AIGC for RF Sensing: The Case of RFID-based Human Activity Recognition

Ziqi Wang and Shiwen Mao
Deptartment of Electrical and Computer Engineering, Auburn University, Auburn, AL 36849-5201, USA
Email: zzw0104@auburn.edu, smao@ieee.org

2024 International Conference on Computing, Networking and Communications (ICNC) | 979-8-3503-7099-7/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICNC59896.2024.10556013

***Abstract*—The performance of deep learning (DL) empowered wireless communications, networking, and sensing depends on the availability of sufficient high-quality radio frequency (RF) data, which is more difficult and expensive to collect than other types. To overcome this obstacle, we propose to harness the power of diffusion models on latent domains to generate hyper-realistic RF data for RF sensing. We develop a novel lightweight AIGC framework centered on latent domains, termed RFID-ACCLDM (Activity Class Conditional Latent Diffusion Model), to generate large quantities of RF data at low cost, conditioned on activity class labels. We demonstrate the high quality of RFID-ACCLDM generated data via the Frechet Inception Distance (FID) metric, along with a representative downstream task of human activity recognition (HAR). The model trained with synthesized data outperforms its counterpart trained by real data.**

***Index Terms*—Artificial intelligence generated content (AIGC), Conditional diffusion, Data augmentation, Human activity recognition (HAR), Radio frequency (RF) sensing.**

I. Introduction

The past decade has witnessed significant progress deep learning (DL)-based wireless communications and networking [1]. However, the availability of vast amounts of high-quality radio frequency (RF) data is a major determinant of the efficacy of most DL-based methods. RF data possesses unique randomness features and is much more difficult to collect than images or texts. First, RF data is very sensitive to the open-space propagation environment; any variation in the transceiver location or the surroundings could create a different data domain. Second, transceiver devices, waveforms, frequency bands, and protocols all have a significant impact on measured RF data. Third, the wireless channel is also time-dependent, exhibiting large variations over the time of the day, day of the week, and months. Because of such temporal, spectral, and spatial dependencies, collecting RF datasets is an extremely costly task, not to mention that a collected RF dataset might only be used to a limited extent in a different setting. As a result, the first obstacle to overcome in making "ML/AI for wireless" successful is obtaining RF data with high fidelity and diversity while keeping costs low.

Artificial intelligence-generated content (AIGC) has emerged as a significant trend recently. Unprecedented systems such as ChatGPT, DALL-E, and Gemini are leading the way towards Artificial General Intelligence (AGI). Transformers and diffusion models are commonly used as the backbone for these applications, which are mostly developed in the context of text-to-image generation or text-prompted AI agents. *Can we harness the power of AIGC to tackle wireless communication problems, especially generating hyper-realistic RF data*? As an earlier generation of AIGC technology, Generative Adversarial Networks (GANs), have been investigated for data augmentation [2]–[4]. However, GANs can only be leveraged as a performance booster via fine-tuning or augmentation with great room for improvement [5]. Synthesized data of low fidelity or low diversity is typically the combined outcome of the stochastic nature of RF data and the difficulty in training GAN models. The low-dimensional and simplistically synthesized data would have limited utility in RF sensing applications, including human activity recognition (HAR) [6], [7].

In the domain of computer vision (CV), 3D *pose animation data* of great fidelity, diversity, and coherence have been generated utilizing diffusion models [8]–[10]. Chen et al. [11] went one step further in generating vivid 3D human motion given a wide array of input prompts by performing diffusion models on the motion latent space. In this paper, we take one step further to propose utilizing diffusion models on latent domains that preserve the time-varying nature of RF sensing data, to generate hyper-realistic *RF sensing data* for HAR. In particular, we shall establish a novel lightweight AIGC framework centered on latent domains for RFID sensing, named RFID-ACCLDM (RFID-based Activity Class Conditional Latent Diffusion Model), to synthesize high-quality and high-diversity RF data at low costs, conditioned on a range of activity classes. We also construct an RFID sensing system that interrogates the RFID tags attached to test subjects' joints for HAR as a representative example of downstream tasks for our AIGC model. The conditional latent diffusion model (CLDM)-based RFID-ACCLDM system will generate massive amounts RF data for training the RFID sensing system, thereby saving the enormous work of collecting training RF data with human labor.

