sensing [39].

The field of 3D human pose estimation also had the opportunity to capitalize on the power of AIGC. A novel human motion diffusion model can generate various 3D human poses of high fidelity and diversity for user entered text descriptions, activity class labels, or poses [40]. However, the unprecedented capabilities of the above models have a solid foundation that cannot be ignored, which is an abundant amount of natural language descriptions and data for training the model. The field of RF activity sensing, on the other hand, has not yet been able to utilize what the best AIGC models can offer due to the lack of RF data and labels, let alone the challenge on prompts to describe RF data. In this paper, we make one step forward, proposing a GAN-based RF activity data generative model. Using a 3D human skeleton and a target activity as prompt, the RF data corresponding to the given subject and activity can be generated. In addition, with simple manipulation of the Blender animation software [41], small disturbances can be introduced to the 3D human pose data for improved diversity.

III. Challenges in RF Sensing Data Collection

Different from vision-based techniques, it is difficult to perform human perception tasks based on RF signals. Due to the complex format and highly random nature of RF signals, the translation from RF signals to human activities is not straight-forward. In recent years, deep learning (DL) models have been considered as effective approaches to extract human activity features from RF signals. With a suitable DL model and sufficient training data, the trained model can translate RF signals into activity information. However, the currently proposed DL-based approaches require an extensive labeled dataset for supervised training, which consists of synchronized activity information (i.e., label) and RF signal, especially for complicated human sensing tasks. Moreover, to achieve adaptability, the training dataset should be collected from a large number of subjects, diverse environments, different RF devices, and RF data modality (e.g., RSSI, phase, range profile, and so on). Training data collection for RF sensing systems is a dauntingly high-cost process.

The challenges of the training data collection for RF sensing mainly come from the following three aspects.

1) High labor and time costs. First, RF sensing data collection is simply labor-intensive and time-consuming. For example, it requires around 6 minutes of data under one setting (the same subject, viewpoint, location/environment, and RF platform) for the RF-Pose model to estimate recognizable poses similar to ground truth poses, based on seen poses and subjects during training, and around 18 minutes for the model to estimate smooth pose movements [22]. For the dataset to be useful, these will be repeated for a large number of different subjects, locations/environments, viewpoints, RF platforms, etc., costing huge amounts of efforts and time. The diversity in training subjects, environments, and activity types also needs to be tested in a cross-dataset scenario.

2) Synchronized multi-modal data. Second, most existing pose sensing models require synchronized multi-modal data, i.e., vision data and RF data, for supervised training, which requires synchronization of the multi-modal data collection devices, usually with different sampling rates. Data collection is needed again if the misalignment in time is larger than half of the duration of a performed activity.

3) Diversity in RF devices. Lastly, considerable diversity exists in the RF signal representations collected from different RF devices, operating on different frequency bands with different protocols, waveforms, and hardware designs. There is no universal data type for various platforms to work at the same time. The same propagation environment will look very different, e.g., RFID in 900 MHz versus FMCW radar in the mmWave band, and the same human activity will be transformed into diverse representations of RF data. Different RF devices also have different issues on data collection. For example, RFID tags need to be attached to human body, while WiFi and FMCW radar are device-free. Due to the interrogation protocol, RFID data is usually extremely noisy and sparse; data imputation is needed to make the collected data useful [22]. On the other hand, 2.4G WiFi CSI data

is highly susceptible to interference and movements in the surroundings, often leading to poor recognition performance.

IV. AIGC for Labeled Human Activity RF Dataset

We propose to leverage the power of data augmentation to effectively reduce the RF data collection efforts. Data augmentation aims to increase the amount of data by adding slightly modified copies of the existing samples or creating new, synthetic data. For image data used in computer vision tasks, an effective method is to perform a combination of affine image transformations such as rotation, reflection, scaling, and color modification (e.g., changing contrast or brightness, white balancing, sharpening, and blurring).

Obviously, such methods will not work for RF data. Pose, however, is more similar to image data and can be manipulated and modified more easily in terms of pose movement variations, body forms, camera viewpoints, and locations. We leverage GANs in our system because they provide a powerful framework for generating high-quality and diverse RF sensing data directly from 3D human pose data. Unlike traditional GANs that generate data from random Gaussian distributions, the autoencoder-based GAN leverages a structured latent space that can be manipulated using pose data, allowing more control over the variations in generated RF data. Additionally, the adversarial training mechanism of GANs enables the model to learn to generate realistic RF data rather than data that might closely resemble the input pose data. Compared to other methods like simulation-based synthesis, our approach simplifies the generation process, requiring less domain-specific knowledge while enabling the generation of diverse and robust datasets.

We find that *joint kinematics of 3D human pose data can be seamlessly mapped into credible RF data by our proposed RF-CRGAN network*. By enhancing the diversity of pose data, we can, in turn, augment RF data by mapping the augmented pose data into high-quality RF data. We propose the end-to-end framework *RF-AIGC*, which involves both augmentation and generative modules. Fig. 1 presents an overview of the RF-AIGC architecture, which comprises two main modules: an RF-CRGAN model and an artificial-poses generating module. We next present the detailed design of RF-AIGC.

A. RF data collection and pre-processing

As in our prior work [42], we aim to develop a general framework that works for different wireless technologies, using RFID, WiFi, and FMCW radar as examples. We first describe how the RF data is collected and preprocessed using these wireless technologies. The RF features for our model to learn are selected and processed. A suitable amount of RF features are then chosen to match the joint rotations of 3D human pose data.

1) RFID platform. We attach 12 passive RFID tags as wearable sensors to the 12 joints of the test subject [22]. While the subject performs various poses, a reader with three antennas queries the tags and collects phase variation data (i.e., the differences between two consecutive phase samples) from tag responses for the kinematics RF mapping network. The phase variations can be expressed as follows:

$$\Delta\phi_{RFID} = \text{mod}\left\{\frac{4\pi(S_t - S_{t-1})f_\alpha}{c}, 2\pi\right\}, \quad (1)$$

where S_t denotes the tag-to-antenna distance for the tth sampled data on channel α, and c is the speed of light. $(S_t - S_{t-1})$ represents the change of relative distance in the last sample, rendering it suitable for tracking the movement of the tag. The sampling rate of RFID phase data is 110 Hz.

2) WiFi platform. A commodity WiFi platform is built to capture WiFi CSI (both in 2.4 GHz and 5 GHz), which is a fine-grained feature representation of the OFDM channel. The CSI data used in this paper refers to the differences of phase values between adjacent antennas for the nth subcarrier:

$$\Delta\phi_{CSI} = (\phi_{k,n} - \phi_{(k+1) \bmod 3,n}) + \epsilon, \quad (2)$$

where k denotes the antenna that collects the phase data, and ϵ is the random noise. We collect 30 subcarrier-level phase information from each antenna to obtain 90 phase difference samples for one time frame of the activity being captured. A mean absolute deviation-based selection method is chosen to filter out the top $k(36)$ most reliable subcarriers for activity sensing. Joint kinematics can be learned from the CSI data because human activities (moving and rotating body parts) can cause considerable variations in the WiFi channel. The sampling rate of CSI phase difference data is 10 Hz.

3) FMCW radar platform. A commodity FMCW radar (TI's IWR1843BOOST) is utilized in this study. The intermediate frequency (IF) signal's (between transmitting and receiving chirps) frequency is $f_{IF} = S2d/c$, i.e., directly proportional to the range between the reflecting object and radar d (where S is the slope of the frequency modulation and c is the speed of light). Therefore, a Range-FFT (1D Fast Fourier Transform (FFT) across multiple IF signal frequencies) can be used to obtain range profile $X[k]$ (i.e., the power of reflected signals) at different distances, through a discrete Fourier transform (DFT) on the N_s sampled IF signal, as

$$X[k] = Ae^{j\phi_{IF}} P_{N_s}\left(\frac{2\pi k}{N_s} - \omega_{IF}\right),\ 0 < k \leq N_s, \quad (3)$$

where ϕ_{IF} and ω_{IF} are the phase and discrete angular frequency of the IF signal, respectively, and $P_N(\omega)$ is the Fourier transform of a square window function of length N.

During the experiment, we find that range profile alone carries enough periodic information of human movements. Microdoppler signatures can capture more movement information, but a format of 2D range-Doppler feature map for one each frame can induce computation complexity, while a format of time-Doppler feature map lacks the characteristics of moving joints across different range bins. One range profile data frame consists of the time duration of active chirp processing, ADC, and DSP (FFT). Due to the 256-point Range-FFT, the range bin resolution is around 0.044m. Such precision is sufficient to capture the movements of different joints of human body. The length of such range bins is adequate for the supervised learning network to learn the transformation. We

This article has been accepted for publication in IEEE Internet of Things Journal. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2024.3482256

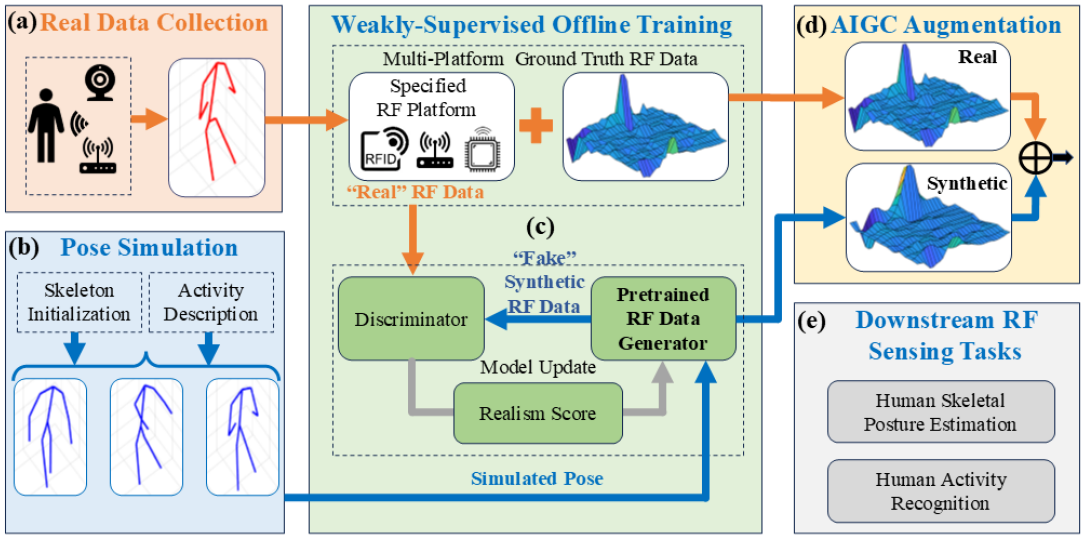


Figure 1. Overview of the architecture of the proposed RF-AIGC system: The primary function is to generate RF sensing data from pose data, and there are five main sections. (a) Real RF and pose data collection from different RF sensing platforms and kinect camera. (b) Online pose data simulation for enhanced diversity. (c) Real RF data and simulated pose goes through a weakly-supervised training: synthesized RF data from simulated pose data undergoes adversarial training with real RF data for a better and more generalized generator. (d) Our synthetic RF data can be used as AIGC data for augmentation to help training the models of different downstream RF sensing tasks as mentioned in (e).

only choose to use 36 points, in which range the test subject is performing the activities. The sampling rate of FMCW radar is also 10 Hz.