The main contributions of this study can be summarized as follows:

- To the best of our knowledge, this is the first work that applies CLDM to generate RF data. Our AIGC data is of higher quality than existing methods in terms of accessibility, quantity, fidelity, and diversity. The proposed AIGC model only requires a minimal quantity of real RF training data, combined with the utilization of latent representations, thus saving a substantial amount of time and computation resources on diffusion training and inference.

- We quantitatively show that the RFID-ACCLDM generated data is of high quality through metric of Frechet Inception Distance (FID) [12].
- Our RFID-ACCLDM generated data is highly effective in boosting the performance of HAR tasks without the need for mitigating the domain gaps using additional real RF data. We demonstrated this by using a representative downstream task of HAR with RFID sensing, where the DL model trained with RFID-ACCLDM generated data outperforms that trained with real RF data.

In conclusion, we address two important problems with an *AIGC for RF sensing approach*: (i) how to save the demanding cost of collecting RF data, and (ii) how to conveniently synthesize large amounts of high-quality RF data for effective training of ML models.

The remainder of this paper is structured as follows. We review related work in Section II and then describe the system design in Section III. Section IV presents our experimental study and Section V summarizes this paper.

II. Related Works

Diffusion-based AIGC applications have largely been explored in the domain of CV. Ground-breaking results on image synthesis were reported in [13]. The fidelity of these generated content and the generalizability and adaptability of diffusion models have inspired several applications utilizing diffusion in fields outside of CV. Cao et al. [14] applied diffusion models in high-frequency spaces and achieved excellent, fast MRI reconstruction performance. Moreover, diffusion models conditioned on inputs such as texts and labels have also been proven to be capable of generating more complex and variant data. A conditional Denoising Diffusion Probabilistic Model (DDPM) was used to generate coarse but complete 3D point clouds based on real-scanned partial 3D point clouds [15], while a conditional Score-based Diffusion model was used for time-series imputation tasks for healthcare and environment data [16].

With the increasingly difficult task of bettering content generation, some researchers have shifted their attention to the simpler, lower-dimensional latent space. The ingenious and natural idea that diffusion models should have even better performance on latent dimensions, stimulates some recent applications across different fields. Latent Diffusion models (LDMs) were first introduced in [17] and has enabled state-of-the-art image synthesis without excessive computations. Then, Blattmann et al. in [18] turned an image-based LDM into an unprecedented high-quality video generator, by inserting a temporal dimension based on temporal attention to the LDM. Vision-based 3D human pose estimations have had prior success using plain diffusion models [8], but the rather complicated human movements created a huge computational overhead for the diffusion model. Instead of directly performing diffusion on human movements, the authors in [11] performed diffusion on the motion latent space. As a result, novel fidelity was achieved on extensive human motion generation with greatly reduced cost. Conditional inputs such as textual descriptions were embedded to enable vivid generation with only users' input prompts.

RFID, WiFi, and FMCW radar have been extensively exploited for HAR [6]. Recently, the incorporation of DL models has helped improve RF sensing performance. However, a massive amount of training data with high quality and diversity is typically needed for the DL models to work [19]. The inherently massive and noisy RF measurements are also subject to the impact of changes in the environment, user location, orientation, and user body shape, leading to a difficult uphill battle for making DL models scalable and generalizable. One direct and effective method to address these challenges is data augmentation, and GAN-based methods have been investigated in this regard [2], [4], [20]. Amplitude-Feature Deep Convolutional GAN (AF-DCGAN) [20] was presented to mitigate the efforts involved in collecting WiFi fingerprints by synthesizing CSI amplitude feature maps. However, any alteration to the indoor environment may cause a degradation in location accuracy. Additionally, a complicated multimodal GAN [21] including two generators and one classification model was designed to synthesize CSI (channel state information) data for addressing the impacts of environmental changes. Despite their effectiveness in boosting sensing performance, most GAN-generated data exhibit a relatively large domain gap from real data, which limits their usefulness. A simple yet powerful data augmentation approach is needed for such RF sensing applications.

III. System Design

As shown in Fig. 1, the proposed RFID-ACCLDM system consists of two stages. The first stage is a recurrent variational autoencoder (R-VAE) that can accurately sample latent distributions and faithfully reconstruct the latent representations back into original RF data. The latent space of RF activity data is compact and lightweight, while capturing a significant amount of features of the raw RF domain. The second stage is a CLDM that performs the diffusion process on the latent dimensions. The trained model is able to mass-generate latent vector representations that can be decoded into realistic and diverse RF data corresponding to different human activities.

A. R-VAE

The RF data corresponding to human activities, i.e., $x_{1:N}^{L} = \{x_i^L\}_{i=1}^{N}$, are 2D time-variant data with numerous features, in which N stands for the time frame number, and L denotes the number of RF features. RF signals are readily impacted by nearby movements, and, when captured by RF devices, behave in a cyclical fluctuation pattern distinctive to different human activities. To learn the time dependencies in temporal RF data and sample latent vectors with time dependencies, we incorporate LSTM (Long Short-Term Memory) units into the VAE encoder and decoder structure, termed LSTM RF encoder ε and LSTM RF decoder ψ, respectively. The encoder ε encodes real RF data $x_{1:N}^{L}$ into a latent vector $z = \varepsilon(x_{1:N}^{L}) \in R^{1\times ld}$, whose dimension is a 1D vector with arbitrary length. The LSTM encoder is fed with the input RF sequence over

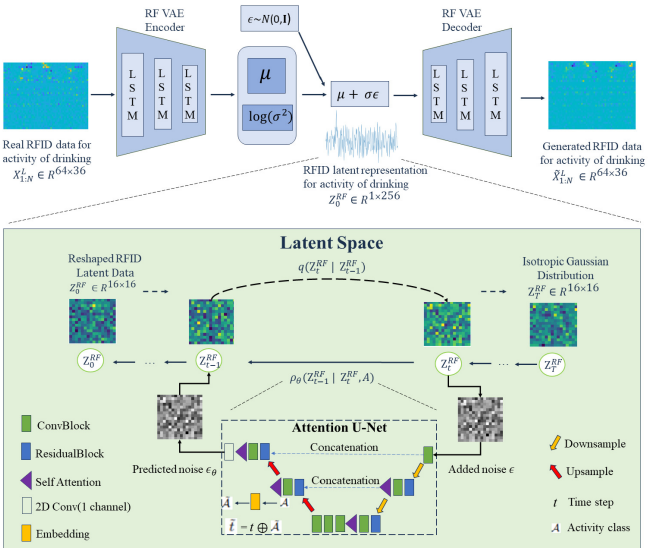

Figure 1. The procedure of conditional RF data generation with RFID-ACCLDM. The reverse process p (see (2)) progressively transforms random Gaussian noises into plausible time series data, conditioned on embedded class labels. The structure of the denoiser, the U-Net model, is also illustrated.

time, and the neural network attempts to store all of its data in its final hidden state c_t (cell state) by encapsulating it. The mean μ and log variance σ^2 can then be obtained after passing through a linear layer. The latent distribution z can be parameterized by a normal distribution with such mean and log variance. To enable back propagation for this random block computation, a reparameterization trick is executed to approximate z as $z = \mu + \tilde{\sigma} \cdot \epsilon$, where $\tilde{\sigma} = e^{0.5 \times \log \sigma^2}$ and ϵ is sampled from a standard normal distribution $\mathcal{N}(0, \mathbf{I})$ with the same shape of the standard deviation $\tilde{\sigma}$. The internal states are then passed onto the decoder ψ consisting of LSTM cells, which will be used to reconstruct the target sequence. The encoder and decoder are implemented by a 3-layer LSTM with a hidden size of 1,024. The latent length of z is set to 256.

The overall training objective is to minimize the total reconstruction error and negative Kullback-Leibler (KL) divergence score, which can be expressed as follows [22]:

$$\min_{\phi,\theta} \mathcal{L}_{R-VAE}(\phi,\theta) \tag{1}$$
$$= \mathbb{E}_{q_\phi(z|x^L_{1:N})}[\log p_\theta(\tilde{x}^L_{1:N}|z)] - KL(q_\phi(z|x^L_{1:N})||p(z)),$$

where $q_\phi(z|x^L_{1:N})$ and $p_\theta(\tilde{x}^L_{1:N}|z)]$ are parametric probability distributions modeling the encoder and decoder, respectively, with ϕ and θ being the variational parameters; $P(z)$ represents the latent distribution of $\mathcal{N}(0, \mathbf{I})$. The first term in (1) is similar to autoencoder's reconstruction loss and can be trained with mean squared error (MSE) $(x^L_{1:N} - \tilde{x}^L_{1:N})^2$. The second term can be transformed to $-0.5 \sum_{l=1}^{ld}(1 + \log(\sigma_l^2) - \mu_l^2 - \exp\log(\sigma_l^2))$. In each epoch, the total loss is calculated through $\sum_{m=1}^{M} x_m$ for M amounts of RF data with $x_m = x^L_{1:N}$ being the RF data for the mth individual activity.