4) Training-ready RF and pose data. Since the frame rate of Kinect video data is 30 Hz, we first synchronize the three types of distinctive RF data with the Kinect data to 10 Hz utilizing recorded timestamps for time dimension alignment. Following that, the background information are effectively removed by Hampel filters. The three types of distinctive RF data are in upstream formats of typical RF data preprocessing, and synthesized data in such format can be later processed further (e.g., Doppler FFT on the range profile data) for more in-depth downstream tasks. We design a universal representation for the RF data: the input to our AIGC network is $S^N_{1:T}$, with N being the RF features, and T being the total number of time frames. We set T to 30 (or 3 seconds). Denote the 3D human pose data as $X^P_{1:T}$, where P represents the number of joint positions. We choose 3 seconds for T because this will includes roughly 1 to 2 cycles of a complete activity performed by the test subject.

B. The proposed RF-CRGAN system design

Our proposed neural Kinematic RF transformer network for RF sensing data synthesis is illustrated in Fig. 2, which consists of two stages: The first stage is to pretrain a baseline Generator that can produce synthesized RF data; then in the second stage, we fine-tune the Generator and start alternately optimizing the Generator $\mathcal{G}$ and Discriminator $\mathcal{D}$. We aim to enrich the diversity of our generative model, instead of generating homogeneous RF data that only closely resemble the ground truth data, which is lacking the ability of out-of-distribution generation. We introduce two effective augmentation methods, i.e., Temporal Gaussian Noise Perturbations (TGNP) and Pose-related characteristics modifications (PoseMod), to increase the diversity of pose data (i.e., pose augmentation). The Generator $\mathcal{G}$ can then transform such pose data into their corresponding RF data (i.e., RF data synthesis). However, there is no such RF data label available for the Discriminator $\mathcal{D}$ to use. Therefore, a weakly-supervised mechanism is incorporated by making the Discriminator $\mathcal{D}$ treat such synthesized RF data as "fake" samples. Intuitively, this enforces the RF features learned from ground truth data (with labels) to be adapted to augmented RF data through adversarial learning.

1) Pretraining the Generator $\mathcal{G}$. Following the GAN design of our previous work [7], we first pretrain the baseline RF feature transformer (i.e., Generator $\mathcal{G}$) leveraging collected ground truth pose and RF data with a supervised one-to-one ratio. An RNN-based autoencoder is deployed to map joint rotations of 3D pose data into RF features, and a 1D convolutional layer-based Discriminator takes the synthetic RF data as "fake" data, in contrast to the real RF data, to learn how to distinguish them for improving the network's generation ability. After training, we simply discard the Discriminator and use the trained RNN-based autoencoder as our Generator $\mathcal{G}$.

2) Generator fine-tuning and adversarial learning. We preserve the weights of the trained Generator during the first stage to maintain knowledge of ground truth data, which captures the most realistic RF distributions. We modify the RNN autoencoder by conditioning on the skeletons $\hat{S}$ and global motions G_t (regarding translations and rotations) through the RNN hidden dimension. 3D pose data is represented during the autoencoder stage as $X^{P'}_t = [L^{P'}_t, G_t \in R^{1\times 3}]$, consisting of local joint motions and global motions, where P' represents the number of joint positions (x, y, z). The RNN encoder and decoder are denoted by ε and ψ, respectively. Conditioned on previous time steps via an RNN hidden representation, we synthesize the current 3D pose at time step t. The temporal consistency in the pose sequences can be captured as

$$h^{\varepsilon}_t = RNN^{\varepsilon}(X^{P'}_t, h^{\varepsilon}_{t-1}; W^{\varepsilon}), \tag{4}$$

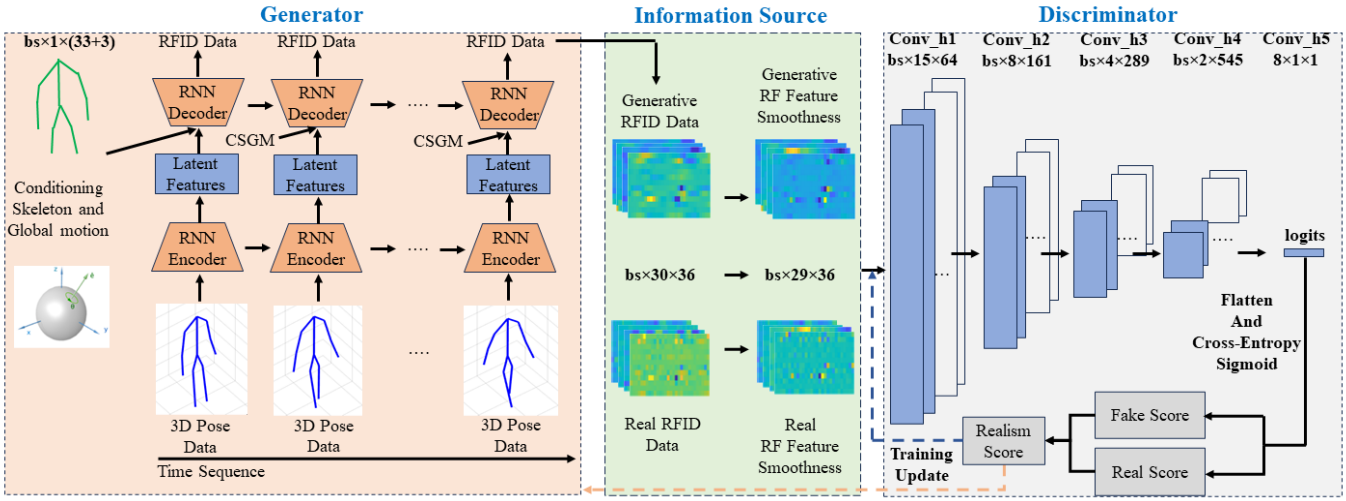


Figure 2. Internal architecture of the RF-CRGAN network.

where $h_t^{\boldsymbol{\varepsilon}}$ is the encoded representation of the input pose up to time step t. The encoded features are then fed to the RNN decoder to transform pose data into RF data through

$$h_t^{\boldsymbol{\psi}} = RNN^{\boldsymbol{\psi}}(\tilde{L}_{t-1}^{P'}, \tilde{G}_{t-1}, \hat{S}, h_t^{\boldsymbol{\varepsilon}}, h_{t-1}^{\boldsymbol{\psi}}; W^{\boldsymbol{\psi}}), \quad (5)$$

$$\tilde{S}_t^N = W^{ST} h_t^{\boldsymbol{\psi}}, \quad (6)$$

where $h_t^{\boldsymbol{\psi}}$ is the hidden representation of $\boldsymbol{\psi}_t$, $\tilde{S}_t^n$ is the transformed RF data at time step t, and $W^{\boldsymbol{\varepsilon}}, W^{\boldsymbol{\psi}}, W^{ST} \in R^{1\times 3P'}$ are learnable parameters. As can be seen in the above equations, the RNN decoder $\boldsymbol{\psi}$ is able to turn local joint motions $\tilde{L}_{t-1}^{p'}$ conditioned on pose skeletons $\hat{S}$, and global motions $\tilde{G}_{t-1}$ into RF data $\tilde{S}_t^N$. Hence our RNN Autoencoder-based Generator F is also called the *Neural Kinematics RF Mapping Network* in this paper.

After pretraining and fine-tuning the Generator $\mathcal{G}$, we alternatively optimize $\mathcal{G}$ and $\mathcal{D}$ as in the following minimax game:

$$\min_{\mathcal{G}} \max_{\mathcal{D}} [\mathcal{L}_{ce}(\mathcal{D}(S_{2:T}^N - S_{1:T-1}^N), 1) + \mathcal{L}_{ce}(\mathcal{D}(\tilde{S}_{2:T}^N - \tilde{S}_{1:T-1}^N), 0)], \quad (7)$$

where the inputs to $\mathcal{D}$ ($S_{2:T}^N - S_{1:T-1}^N$ and $\tilde{S}_{2:T}^N - \tilde{S}_{1:T-1}^N$) are the RF feature differences between two adjacent time steps, and the fake samples $\tilde{S}_{1:T}^N$ are generated through $F(X_{1:T}^{P'})$. $\mathcal{L}_{ce}$ is the cross entropy loss between the logits calculated by $\mathcal{D}$ and the actual labels y, defined as

$$\mathcal{L}_{ce}(logits, y) = -y\log(logits) - (1-y)\log(1-logits). \quad (8)$$

The RF feature differences help the adversarial training generate RF data with not only high fidelity but also temporal coherence comparable to real RF sensing data. During training and within each batch, half of the samples are fake data generated by $\mathcal{G}$ given an augmented pose through augmentation methods (either by TGNP or PoseMod), and the rest are real samples. This way, a more generalized discriminator is trained, which in turn helps to train a better generator. On the other hand, the Generator $\mathcal{G}$ tries its best to generalize an RF-activity-sensing plausible data for a given augmented pose to fool the Discriminator $\mathcal{D}$ through minimizing the following:

$$\min \mathcal{L}_{ce}(\mathcal{D}(\tilde{S}_{2:T}^N - \tilde{S}_{1:T-1}^N), 1). \quad (9)$$

In summary, the superiority of our adversarial learning lies in that, by learning a better discriminator with a weakly-supervised learning manner, we boost the ability of our neural kinematics RF transformer to generalize RF activity data with unprecedented diversity, while also creating labels (i.e., augmented poses) for generative RF data. During the learning process, the initial synthesized RF data from augmented poses are visually and physically invalid in the meaning of RF sensing, and are easily distinguishable from the ground truth RF data. Hence, there is high incentive to improve the generator $\mathcal{G}$ to better fool $\mathcal{D}$ and generate results of better quality. We also find that the training process would converge faster and get relatively better performance with a pretrained generator.