B. RFID Data Generation with Conditional Latent Diffusion

Denoising diffusion probabilistic models [23] progressively perturb data with random noises (termed the "forward diffusion" process), and then remove noises in succession to generate new data samples (termed the "reverse diffusion" process). The former can be designed with a T-length Markov chain with fixed-variance scheduler to alter data distribution into an Isotropic Gaussian distribution, whereas the latter also utilizes a T-length Markov chain to reverse the Gaussian corruption by learning the transitional kernels parametrically modeled by a neural network $\epsilon_\theta(x_t, t)$ such as the U-Net [24].

Nevertheless, raw RF data typically have sophisticated motion-specific features over time coupled with high-frequency outliers, which hinder the diffusion model to learn the true data distribution. With increasing variations of activity classes, a base diffusion setup with DDPM schedules and a U-Net will have difficulties generating realistic RF data true to their class labels (i.e., human activities), while at the same time consuming more computational time and resources. Therefore, we propose to carry out the diffusion process on a representative and low-dimensional RF latent space, i.e., $z \in R^{1\times256}$, to reduce the cost and enhance the generative quality. To meet the input dimensions of the U-Net, we first reshape the latent space into a 2D representation of size $1\times16\times16$. The proposed RFID-ACCLDM system is capable of generating RFID data of high fidelity and diversity that closely aligns with various activity classes. The impressively realistic data samples vividly capture long-range correlation of movement trajectories as well as short-range delicate movement information of human joints.

In RFID-ACCLDM, the latent vector is denoted as z_t^{RF} for convenient reference at any time step within the forward and reverse diffusion process. Following the notation, $z_0^{RF} = \varepsilon(x^L_{1:N})$ is the first and pre-noising sample in the forward process, as well as the final sampled latent vector. The forward diffusion on latent space can be modeled as a Markov nosing process as follows:

$$q(z_t^{RF}|z_{t-1}^{RF}) = \mathcal{N}(z_t^{RF}; \sqrt{\alpha_t} z_{t-1}^{RF}, 1-\alpha_t \mathbf{I}).$$
$$q(z_{1:T}^{RF}|z_0^{RF}) = \prod_{t=1}^{T} q(z_t^{RF}|z_{t-1}^{RF}),$$

in which the constant $\alpha_t \in (0,1)$ is a hyper-parameter for noising and sampling, and α_t is calculated as $1-\beta t$.

Furthermore, we use $\mathcal{A}$ to designate the class label of human activities ranging from simple one-limb activities (e.g., drinking) to complex full-body activities (e.g., fortnite dancing). To enable conditional latent diffusion, we design a reverse diffusion process tailored to the latent space of RFID sensing, and a supervised training method. The class label $\mathcal{A}$ is taken as the conditioning input. The Markov chain for the reverse process of RFID-ACCLDM is defined as:

$$p_\theta(\mathbf{z}_{t-1}^{RF}|\mathbf{z}_t^{RF}, \mathcal{A}) = \mathcal{N}(\mathbf{z}_{t-1}^{RF}; \boldsymbol{\mu}_\theta(\mathbf{z}_t^{RF}, t\,|\,\mathcal{A}), \boldsymbol{\Sigma}_\theta(\mathbf{z}_t^{RF}, t\,|\,\mathcal{A}))$$
$$p_\theta(\mathbf{z}_{0:T}^{RF}|\mathcal{A}) = p(\mathbf{z}_T^{RF}) \prod_{t=1}^{T} p_\theta(\mathbf{z}_{t-1}^{RF}|\mathbf{z}_t^{RF}, \mathcal{A}).$$

Next, we define a new denoiser U-net $\epsilon_\theta(z_t^{RF}, t \,|\, \mathcal{A})$, using activity class labels as the conditional input. The parameterization of $p_\theta(z_{t-1}^{RF} \,|\, z_t^{RF})$ is is given by:

$$\mu_\theta(z_t^{RF}, t) = \frac{1}{\sqrt{\alpha_t}} \left(z_t^{RF} - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}} (\epsilon - \epsilon_\theta(z_t^{RF}, t \,|\, \mathcal{A}) \right),$$