C. Augmentation techniques for poses

The privacy of test subjects is largely preserved during our data collection process, since our ground truth 3D pose data are in the format of 3D joint locations (x, y, z). The Python Numpy files are converted to animation frames using the Blender software [41] with bones and meshes that can be freely manipulated. These pose data will then go through pose augmentation for enhanced diversity. We design two methods for pose augmentation, one with an emphasis on automation (termed TGNP), and the other with a focus on richer diversity (termed PoseMod).

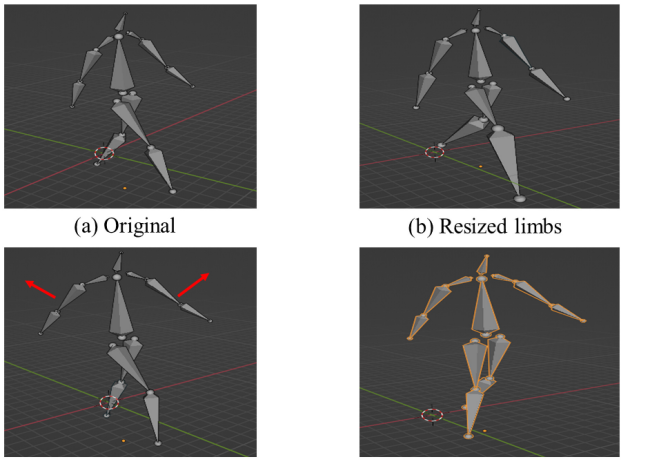
1) TGNP. The TGNP augmentation method aims to make the augmentation process simple, swift, and virtually automatic, yet effectively introducing pose diversity and model robustness. TGNP has real-life inspirations in which human motions in videos can often be occluded for a period of time, resulting in a noisy 3D pose. Given a ground truth 3D pose $X_{1:T}^P$ over time frames T, we randomly perturb $p < P$ joints within $j \leq T$ frames with additive Gaussian noise $\mathcal{N}(0, \sigma^2)$ to create $X_{1:T}^{P'}(TGNP)$. The variance needs to be relatively small, so that the contaminated signal still have an SNR in the range of 30 dB to 50 dB. An SNR smaller than 30 dB will cause a large disturbance to joint locations and hence result in uncanny poses, but an SNR larger than 50 dB will

only introduce negligible perturbations. On the other hand, Gaussian noises added to all time frames could introduce too much jittering, rendering the pose and potentially the transformed RF data lacking temporal smoothness. In addition, we also briefly modify the skeleton sizes after a certain amount of time to induce proper diversity at low cost when operating this augmentation method.

2) PoseMod. The PoseMod augmentation method emphasizes curating poses to introduce richer and more sophisticated diversity. We try to introduce perturbations to various skeletons and activity-specific pose movements, and find that skeleton sizes, pose movement variations, and frequencies, along with locations and camera viewpoints, are the key features for augmentation. By artificially introducing small perturbations to these features, PoseMod can increase the diversity of generated RF data, thereby induce higher robustness in RF-CRGAN training. Unlike TGNP that only introduce small perturbations to a pose, PoseMod can also generate new poses, which can then be transformed into plausible RF data seamlessly corresponding to the augmentation operations. There are mainly three operations for PoseMod as illustrated in Fig. 3.

- Skeleton and limb size: The overall size of a skeleton (height and width) and the lengths of limbs can be freely modified using Blender [41]. Such modifications correspond to the use of different test subjects.
- Pose movement variations: Modify the positions of certain joints during the key frames of a pose (the frames at which a key activity-specific movement is being performed, e.g., the locations of arms and legs reaching a peak location before moving backward). Furthermore, the frequencies of pose movements can also be adjusted. Our arguments for such operations are that the RF-CRGAN network is intended to learn the joint kinematics of poses across time steps, hence the movement variations and frequencies of the poses are the most dominant features, especially at joint locations such as arms and legs where the motions are the most obvious. This operation offers simple and meticulous options during augmentation without sacrificing the fidelity of natural pose movements.
- Global motions: Adjust translation regarding the physical x-y plane, and rotation regarding the z plane. This is to simulate a real-life data collection scenario where the target subject needs to change locations and orientations in front of both RF devices and Kinect camera. Such modifications make the pose data to be situation aware, which can then be transferred into RF data for more generalized data synthesis.

For each activity class, we perform the second and third operations on various modified subject skeletons achieved by the first operation. We term the augmented pose through PoseMod $X^P_{1:T}(PoseMod)$. We then define a general augmented pose using either of the methods for training the RF-CRGAN model in Section IV-B as $X^{P'}_{1:T} = \{X^P_{1:T}(TGNP), X^P_{1:T}(PoseMod)\}$. Our pose augmentation techniques conform to the designs in several state-of-the-art 3D pose estimation works [43]–[45], while having the unique novelty in that our augmented poses are for generating new AIGC RF data, instead of only aiming at improving the generalization of the 2D-to-3D pose estimation network.



(a) Original (b) Resized limbs (c) Altered movement variations (d) Different locations and viewpoints

Figure 3. Examples of the PoseMod augmentation method.

V. Implementation and Experimental Study

A. Prototype system

To evaluate the proposed RF-AIGC framework, a prototype is developed using several representative RF technologies, including UHF RFID, 2.4GWiFi, 5G WiFi, and FMCW radar. An off-the-shelf Impinj R420 reader, passive ALN-9634 (HIGG-3) tags, and three S9028PCR polarized antennas make up the RFID platform. The RFID system hops among 50 channels from 902 MHz to 928 MHz, and it remains on each channel for 0.2 s. A standard Intel 5300 network interface card (NIC) operating at either 2.412 GHz or 5.3 GHz is used as the WiFi CSI platform. Finally, an IWR1843 Boost single-chip FMCW mmWave sensor operating at 76 $\sim$ 81 GHz is deployed in the mmWave platform. We use a Lenovo laptop with a GTX 1660 Ti GPU for signal processing, model training, and inference. The system setup is illustrated in Fig. 4.

To demonstrate the advantages of the data augmentation strategy, the synthesized RF data are used for training the models of two different downstream tasks: (i) 3D pose estimation (a regression task) and (ii) Human activity classification (HAC, a classification task). We train the same machine learning model in these two tasks using synthesized data, augmented data, and real data, respectively, and compare the performance of the trained models.

B. Dataset construction and composition

RF data is collected by sampling activities performed by a test subject in front of the RF sensing platforms and a Kinect 2.0 device. The individual performs eleven types of distinct activities, including drinking (DK), raising the left arm (LA), raising the left leg (LA), standing still (ST), squatting (SQ), boxing (BX), twisting (TW), walking (WA), kicking (KI), waving up and down (UD), and weight lifting (WL). Data is sampled when the participant continuously repeats the full range of specific activities. The configurations for

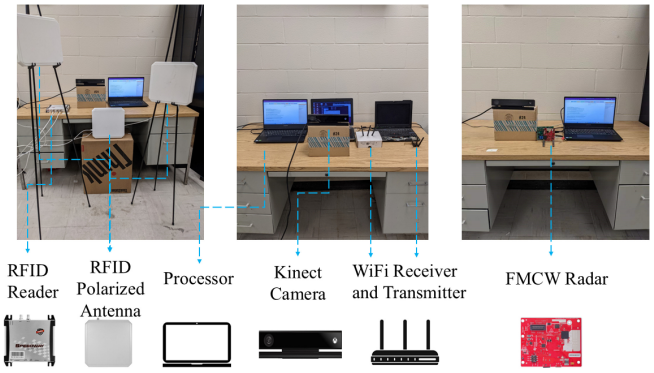


Figure 4. The configuration of the experimental RF-AIGC system.

three RF sensing platforms are as follows. We conduct RFID-based sampling by attaching 12 passive RFID tags to 12 selected joints on the subject's body, including the neck, left shoulder, left elbow, left wrist, right shoulder, right elbow, right wrist, pelvis (root joint), left hip, left knee, right hip, and right knee. We use a reader with three polarized antennas to interrogate the tags, which helps make sure that each RFID tag is covered by at least one antenna. As for WiFi platforms, we set the WiFi transmitter to the *injection* mode, and the receiver to the *monitor* mode. The frequency bands used are 2.412 GHz and 5.3 GHz, respectively. This setup enables us to measure the impacts of different bands under the same WiFi platform. Last but not least, the FMCW radar of model IWR1843 Boost is used to generate range profiles for the scanned area where the test subject locates. These four settings can all independently capture human activities well. The Kinect device will capture vision data that is synchronized with the RF data for supervised training of the models.

To comprehensively examine the performance of the RF-AIGC framework, a holistic evaluation plan of real and synthesized data is designed. For real data, we use limited and sufficient amounts of training data as baselines. If training data is limited and homogeneous (e.g., little diversity within the dataset), the performance is less than satisfactory, while a model trained on sufficient and heterogeneous real data (higher diversity including subjects, activities, and locations) can achieve excellent results. The generated new RF sensing data are synthesized data by RF-AIGC. The downstream task models, i.e., for 3D pose prediction and HAC, trained on synthesized data alone can have adequate performance. Synthesized data is also utilized to augment the real data for better and more robust model performance. It often takes a few times more synthesized data for the model performance boost to be effective due to the domain gap problem of GANs.

For simplicity, we would abbreviate these dataset terms when needed in figures or tables: "lim." for limited, "suffi." for sufficient, and "synth." for synthesized. Six test subjects are involved in the collection of real RF and Kinect data, denoted by S1, S2, ..., S6, respectively. S1, S2, and S3 are collected under homogeneous settings (i.e., similar body shapes, viewpoints, and locations), while S4, S5, and S6 are collected under heterogeneous settings. S4 and S5 are of similar body shapes and at the same location, which is different from S1 to S3. S6 has a different body shape, activity-related movement variations, locations, and viewpoints compared to S1 to S3 or S4 to S5. For each of the test subjects, we collect data for the eleven activities each for around 7.8 minutes, reaching a total of 46.9 minutes. In order to capture the temporal dependencies, a sliding window of 3 seconds with a sliding rate of 1 second is used on collected and processed data files to obtain 5 basic data units of the same length of 3 seconds. The 1-second sliding factor is chosen to create more diversity within the collected data since activities can change moderately in this time window. In the end, we obtain a total of 99 minutes of training-ready data (9 minutes for each activity) for both Kinect and RF data. Test data are separate from training data and are dynamic in different testing scenarios.

C. Quality of synthesized RF data

We first examine the quality of RF-AIGC generated data. The Structural Similarity Index (SSIM) [46] has been used in computer vision to evaluate the brightness, contrast, and structural quality of reconstructed images. Since SSIM uses the structural index to measure not only the mean intensity and standard deviation, but also the specifics and overall pattern of features inside a picture, it has also been acknowledged as a valuable metric for assessing how similar the synthesized data is to the real data. Similar to how the structural index is naturally suited for locating the key features across all pixels in an image, it can also be used to analyze the pattern of movement features across time frames and RF features. For real data x and synthesized data x', SSIM is defined as

$$SSIM(x, x') \triangleq \frac{(2\mu_x\mu_{x'} + C_1)(2\sigma_{xx'} + C_2)}{(\mu_x^2 + \mu_{x'}^2 + C_1)(\sigma_x^2 + \sigma_{x'}^2 + C_2)}, \quad (10)$$

where μ_x and $\mu_{x'}$ denote the mean intensity, and σ_x and $\sigma_{x'}$ the standard deviation of real and synthesized data, respectively, and $\sigma_{xx'}$ is their covariance. C_1 and C_2 are constants. SSIM takes a value between 0 and 1. A value of 1 indicates full structural similarity between real and synthesized data.

As can be seen in Fig. 5, for a specific activity of boxing, the RF-AIGC synthesized RF data visually conforms to the ground truth RF data in terms of sharpness, contrast, and brightness to a large extent across all the four RF platforms. Furthermore, we utilize SSIM score to measure the quality of our synthesized RF data compared to our collected ground truth RF data. SSIM score involves the product of three measurement metrics including luminance, contrast, and structure. Luminance and contrast seamlessly grasp how similar the values learned by RF-AIGC are to the ground truth RF data in terms of mean and variance. The structure index, by utilizing covariance, measures how well our RF-AIGC network learns the short-time delicate movement details and long-time frequency patterns of the collected ground truth RF data. The SSIM map shown in Fig. 5, where each pixel is evaluated with a score (yellow denotes high similarity, and blue for low similarity), visually and quantitatively confirms the quality of these synthesized samples.

As shown in Table I, our proposed RF-AIGC system is able to achieve satisfactory SSIM scores across all the four

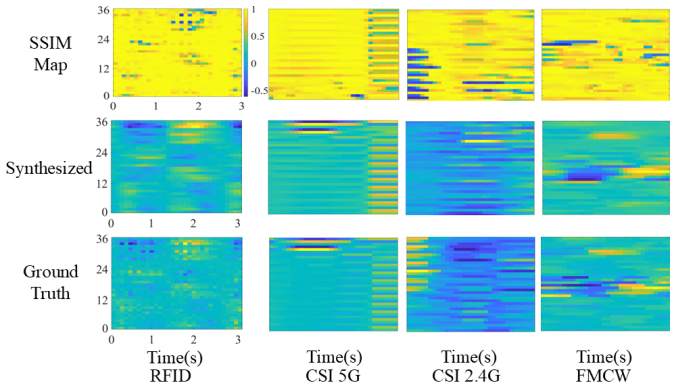


Figure 5. Demonstration of real RF signal examples, the corresponding synthesized RF signal examples, and SSIM score maps that specifically show how well the synthesized RF data match the real RF data.

Table I
SSIM SCORES ACHIEVED BY RF-AIGC FOR THE FOUR RF PLATFORMS

RF Platforms	SSIM Score ↓	SSIM Structure Score ↓
RFID	0.8995	0.9310
5G WiFi	0.8363	0.8675
FMCW Radar	0.8282	0.8563
2.4G WiFi	0.7473	0.7718

RF platforms. The RFID Platform achieves a high score close to 90% and a high structure score of 93.1%. 5G WiFi and FMCW radar platforms have similar scores, with 5G WiFi being slightly better. 2.4G WiFi, however, has a relatively lower score albeit visually still quite competitive in sharpness and brightness. This is because the 2.4 GHz WiFi has a larger coverage and is more susceptible to interference and movements in the surroundings.

In our RF-AIGC, PoseMod allows users to precisely control and modify many factors of the pose data, and TGNP to introduce Gaussian noises to the pose data. As a result, in addition to performing exceptionally well at generating homogeneous data that is similar to the training set (which results in high SSIM scores), our model also produces high-quality RF data with *considerable diversity*, which is necessary to train a strong model. Since SSIM is not adequate to capture such diversity, we employ the Frechet Inception Distance (FID) [47] to assess how close the distributions of generated and actual RF data are through the distance in the high dimensional latent space between the feature vectors from both parties. The lower the FID score, the higher the fidelity of the generated RFID data as compared to the real data. We also incorporate two additional metrics, termed *diversity* and *multi-modality*, to measure the inter-activity-class and intra-activity-class variance within generated and real RF data. Contrary to FID, the higher the metric value, the stronger the diversity and multimodality. FID, diversity, and multimodality, together as a holistic metric system, can help to comprehensively quantify the model performance on synthesizing RF data. The three metrics are defined as:

$$\text{FID} = \|\mu - \mu'\|_2^2 + \text{Tr}(\Sigma + \Sigma' - 2\sqrt{\Sigma \times \Sigma'}) \tag{11}$$

$$\text{Diversity} = \frac{1}{S_{div}} \sum_{i=1}^{S_{div}} \|f_i - f_i'\|_2 \tag{12}$$

Table II
COMPARISON OF FID, DIVERSITY, AND MUTLIMODALITY SCORES FOR GENERATED AND REAL RF DATA

	FID ↓	Diversity ↓	Multimodality ↓
PoseMod Synth.	$58.128^{\pm 0.103}$	$10.843^{\pm 0.266}$	$9.008^{\pm 0.317}$
TGNP Synth.	$50.500^{\pm 0.091}$	$9.594^{\pm 0.287}$	$8.058^{\pm 0.414}$
Sufficient Real	$6.216^{\pm 0.025}$	$9.329^{\pm 0.230}$	$8.392^{\pm 0.391}$
Limited Real	$4.548^{\pm 0.008}$	$8.584^{\pm 0.243}$	$7.353^{\pm 0.409}$

Table III
COMPARISON OF FID SCORES FOR SELECTED SYNTHESIZED ACTIVITY CLASSES

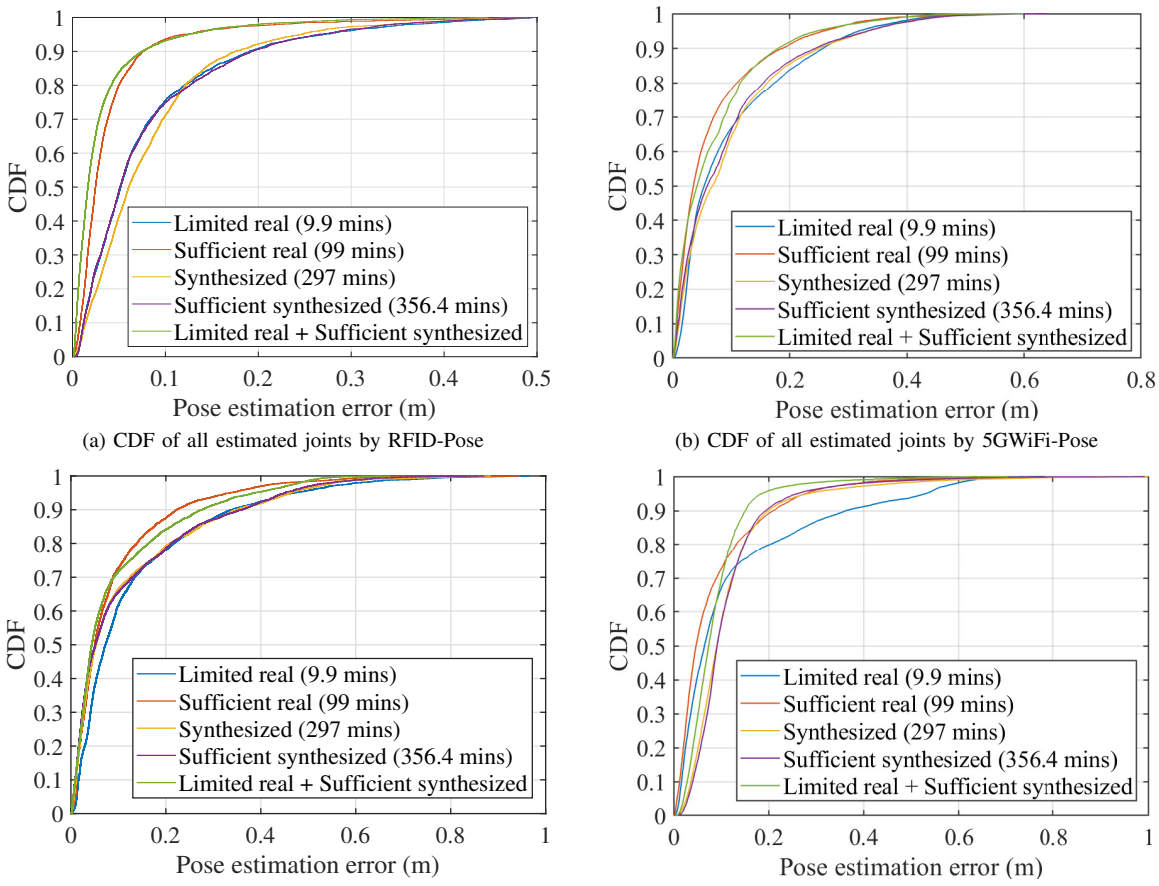
Standing still	Waving	Walking	boxing
$35.293^{\pm 0.034}$	$31.564^{\pm 0.041}$	$44.698^{\pm 0.096}$	$70.344^{\pm 0.113}$

$$\text{Multimodality} = \frac{1}{Z S_{mul}} \sum_{z=1}^{Z} \sum_{i=1}^{S_{mul}} \left\| f_{z,i} - f'_{z,i} \right\|_2, \tag{13}$$

where Σ and Σ' denote the covariance matrices of the real and generated feature vectors, respectively, μ and μ' denote the feature-wise means of the real and generated feature vectors, respectively, and Tr$(\cdot)$ refers to the trace linear algebra operation. The feature vectors between the two distributions are obtained by the `inceptionv3` neural network model [48], and f_i and f_i' represents the feature vectors. For diversity, we randomly select two subsets of samples of the same size S_{div}, and calculate the variance of the RF data across all activity classes. S_{div} is set to 200 in our experiments. For multimodality, we choose a set of RF data with Z action types. For each action z, we randomly sample two subsets with the same size of S_{mul} to calculate the diversity within each class.