where z_t^{RF} is defined as $\sqrt{\bar{\alpha}_t} \cdot z_0^{RF} + \sqrt{1-\bar{\alpha}_t} \cdot \epsilon_0$ with $\epsilon_0 \sim \mathcal{N}(0, \mathbf{I})$ and $\bar{\alpha}_t = \prod_{\tau=0}^{t} \alpha_\tau$. As in [23], the reverse process of our RFID-ACCLDM system can trained by solving the following optimization problem:

$$\min_\theta \; \mathcal{L}_{RFID-ACCLDM}(\theta) \tag{2}$$
$$= \mathbb{E}_{t, \epsilon \sim \mathcal{N}(0,\mathbf{I}), z_0^{RF} \sim q(z_0^{RF})} \left\| (\epsilon - \epsilon_\theta(z_t^{RF}, t \,|\, \mathcal{A})) \right\|^2.$$

The denoising function $\epsilon_\theta(z_t^{RF}, t \,|\, \mathcal{A})$ estimates the noise vector ϵ that was introduced to its noisy latent vector input z_t. During the training of the U-Net, the encoder $\boldsymbol{\varepsilon}$ can be frozen to compress motion into z_0^{RF}. If the overhead of computation is troublesome for computing devices with limited power, latent vectors of RF data from different activity classes can be computed before the diffusion at the cost of scalability and convenience for the entire system. During the reverse diffusion stage, $\epsilon_\theta(z_t^{RF}, t \,|\, \mathcal{A})$ first predicts z_0^{RF} with T successive denoising steps. Then the decoder $\boldsymbol{\psi}$ reshapes and decodes z_0^{RF} back to RF data corresponding to specific human activity.

The U-Net, deployed as the denoiser network for the diffusion process of RFID-ACCLDM, is based on a wide ResNet. We choose U-Net since it can compress and reconstruct a noisy latent input at time step t to predict the noise that has been added to the latent input, hence achieving the effect of "denoising," which is one step of the reverse diffusion process of generating new samples. The training objective in each epoch can be conveniently modeled by $(\epsilon_\theta - \epsilon)^2$, i.e., the MSE function between the predicted noise ϵ_θ and the introduced noise ϵ. To capture the time step t when the latent representation within a batch is currently computed for the U-Net, we apply sinusoidal positional encodings to encode the noise level and time step t. To incorporate activity class conditioned diffusion generation, we first embed the class labels using an MLP (multilayer perceptron) layer, which can be easily implemented through a Pytorch function. The class embedding is then integrated into the U-Net by concatenating the embedded label with time step t. We denote the resulting time step as $\tilde{t}$. The implementation of our U-Net network is shown in the lower part of Fig. 1. We use a basic architecture of *U-Net model for diffusion* including residual blocks and the self-attention mechanism. The encoder compresses our reshaped latents $z_0^{RF} \in R^{16\times16}$ to as small as $R^{4\times4}$.

IV. Experimental Study

A. Implementation and Experiment Setting

As a paradigmatic downstream task, we design a holistic RFID-based HAR system to evaluate the performance and advantages of our generative network model. As shown in Fig. 2, an off-the-shelf Impinj R420 reader, passive ALN-9634 (HIGG-3) tags, and three S9028PCR polarized antennas are used in the system. We attach 12 RFID tags to the test subject's joints, including the hip, neck, left upper leg, left knee, right upper leg, right knee, left shoulder, left arm, left forearm, right shoulder, right arm, and right forearm. A Lenovo Legion gaming laptop with an Nvidia GTX 1660 Ti GPU and an Intel Core i7-9750H CPU is used to process raw RF signals and train diffusion models.

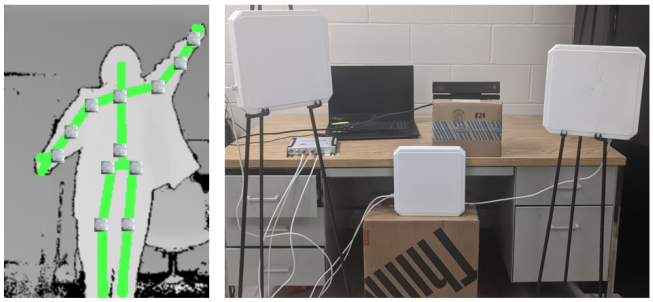

Figure 2. The setup of the RFID-based HAR experimental system.

We collect RFID phase data from three antennas of the reader An Xbox Kinect 2.0 device is used to obtain vision data, which is used as labels for supervised training in the original baseline system. The variations between RFID phase values from two successive time frames are computed as the ground truth RF data. The sampling rate of RFID phase data is around 110 Hz, while the frame rate of Kinect is 30 frames per second (fps). Every collected data sample is preprocessed and synchronized prior to being downsampled to 7.5 Hz.