The overall metrics for all activity classes are shown in Table II. For fair comparison, each experiment is repeated 25 times, and a statistical interval with 95% confidence is presented. Among the four metrics, FID is the most important indicator in evaluating the overall performance of a model. A better FID helps different downstream tasks tremendously, but diversity and multimodality are also important especially when FID is not optimal. The table shows that limited real data has the lowest diversity and multimodality score, while sufficient real data has great overall scores regarding a low fidelity, and higher diversity and multimodality scores. Synthesized data from either augmentation method have evidently lower fidelity scores in exchange for better diversity and multimodality. PoseMod offers more adaptability than TGNP, and poses more challenges for GAN to synthesize, hence the lower fidelity. There is still a non-negligible domain gap between generated data and real data. Generally, better diversity and multimodality mean better generalization. So with the cost of using computers to generate a large amount of synthesized data, the augmented dataset can achieve a performance on par with sufficient real data. We randomly select 80 synthesized RF data for four representing activity classes and demonstrate their FID scores in Table III. For a simpler activity involving only limb movements such as waving up and down, the FID is relatively better, while complex activities involving four limbs have worse FID scores.

This article has been accepted for publication in IEEE Internet of Things Journal. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2024.3482256



(a) CDF of all estimated joints by RFID-Pose

(b) CDF of all estimated joints by 5GWiFi-Pose

(c) CDF of all estimated joints by 2.4GWiFi-Pose

(d) CDF of all estimated joints by FMCW-Pose

Figure 6. Pose estimation errors achieved by models trained on five different combination of real and synthesized data for different sensing methods.

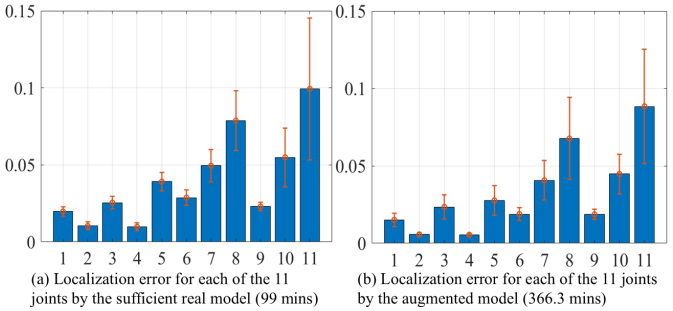
D. Downstream task I: RF-Pose prediction

For 3D pose estimation task, the mean per joint position error (MPJPE) for each time frame evaluates the distance between the RF-predicted joint positions and the ground truth after aligning the root joint. The precision indicates the fine-grained RF features learned by our synthesized RF data, which can be transformed into plausible 3D geometric structures. To evaluate how PoseMod augmentation affect MPJPE, we first perform testing on seen data classes during training (i.e., S1, S2, and S3) with five different combinations of training data. From Fig 6, we can see that the RFID platform has the best overall median estimation error across 4 RF platforms. This can also be seen from the largest median error being less than 50cm, while other platforms can be as large as 100cm, causing total joint displacements. The reason is that there are a fair amount of outlier joints for the estimated pose. The blue curve and red curve denote the limited real model (S1, S2, and S3) and the sufficient real (S1, S2, S3, S4, S5, and S6) model, respectively. They serve as baselines for limited (5.62cm) and optimal (2.26cm) performance achieved by real data only. The yellow and purple curves indicate synthesized training models with increasing data. The green curve highlights the superiority of our synthesized RF data in that combined with only a limited amount of real data, the augmented model achieves performance better than sufficient real model, reaching 1.72cm. This is where diversity and multimodality come into play since without PoseMod, median error is 3.78cm even when there are only five types of activities [8]. Fig. 7 shows that the position error for each of the 11 joints of the augmented model is smaller than that of the sufficient real model. It is important to note that for synthesized model, satisfactory pose estimation performance is achieved (6.08cm), but with flaws such as a lack of temporal smoothness and sometimes, extreme outliers for some joint positions. This is largely due to the domain gap between synthetic and real data, as can be inferred from the FID score.

Because of the limited set up for 3D pose estimation (lack of transcievers for an all-around scanning of the human body) for the other three platforms, the baseline estimation performance is suboptimal. However, it is evident that augmented models are able to achieve performance boost (more than 20%) over limited models for 5GWiFi-Pose and 2.4GWiFi-Pose. As for the FMCW platform, the augmented model achieves a worse median error than limited real model, but still excels overall where the 90 percentile error is 15 cm, significantly better than 35.97cm from the limited model.

One of the main purposes of the proposed system is to improve the pose estimation performance across seen skeletons and activity types when there is only a limited number of real data available, and the above results have validated this design goal. However, the more impressive characteristics of the system is to surgically augment desired data under different scenarios. Activity classes and subject skeletons are the two

This article has been accepted for publication in IEEE Internet of Things Journal. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2024.3482256



(a) Localization error for each of the 11 joints by the sufficient real model (99 mins)

(b) Localization error for each of the 11 joints by the augmented model (366.3 mins)

Figure 7. A comparison of localization error for each of the 11 joints for the activity type of walking between sufficient real and augmented model.

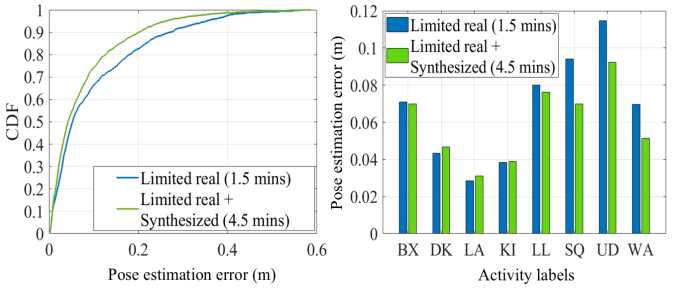


Figure 8. The advantages of augmentation illustrated via cross activities testing (test the trained RF-Pose network on unseen activities) with a comparison of CDF curves and bar graphs on 8 activity classes.

most important factors in pose estimation tasks, and the experiments below can effectively showcase such advantages.

1) Cross-activity testing. When a model is specifically trained on certain activities, it is possible that the model learns to only reconstruct the similar activity classes, and might break down when it comes to unseen activity classes during training. In Fig. 8, we plot the CDF curves and bar graph results of the two models trained with different training data. The baseline model is trained with 5 classes of real activities, and the augmented model is trained with both real activities and 3 unseen classes of synthesized activities. There are 1.5 minutes of data in each real class, and 4.5 minutes in each synthesized class. Section V-D has shown the abilities of the pose augmentation techniques for elevating the pose estimation through an extensive amount of synthesized data, while the results here demonstrate that our data augmentation can be quick and effective for improving the generalization of pose estimation models on untrained activity classes with a relatively lower cost of generating synthesized data. As the CDF curve shows, the model performs worse when the activity class is not seen during training. In the bar graph, unseen activity classes can have errors higher than 10 cm, while none of them estimated by augmented models reach higher than 10 cm, and it is clear that the augmented model achieves a performance boost on each of the unseen activity classes.

2) Cross-skeleton testing. Subject skeletons play a crucial role in RF-Pose estimation tasks. We conduct experiments using four subjects to study how unseen skeletons affect the pose estimation performance and how pose augmentation mitigates the performance degradation. The training dataset is made up of the homogeneous S1, S2, and S3 data, while one untrained subject S4 is used for testing. The model trained on S1, S2, and

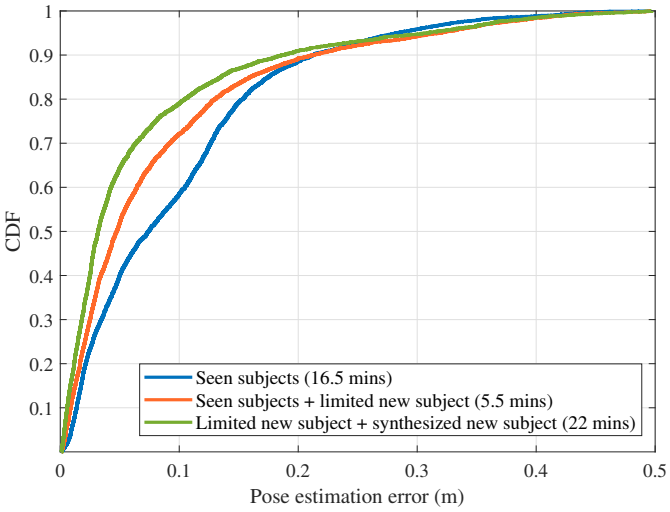


Figure 9. The advantages of augmentation illustrated via cross-skeleton testing (test the trained RFPose predictor on unseen subjects) with a comparison of CDF curves across all activity classes.

S3 suffers greatly and only achieves a median error of 7.41cm. Furthermore, even with the addition of a limited amount of real data from S4, the median error of 4.72cm is still not optimal. However, when we add 22 minutes of synthesized data of S4, the model can attain a median error of 3.19cm, improving considerably. We can see from Fig. 9 that the overall CDF is worse than that of Fig 6 considering the less amount of data we use. This experiment helps validate the cost-effectiveness of our synthesized data on unseen test subjects.

Holistic results of pose augmentation on RF-Pose estimation performance are shown in Fig. 10. The baseline results (Limited and Sufficient real model) are shown in light blue and red curves. When no augmentations are applied, median errors of 5.62 cm and 2.26 cm are achieved by the limited real and Sufficient real model, respectively. The failure of limited real model is caused by overfitting over the small number of training examples. The sufficient real model has excellent performance at the cost of far more data collection time. The augmented results improve as the number of TGNP synthesized training examples increase, up to the saturation of around 2.51 cm where adding more synthesized data with different skeletons and iterations of temporal Gaussian noises fail to further improve the classification results, and the augmented results do not surpass the Sufficient real model. The cyan and dark blue curves show that it takes the more diverse PoseMod augmentation method and around 3 times the amount of synthesized training data to achieve a superior performance of 1.72cm all the way from 5.62cm. In [49], the authors showed recent text-based generative models also experience the same synthetic data efficiency problem.

E. Downstream task II: RF-HAC classifier

Our synthesized RF data can also be deployed effectively for Human Activity Classification (HAC), as they consist of high-quality differentiable features for distinct human activities. The recognition accuracy indicates the correlation of the RF sensing data and its activity class.

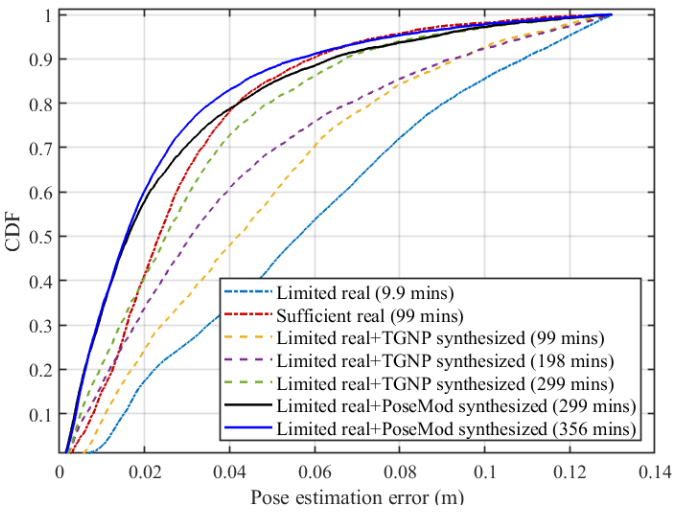
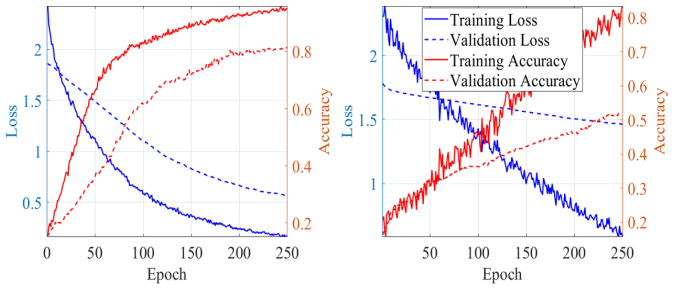


Figure 10. CDF plots of model performance with the increase of training set size and diversity



(a) Limited real+TGNP synthesized (99 mins) HAR (b) Limited real (10.5 mins) HAR

Figure 11. RF-HAC training and validation performance across 250 epochs with the RFID platform.

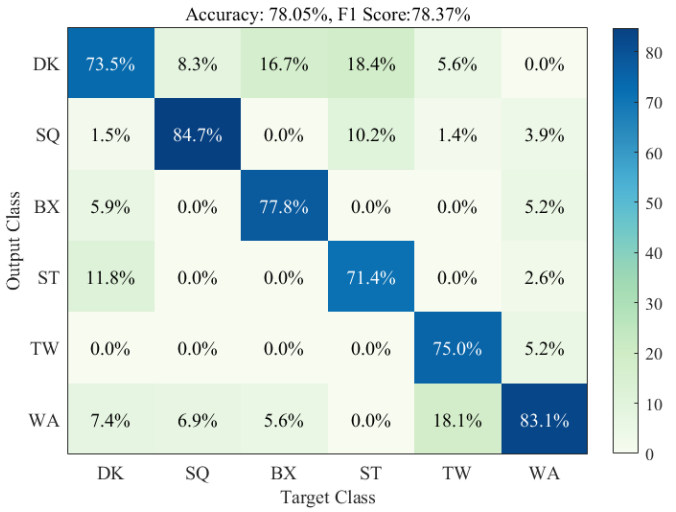


Figure 13. Synthesized model performance with 5-shot of real data per class

In this study, a custom 2D Convolutional Neural Network (CNN) model is first utilized to extract salient features of the RF data, and then a Bidirection LSTM network is employed for the final human activities features classification. In the feature extraction network, there are 4 Convolutional layers with a filter size of 3 and the numbers of filters are 16, 32, 64, and 1, respectively. All Convolutional layers use the "relu" activation function. Convolutional layers 1-3 use "same" padding whereas Convolutional layer 4 applies "valid" padding. The Convolutional layer 2 is followed by a Maxpooling2D layer with a pool size of 2, further followed by a dropout layer 0.5. Every other Convolutional layer is tailed by only the dropout layer. Convolutional layer 4 with a filter size of 1, along with a reshape layer, is used for dimensionality reduction which prepares the output for bi-directional LSTM networks. This is because LSTM networks require the input to be 3 dimensional with the format of samples, time steps, and features, while the output of a typical 2D CNN network consists of an extra dimension of channels. The LSTM classifier network has a relatively simpler architecture where two LSTM layers work in parallel to make up a bi-directional LSTM layer. In contrast to the unmodified input that is fed into the first layer, the input to the second layer is a reverse replica of the data. A bi-directional LSTM layer can help capture the information from both the future and the past. Both of the networks use the Adam optimizer with a learning rate of 0.0001. The batch size is set to 16. Dropout layers are important for preventing the model from overfitting.

We find RFID is more resilient to environment interference. Such advantages enable RFID data to have richer features of human motions. Furthermore, more reliable human movement features can be better conveyed through RFID tags attached to the subjects joints than by other platforms, especially in dynamic environments. Just as the RF-Pose predictor has better performance on RFID data, RF-HAC achieves superior classification performance on both accuracy and F1 score for the RFID platform. For the RFID-based human activity feature classifier, a simple dense layer with a softmax activation function without bi-directional LSTM layers is capable of achieving high classification accuracy as well. The training and validation performance can be found in Fig. 11, where the augmented model (using 10.5 mins of real data and 99 mins of synthetic data) converges much faster than the limited real model (using 10.5 mins of real data). For the limited real model, the validation accuracy only reaches lower 50% when the training accuracy is 84%. This is a serious overfitting problem. The training curves perfectly explain why the limited real model suffers when it comes to HAC. On the other hand, even though the overfitting problem is still present, the gap between the training and validation performance has reduced significantly. There are 32 minutes of test data in total for the 6-class RF-HAC task. The test data are real data separately collected from training data and belong to S1, S2, S3, S4, and S5.

The confusion matrices are presented in Fig. 12. The augmented model exhibits exceptional performance that is on par with the sufficient real model regarding both Accuracy and F1 score across the four platforms. However, it is worth noting that the synthesized models fail to achieve a decent performance on test data and have an overfitting problem. This is largely due to the domain gap issue discussed in various works using GAN-based synthesized data for augmentation [50], which is evidenced by the low FID scores. We synthesize new RF data using TGNP or PoseMod to obtain better diversity at the cost of fidelity since we do not use one-to-one mapped RF data from ground truth data. Sometimes, the synthesized model

This article has been accepted for publication in IEEE Internet of Things Journal. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2024.3482256