In the R-VAE and diffusion training, we set the length of RFID activity data to 64 samples (or, 8.53 seconds). A window of 30 frames (4 seconds) with a sliding factor of 10 frames (1.3 seconds) is slid across 64 frames to obtain 4 RFID data units of 30 frames, which are used for the downstream task of HAR. The total dimension of RF data is 30×36 where 30 refers to the number of time frames, and 36 denotes the number of RF features.

We use six RFID data files with a length of 64 frames per activity class as the U-Net's training data. These data were captured from three test volunteers with similar body shapes. All our models are trained with the AdamW optimizer with a batch size of 4. As for the R-VAE training, the learning rate is set to 0.0001. The training task lasts 4 hours. A linearly scaled variance β_t is chosen from $\beta_0 = 10^{-4}$ to $\beta_T = 0.02$ for the diffusion training. The number of noising steps T is set to 1,000. We utilize a cyclical learning rate mechanism with the maximum learning rate set to 0.005. For run time, the U-Net is trained for 12 hours, while the diffusion training on raw RFID data takes 16 hours. For sampling, our latent diffusion technique only takes 4 seconds to generate one sample, while the diffusion model on raw data takes nearly 40 seconds. Classifier-free guidance [25] is implemented to improve data generation and prevent the model from synthesizing images of conflicting classes.

Table I
COMPARISON OF FID SCORES: RFID-ACCLDM VS. RFPOSE-GAN

	Standing still	Waving	Walking
RFID-ACCLDM (proposed)	4.5583	7.0073	3.6421
RFPose-GAN [4]	36.1981	32.2464	45.3412

B. Quality of Synthesized RF Data

A distinctive advantage of the proposed RFID-ACCLDM model is that it produces high-quality RFID data with *great diversity*, as opposed to merely producing data that is homogeneous and similar to the training set. Such diversity is highly desirable for training robust DL models. In this study, we employ the Frechet Inception Distance (FID) [12] to evaluate the *distribution similarity* between collections of generated and real RFID data. The FID score quantifies the distance between feature vectors in a high-dimensional latent space. A lower FID score indicates that the generated RFID data is more faithful to the real data (higher fidelity).

We randomly sample 80 latent diffusion generated and real RFID data from different activities for FID calculation and comparison. The activities range from a simple activity of standing still to a complicated activity of boxing that involves all the body parts. As can be seen in Table I, superior FID scores are achieved by our proposed model over our previous work RFPose-GAN [4]. RFPose-GAN uses a supervised GAN to map a specific 3D pose data to its corresponding synthesized RFID data. It may be hard to train such GAN models, where only some parts of the data distributions were learned sometimes. Consequently, it is rather challenging to synthesize specific activities with minimal variations over time under noise and interference from the environment, which result in the high FID scores of RFPose-GAN. On the other hand, the significantly lower FID scores of RFID-ACCLDM demonstrate the high fidelity of its generated RFID data. Such a caliber of FID scores is on a par of state-of-the-art image synthesis works [17].

C. Human Activity Recognition Results

As a final test of RFID-ACCLDM, we use its synthesized RFID data to train a downstream task's DL model. In this study, the quality of our generated data is tested using an RFID-based HAR system with six activity classes. We deploy a straightforward CNN model for the classification task, which consists of four 2D convolutional layers each accompanied by a dropout layer to help reduce overfitting. The second, third, and fourth convolution layer is followed by a maxpooling2D layer. For the purpose of calculating final accuracy, the convolution output is flattened and fed into a fully connected layer. Given that the The test data are from the collected ground truth data including two different subjects at locations slightly different from where the training data was collected. They are also processed with time windows starting and ending at random time frames to try to replicate a real-life scenario.

In Fig. 3, three confusion matrices for RFID-based HAR are presented. They are obtained by training on 32 minutes of real data (left), 16 minutes of RFID-ACCLDM generated data (middle), and 64 minutes of RFID-ACCLDM generated data (right), respectively. Despite using synthesized data that is only half the amount of real data, the accuracy and F1 score are slightly better than training with real data. This is because our synthesized data offers more fine-grained diversity while reaching the same level of fidelity as real data. Furthermore, with the addition of another 48 minutes of RFID-ACCLDM synthesized data, both the accuracy and F1 score completely outperform the case of training with real data by a large margin (reaching 91.80% and 91.56%, about a 9.7% improvement). This result proves the superiority of our AIGC model because it only takes us about 36 minutes to create this amount of synthesized data. It is important to note that the CNN designed for the classification task is only to showcase the effects of our generated data, but not to bring out the full potential of such data and real data. Future work will involve a more comprehensive system of classifiers for an ablation study.