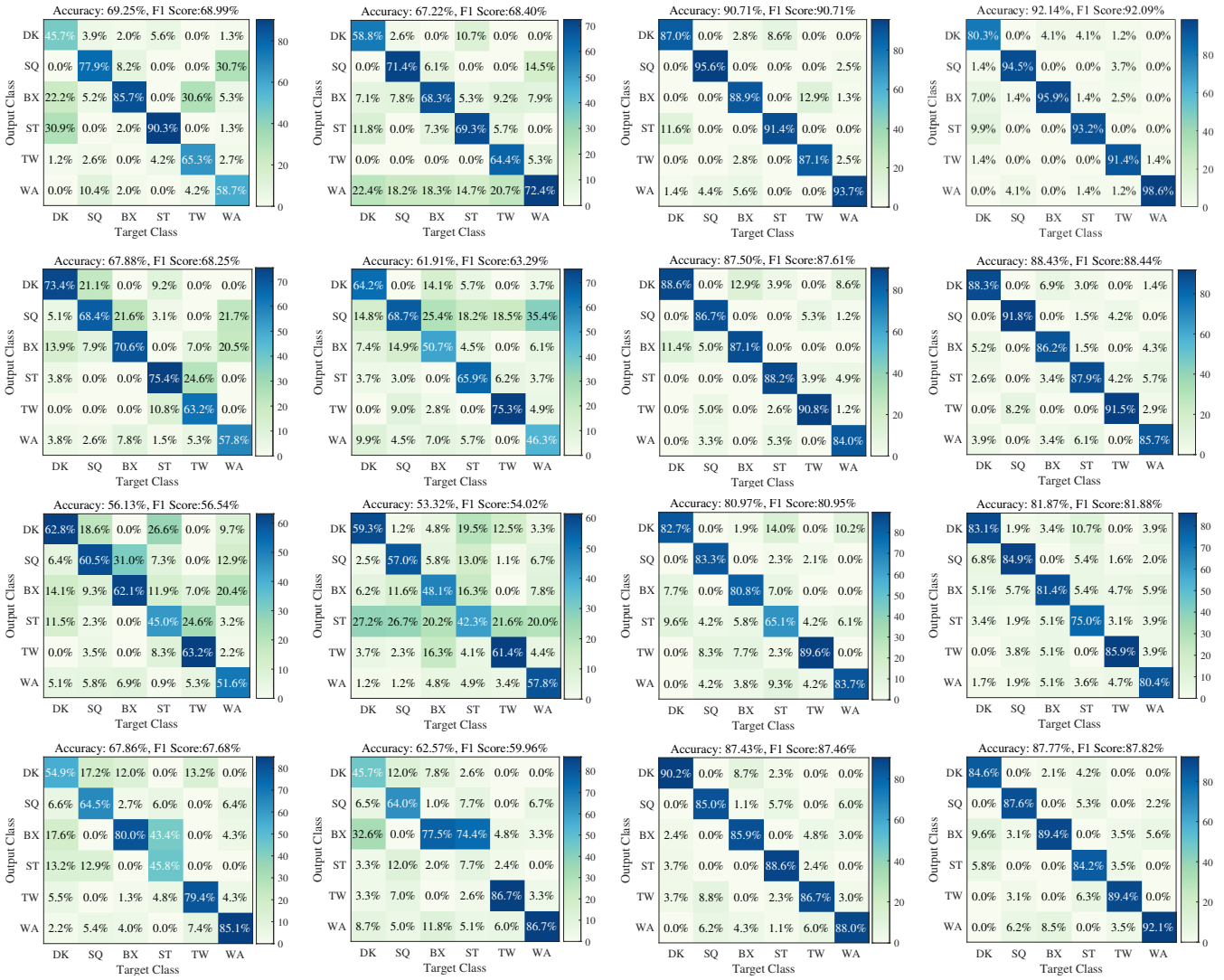


Figure 12. Confusion matrix of models with limited real, synthesized, sufficient real, and augmented data (limited real+PoseMod synthesized data), respectively from left to right, and in the order of RFID, WiFi 5G, 2.4G, and FMCW platforms from top to bottom.

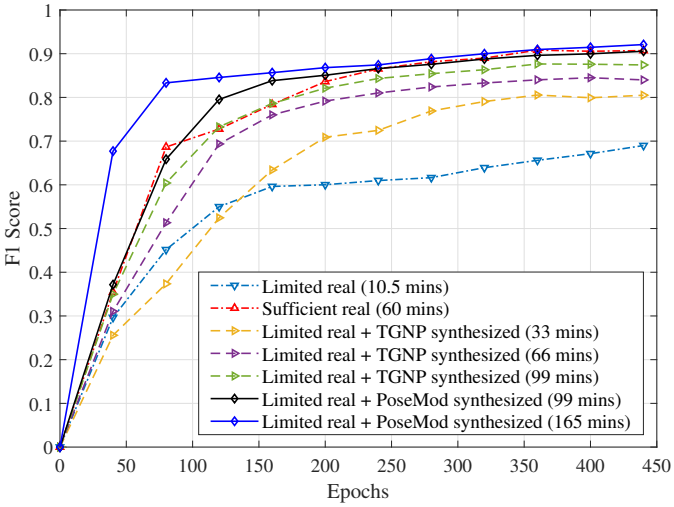


Figure 14. F1 scores of the models trained with the progressive increase of training dataset size and diversity

would collectively misclassify one particular class (walking) as shown in the top row for RFID platform in Fig. 12, and sometimes the model would have a considerable bias on one or two classes shown in the second row for the FMCW radar platform. The synthesized models still function well for some of the classes. When the training and testing classes are reduced to 3 classes, results of up to 85% can be achieved.

The synthesized model can be improved effectively with 5-shots of real RF data which are easily to collect. The metrics are raised to the upper 70% level as shown in Fig. 13, and the overfitting problem is effectively alleviated where one particular class is not dominantly recognized anymore. Despite the inadequate performance of the synthesized model, the five-shot classification is enabled on the basis of a large quantity of diversely synthesized data, which further demonstrates the benefits of our augmentation framework. RF-HAC requires less synthesized data than RF-Pose due to the reduced difficulty of the task from pose tracking to classification.

Fig. 14 shows a comparison of the F1 scores of the RF-HAC classifier models at different scales of augmented training samples and diversity across training epochs. Light blue and red curves are again the baseline for limited and sufficient model performance with real data. We find that adding more

TGNP synthesized data improves F1 scores to a saturated score of around 87%. Only with more diverse synthesized data (PoseMod), and almost 3 times the amount of training data as sufficient real data, can the augmented model truly achieve a superior performance. This conforms to the synthesized data efficiency as the performance of RF-Pose. When augmented with 331.2 minutes of synthesized data, the F1 curve reaches 0.92. Compared with the limited real model, there is around a 30% gain in F1 scores across almost all epochs. This experiment validates that the proposed RF-AIGC system can effectively improve the accuracy of CNN-based HAC.

VI. Conclusions

In this paper, we proposed an *AIGC for RF sensing* approach to address the challenge of lacking RF data, to enable more free-form RF data generation. The proposed RF-AIGC framework utilizes a recurrent GAN model conditioned on 3D human pose data to generate RF sensing data. The high quality and functionalities of the synthesized data by the proposed RF-AIGC system were demonstrated through the metrics of SSIM, FID, diversity, and multimodality, as well as two representative downstream tasks. The proposed RF-AIGC system not only achieved the generation of RF data with high quality and diversity across four different RF sensing platforms, but also significantly mitigated the prohibitive costs associated with traditional RF data collection methods.

References

[1] Y. Sun, M. Peng, Y. Zhou, Y. Huang, and S. Mao, "Application of machine learning in wireless networks: Key technologies and open issues," *IEEE Commun. Surveys Tuts.*, vol. 21, no. 4, pp. 3072–3108, Fourth Quarter 2019.

[2] C. Li, Z. Cao, and Y. Liu, "Deep AI enabled ubiquitous wireless sensing: A survey," *ACM Comput. Surv.*, vol. 54, no. 2, pp. 1–35, Mar. 2021.

[3] J. Zhang, F. Wu, B. Wei, Q. Zhang, H. Huang, S. W. Shah, and J. Cheng, "Data augmentation and dense-LSTM for human activity recognition using WiFi signal," *IEEE Internet of Things J.*, vol. 8, no. 6, pp. 4628–4641, Mar. 2021.

[4] X. Pan, A. Tewari, T. Leimkühler, L. Liu, A. Meka, and C. Theobalt, "Drag your GAN: Interactive point-based manipulation on the generative image manifold," in *Proc. ACM SIGGRAPH 2023*, Los Angeles, CA, July/Aug. 2023, pp. 1–11.

[5] D. Wang, J. Yang, W. Cui, L. Xie, and S. Sun, "Multimodal CSI-based human activity recognition using GANs," *IEEE Internet of Things J.*, vol. 8, no. 24, pp. 17 345–17 355, Dec. 2021.

[6] P. F. Moshiri, H. Navidan, R. Shahbazian, S. A. Ghorashi, and D. Windridge, "Using GAN to enhance the accuracy of indoor human activity recognition," *arXiv preprint arXiv:2004.11228*, Apr. 2020. [Online]. Available: https://arxiv.org/abs/2004.11228

[7] Z. Wang, C. Yang, and S. Mao, "Data augmentation for RFID-based 3D human pose tracking," in *Proc. IEEE VTC-Fall 2022*, London, UK, Sept. 2022, pp. 1–2.

[8] C. Yang, Z. Wang, and S. Mao, "RFPose-GAN: Data augmentation for RFID based 3D human pose tracking," in *IEEE RFID-TA 2022*, Cagliari, Italy, Sept. 2022, pp. 1–4.

[9] J. Wang, D. Vasisht, and D. Katabi, "RF-IDraw: Virtual touch screen in the air using RF signals," *SIGCOMM Comput. Commun. Rev.*, vol. 44, no. 4, pp. 235–246, Aug. 2014.

[10] H. Li, C. Ye, and A. P. Sample, "IDSense: A human object interaction detection system based on passive UHF RFID," in *Proc. ACM CHI 2015*, Seoul, Republic of Korea, Apr. 2015, pp. 2555–2564.

[11] H. Ding, J. Han, L. Shangguan, W. Xi, Z. Jiang, Z. Yang, Z. Zhou, P. Yang, and J. Zhao, "A platform for free-weight exercise monitoring with passive tags," *IEEE Trans. Mobile Comput.*, vol. 16, no. 12, pp. 3279–3293, Dec. 2017.

[12] Y. Wang, J. Liu, Y. Chen, M. Gruteser, J. Yang, and H. Liu, "E-Eyes: Device-free location-oriented activity identification using fine-grained WiFi signatures," in *Proc. ACM Mobicom 2014*, Maui, HI, Sept. 2014, pp. 617–628.

[13] K. Qian, C. Wu, Z. Zhou, Y. Zheng, Z. Yang, and Y. Liu, "Inferring motion direction using commodity Wi-Fi for interactive exergames," in *Proc. ACM CHI 2017*, Denver, CO, May 2017, pp. 1961–1972.

[14] Y. Ren, S. Tan, L. Zhang, Z. Wang, Z. Wang, and J. Yang, "Liquid level sensing using commodity WiFi in a smart home environment," *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 4, no. 1, pp. 1–30, Mar. 2020.

[15] S. Tan and J. Yang, "WiFinger: Leveraging commodity WiFi for fine-grained finger gesture recognition," in *Proc. ACM MobiHoc 2016*, Paderborn, Germany, July 2016, pp. 201–210.

[16] K. Chetty, G. E. Smith, and K. Woodbridge, "Through-the-wall sensing of personnel using passive bistatic WiFi radar at standoff distances," *IEEE Trans. Geosci. Remote Sens.*, vol. 50, no. 4, pp. 1218–1226, 2012.

[17] H. Li, A. Shrestha, H. Heidari, J. L. Kernec, and F. Fioranelli, "Activities recognition and fall detection in continuous data streams using radar sensor," in *Proc. 2019 IEEE MTT-S Int. Microwave Biomedical Conf.*, Nanjing, China, May 2019, pp. 1–4.

[18] F. Adib, Z. Kabelac, D. Katabi, and R. C. Miller, "3D tracking via body radio reflections," in *Proc. USENIX NSDI 2014*, Seattle, WA, Apr. 2014, pp. 317–329.

[19] D. Zhang, K. Masahiko, and I. Takayuki, "FMCW radar for small displacement detection of vital signal using projection matrix method," *Hindawi Int. J. Antennas Propag.*, vol. 2013, Nov. 2013.

[20] P. Molchanov, S. Gupta, K. Kim, and K. Pulli, "Short-range FMCW monopulse radar for hand-gesture sensing," in *Proc. 2015 IEEE Radar Conf.*, Arlington, VA, May 2015, pp. 1491–1496.

[21] S. Tan, Y. Ren, J. Yang, and Y. Chen, "Commodity WiFi sensing in ten years: Status, challenges, and opportunities," *IEEE Internet of Things Journal*, vol. 9, no. 18, pp. 17 832–17 843, Sept. 2022.

[22] C. Yang, X. Wang, and S. Mao, "RFID-Pose: Vision-aided 3D human pose estimation with RFID," *IEEE Transactions on Reliability*, vol. 70, no. 3, pp. 1218–1231, Sept. 2021.

[23] W. Jiang, H. Xue, C. Miao, S. Wang, S. Lin, C. Tian, S. Murali, H. Hu, Z. Sun, and L. Su, "Towards 3D human pose construction using WiFi," in *Proc. ACM Mobicom 2020*, London, UK, Apr. 2020, pp. 1–14.

[24] 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.

[25] X. Li, Y. Zhang, I. Marsic, A. Sarcevic, and R. S. Burd, "Deep learning for RFID-based activity recognition," in *Proc. ACM SenSys 2016*, Stanford, CA, Nov. 2016, pp. 164–175.

[26] E. Shalaby, N. ElShennawy, and A. Sarhan, "Utilizing deep learning models in CSI-based human activity recognition," *Springer Neural Comput. Appl.*, vol. 34, no. 1, pp. 5993–6010, Jan. 2022.

[27] A. D. Singh, S. S. Sandha, L. Garcia, and M. Srivastava, "RadHAR: Human activity recognition from point clouds generated through a millimeter-wave radar," in *Proc. 3rd ACM Workshop Millimeter-Wave Netw. Sensing Syst.*, Los Cabos, MX, Oct. 2019, pp. 51–56.

[28] H. Kwon, C. Tong, H. Haresamudram, Y. Gao, G. D. Abowd, N. D. Lane, and T. Plötz, "IMUTube: Automatic extraction of virtual on-body accelerometry from video for human activity recognition," *Proc. ACM Interactive, Mobile, Wearable and Ubiquitous Technol.*, vol. 4, no. 3, pp. 1–29, Sept. 2020.

[29] Z. Wang, D. Jiang, B. Sun, and Y. Wang, "A data augmentation method for human activity recognition based on mmwave radar point cloud," *IEEE Sensors Lett.*, vol. 7, no. 5, pp. 1–4, May 2023.

[30] J. Zhang, F. Wu, B. Wei, Q. Zhang, H. Huang, S. W. Shah, and J. Cheng, "Data augmentation and dense-LSTM for human activity recognition using WiFi signal," *IEEE Internet of Things J.*, vol. 8, no. 6, pp. 4628–4641, 2021.

[31] X. Zhang, Z. Li, and J. Zhang, "Synthesized millimeter-waves for human motion sensing," in *Proc. ACM SenSys 2022*, Boston, MA, Nov. 2023, pp. 377–390.

[32] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proc. NIPS 2014*, Montreal, Canada, Dec. 2014, pp. 2672–2680.

[33] M. Patel, X. Wang, and S. Mao, "Data augmentation with Conditional GAN for automatic modulation classification," in *Proc. ACM WiseML 2020*, Linz, Austria, July 2020, pp. 31–36.

[34] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *arXiv preprint arxiv:2006.11239*, Dec. 2020. [Online]. Available: https://arxiv.org/abs/2006.11239

[35] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. 2019 Conf. North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Minneapolis, MN, June 2019, pp. 4171–4186.

[36] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *Proc. ICML 2021*, Virtual Event, July 2021, pp. 8748–8763.

[37] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proc. IEEE/CVF CVPR 2022*, New Orleans, LA, June 2022, pp. 10 684–10 695.

[38] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, and I. Sutskever, "Zero-shot text-to-image generation," in *Proc. ICML 2021*, Virtual Event, July 2021, pp. 8821–8831.

[39] J. Wang, H. Du, D. T. Niyato, Z. Xiong, J. Kang, B. Ai, Z. Han, and D. I. Kim, "Generative artificial intelligence assisted wireless sensing: Human flow detection in practical communication environments," *IEEE J. Sel. Areas Commun.*, vol. 42, no. 10, pp. 2737–2753, Oct. 2024.

[40] G. Tevet, S. Raab, B. Gordon, Y. Shafir, D. Cohen-or, and A. H. Bermano, "Human motion diffusion model," in *Proc. ICLR 2023*, Kigali, Rwanda, May 2023, pp. 1–16.

[41] B. O. Community, *Blender - A 3D modelling and rendering package*, Blender Foundation, Stichting Blender Foundation, Amsterdam, 2018. [Online]. Available: http://www.blender.org

[42] C. Yang, X. Wang, and S. Mao, "TARF: Technology-agnostic RF sensing for human activity recognition," *IEEE Journal of Biomedical and Health Informatics*, vol. 27, no. 2, pp. 636–647, Feb. 2023.

[43] K. Gong, J. Zhang, and J. Feng, "PoseAug: A differentiable pose augmentation framework for 3D human pose estimation," in *Proc. IEEE/CVF CVPR 2021*, Nashville, TN, June 2021, pp. 8571–8580.

[44] S. Li, L. Ke, K. Pratama, Y.-W. Tai, C.-K. Tang, and K.-T. Cheng, "Cascaded deep monocular 3D human pose estimation with evolutionary training data," in *Proc. IEEE/CVF CVPR 2020*, June 2020, pp. 6172–6182.

[45] H. Rhodin, M. Salzmann, and P. Fua, "Unsupervised geometry-aware representation learning for 3D human pose estimation," in *Proc. ECCV 2018*, Munich, Germany, Sept. 2018, pp. 765–782.

[46] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

[47] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "GANs trained by a two time-scale update rule converge to a local Nash equilibrium," in *Proc. NIPS 2017*, Long Beach, CA, Dec. 2017, pp. 6629–6640.

[48] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE/CVF CVPR 2016*, Las Vegas, NV, June 2016, pp. 2818–2826.

[49] R. He, S. Sun, X. Yu, C. Xue, W. Zhang, P. Torr, S. Bai, and X. QI, "Is synthetic data from generative models ready for image recognition?" in *Proc. ICLR 2023*, Kigali, Rwanda, May 2023, pp. 1–13.

[50] S. Sankaranarayanan, Y. Balaji, C. D. Castillo, and R. Chellappa, "Generate to adapt: Aligning domains using generative adversarial networks," in *Proc. IEEE/CVF CVPR 2018*, Salt Lake City, UT, June 2018, pp. 8503–8512.

Ziqi Wang [S'23] received the B.S. degree in electrical engineering from Auburn University, Auburn, AL, USA in 2022. He has been pursuing a PhD degree in the department of Electrical and Computer Engineering at Auburn University since 2023. His current research focuses on Artificial Intelligence of Things (AIoT) and wireless sensing. He is a recipient of IEEE ICC 2024 NSF student travel grant, and a co-recipient of Best Demo Award of IEEE INFOCOM 2024.

Chao Yang [S'18-M'22] Received his M.S. and Ph.D. degree in Electrical and Computer Engineering (ECE) from Auburn university, Auburn, AL. in 2017 and 2022, respectively. He is currently an Associate Professor with Hangzhou Institute of Technology, Xidian University, Hangzhou, China. His current research interests include wireless sensing, system security and wireless networks. He is a co-recipient of the Best Paper Award of IEEE GLOBECOM 2019, Best Demo Award of IEEE INFOCOM 2022, and 2022 Best Journal Paper Award of IEEE ComSoc eHealth Technical Committee.

Shiwen Mao [S'99-M'04-SM'09-F'19] is a Professor and Earle C. Williams Eminent Scholar, and Director of the Wireless Engineering Research and Education Center at Auburn University. Dr. Mao's research interest includes wireless networks, multimedia communications, and smart grid. He is the editor-in-chief of IEEE Transactions on Cognitive Communications and Networking, a Member-at-Large of IEEE Communications Society Board of Governors, and Vice President of Technical Activities of IEEE Council on Radio Frequency Identification (CRFID). He was the General Chair of IEEE INFOCOM 2022, a TPC Chair of IEEE INFOCOM 2018, and a TPC Vice-Chair of IEEE GLOBECOM 2022. He received the IEEE ComSoc MMTC Outstanding Researcher Award in 2023, the SEC 2023 Faculty Achievement Award for Auburn, the IEEE ComSoc TC-CSR Distinguished Technical Achievement Award in 2019, the Auburn University Creative Research & Scholarship Award in 2018, and the NSF CAREER Award in 2010, as well as several IEEE service awards. He is a co-recipient of the 2022 Best Journal Paper Award of IEEE ComSoc eHealth Technical Committee, the 2021 Best Paper Award of Elsevier/KeAi Digital Communications and Networks Journal, the 2021 IEEE Internet of Things Journal Best Paper Award, the 2021 IEEE Communications Society Outstanding Paper Award, the IEEE Vehicular Technology Society 2020 Jack Neubauer Memorial Award, the 2018 Best Journal Paper Award and the 2017 Best Conference Paper Award from IEEE ComSoc MMTC, and the 2004 IEEE Communications Society Leonard G. Abraham Prize in the Field of Communications Systems. He is a co-recipient of the Best Paper Awards from WCSP 2024, IEEE GLOBECOM 2023 (two), 2019, 2016, and 2015, IEEE ICC 2022 and 2013, and IEEE WCNC 2015, and the Best Demo Awards from IEEE INFOCOM 2024, IEEE INFOCOM 2022, and IEEE SECON 2017.