Fig. 4 shows a comprehensive comparison of F1 scores obtained through our proposed model by progressively synthesizing larger amounts of data at different training epochs. It can be seen that the F1 score is steadily improved as more synthesized data are used in model training. The F1 curve is able to surpass the model trained on 32 minutes of real data when only 16 minutes of generated data are used after 480 epochs of pursuit. With 64 minutes of synthesized samples, the F1 curve becomes higher than that of training on 32 minutes of real data for the entire training process. The models trained on 128 minutes of synthesized data converge to a high-performance state after merely 40 epochs, and its F1 curve reaches a new height of 93.05%. This demonstrates the greatly reduced domain gap between real and generated data, which is very common in the case of GAN generated data.

It is important to highlight that the superior F1 scores are obtained by only using synthesized data: *this is an AIGC method, rather than a data augmentation method*. This experiment proves that the data generated by the proposed RFID-ACCLDM method are can replace real data for CNN-based HAR. The fidelity and diversity of the AIGC RFID data synthesized by our model are validated.

V. CONCLUSIONS

In this paper, we proposed an *AIGC for RF sensing* approach to address the challenge of lacking RF data. The proposed RFID-ACCLDM framework utilizes a latent diffusion model conditioned on activity class labels to generate RFID sensing data. We demonstrated the high quality and usefulness of the synthesized data by the proposed RFID-ACCLDM system through the metric of FID, followed by a representative downstream task. The proposed AIGC for RF sensing approach offered a convincing solution to the pressing issues of how to obtain high-quality RF data and minimize the high expense of RF data acquisition.

Acknowledgment: This work is supported in part by the NSF under Grants CNS-2107190, CNS-2319342, and IIS-2306789.

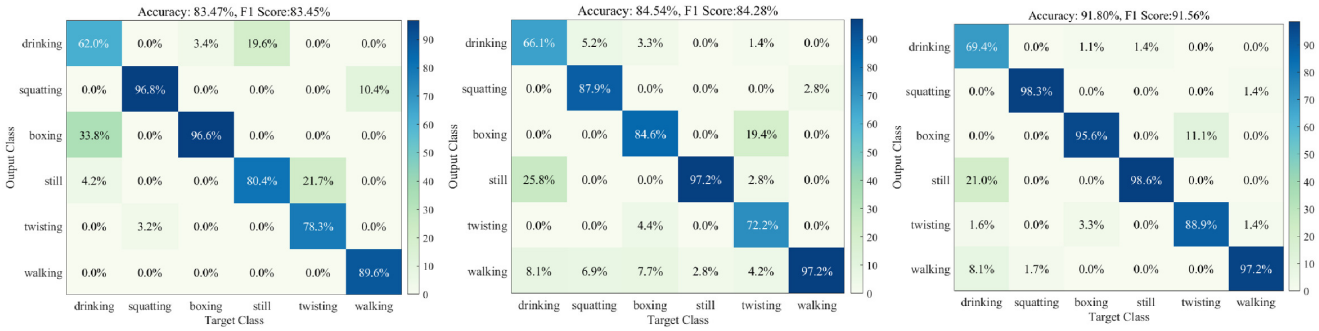

Figure 3. The confusion matrices obtained with CNN models trained on 32 minutes of real data (left), 16 minutes of RFID-ACCLDM generated data (middle), and 64 minutes of RFID-ACCLDM generated data (right).

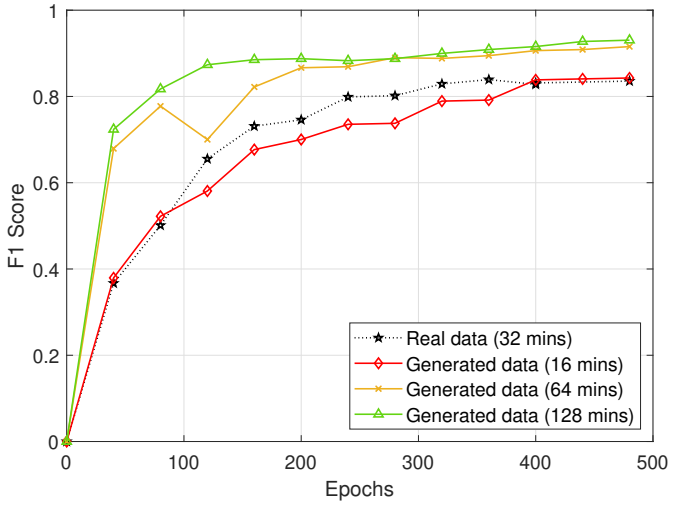

Figure 4. F1 scores of human activity classification when the quantity of RFID-ACCLDM generated data is progressively increased.

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 Communications Surveys and Tutorials*, vol. 21, no. 4, pp. 3072–3108, Fourth Quarter 2019.

[2] 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.

[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] 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.

[5] J. Wang, L. Zhang, C. Wang, X. Ma, Q. Gao, and B. Lin, "Device-free human gesture recognition with generative adversarial networks," *IEEE Internet of Things Journal*, vol. 7, no. 8, pp. 7678–7688, Aug. 2020.

[6] 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.

[7] ——, "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.

[8] 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.

[9] W. Shan, Z. Liu, X. Zhang, Z. Wang, K. Han, S. Wang, S. Ma, and W. Gao, "Diffusion-based 3D human pose estimation with multi-hypothesis aggregation," *arXiv preprint arXiv:2303.11579*, Aug. 2023. [Online]. Available: https://arxiv.org/abs/2303.11579

[10] C. Rommel, E. Valle, M. Chen, S. Khalfaoui, R. Marlet, M. Cord, and P. Perez, "DiffHPE: Robust, coherent 3D human pose lifting with diffusion," in *Proc. IEEE/CVF ICCV Workshops*, Paris, France, Oct. 2023, pp. 3220–3229.

[11] X. Chen, B. Jiang, W. Liu, Z. Huang, B. Fu, T. Chen, and G. Yu, "Executing your commands via motion diffusion in latent space," in *Proc. IEEE/CVF CVPR 2023*, Vancouver, Canada, June 2023, pp. 18 000–18 010.

[12] 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.

[13] P. Dhariwal and A. Nichol, "Diffusion models beat GANs on image synthesis," in *Proc. NeurIPS 2021*, Virtual Conference, Dec. 2021, pp. 8780–8794.

[14] C. Cao, Z.-X. Cui, S. Liu, H. Zheng, D. Liang, and Y. Zhu, "High-frequency space diffusion models for accelerated MRI," *arXiv preprint arXiv:2208.05481*, Dec. 2022. [Online]. Available: https://arxiv.org/abs/2208.05481

[15] Z. Lyu, Z. Kong, X. XU, L. Pan, and D. Lin, "A conditional point diffusion-refinement paradigm for 3D point cloud completion," in *Proc. ICLR 2022*, Virtual Conference, Apr. 2022, pp. 1–24.

[16] Y. Tashiro, J. Song, Y. Song, and S. Ermon, "CSDI: Conditional score-based diffusion models for probabilistic time series imputation," in *Proc. NeurIPS 2021*, Virtual Conference, Dec. 2021, pp. 1–13.

[17] 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.

[18] A. Blattmann, R. Rombach, H. Ling, T. Dockhorn, S. W. Kim, S. Fidler, and K. Kreis, "Align your latents: High-resolution video synthesis with latent diffusion models," in *Proc. IEEE/CVF CVPR 2023*, Vancouver, Canada, June 2023, pp. 22 563–22 575.

[19] 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.

[20] Q. Li, H. Qu, Z. Liu, N. Zhou, W. Sun, S. Sigg, and J. Li, "AF-DCGAN: Amplitude feature deep convolutional GAN for fingerprint construction in indoor localization systems," *IEEE Trans. Emerg. Topics Comput. Intell.*, vol. 5, no. 3, pp. 468–480, June 2021.

[21] 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.

[22] D. P. Kingma and M. Welling, *An Introduction to Variational Autoencoders*. Boston, MA: Now Publishers, 2019.

[23] 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

[24] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Medical Image Comput. Computer-Assisted Intervention 2015*, Munich, Germany, Oct. 2015, pp. 234–241.

[25] J. Ho and T. Salimans, "Classifier-free diffusion guidance," in *Proc. NeurIPS 2021 Workshops*, Virtual Conference, Dec. 2021, pp. 1–8.