Synthetic Signals: Augmenting WiFi-based Human Activity Recognition

Raed Abdulrahim Abumostafa
Department of Information and Computer Science
King Fahd University of Petroleum and Minerals
Dhahran, Saudi Arabia
202443380@kfupm.edu.sa

Ahemd Fatani
College of Computer Science and Engineering
University of Jeddah
Jeddah, Saudi Arabia
AIahemdFatani33@gmail.com

Abdullah Alnaqbi
Department of Mechanical Engineering
Higher Colleges of Technology – Fujariah
Fujariah, Unitied Arab Emirates
H00461581@hct.ac,ae

Supervised by: Mubashar Muhammed

Abdulmajeed Alqarni
College of Computer and Information Sciences
Prince Sultan University
Riyadh, Saudi Arabia

***Abstract*—This project explores the application of the RF-Diffusion model to human activity recognition (HAR) datasets, with the goal of improving the performance of downstream models such as Vision Transformers (ViT) and self-supervised learning (SSL) methods. By leveraging RF-Diffusion for generating and augmenting data, we investigated whether the model could enhance the accuracy of HAR classification tasks. Experiments were conducted on standard HAR datasets, and performance metrics were compared between models trained with and without RF-Diffusion augmentation. The results indicate that RF-Diffusion can provide meaningful improvements in downstream model accuracy, suggesting its potential as a tool for enhancing HAR model performance. These findings highlight the utility of diffusion-based generative models in signal processing applications and open avenues for further research in synthetic data augmentation for HAR systems.**

Keywords—RF-Diffusion, Human Activity Recognition (HAR), Vision Transformer (ViT), Self-Supervised Learning (SSL), Data Augmentation, Downstream Model Accuracy, Generative Models, Synthetic Data

I. INTRODUCTION

Human Activity Recognition (HAR) is a multidisciplinary field focused on automatically identifying and interpreting human actions using data collected from various sensors and devices. HAR systems have been widely explored using wearable sensors, such as accelerometers and gyroscopes, which provide precise motion information but require users to carry or wear devices consistently. Vision-based methods use cameras to capture human movements, offering high accuracy but raising privacy concerns and being sensitive to lighting or occlusion. Contact-based sensors, such as pressure mats or instrumented floors, provide localized activity information but are limited to specific environments.

Recently, wireless signal-based HAR, particularly using Wi-Fi, has emerged as a practical and effective middle ground. Wi-Fi-based methods leverage the interaction between radio-frequency (RF) signals and human bodies to infer activities without requiring cameras or wearable devices. This approach preserves privacy, covers larger areas, and allows non-intrusive monitoring, making it a popular choice for real-world HAR applications. The field has evolved from early methods relying on handcrafted features and classical machine learning to deep learning models, which automatically extract complex representations from raw sensor data. CNNs were among the first deep learning architectures applied to HAR, followed by Vision Transformers (ViT) and self-supervised learning (SSL) methods. These state-of-the-art models can learn rich features and capture long-range dependencies, significantly improving activity recognition performance.

Despite these advances, HAR still faces challenges such as limited labeled data, imbalanced activity classes, and difficulty generalizing across environments or subjects. Generative models, including diffusion-based approaches like RF-Diffusion, offer a potential solution by creating realistic synthetic data to augment existing datasets. This project aims to investigate the effectiveness of RF-Diffusion in enhancing HAR performance, focusing on its impact on downstream models such as ViT and SSL methods. By evaluating whether synthetic RF data can improve accuracy, robustness, and generalization, this study seeks to demonstrate the potential of generative augmentation for advancing state-of-the-art HAR systems.

XXX-X-XXXX-XXXX-X/XX/$XX.00 ©20XX IEEE

II. METHODS AND DATASETS

Our approach builds upon a diffusion-based generative model originally introduced in the RF-Diffusion paper, which we have heavily modified and adapted for Wi-Fi channel state information (CSI) signals. The architecture follows the standard diffusion process, where random noise is progressively denoised into structured signal samples that resemble real Wi-Fi measurements. These synthetic signals are then used to augment existing datasets, with the goal of improving the performance of downstream models such as Vision Transformers (ViT) and self-supervised learning (SSL) methods. For evaluation, we focus on two benchmark Wi-Fi-based HAR datasets: NTU Wi-Fi HAR (NTU-RF), a large-scale dataset that captures diverse human activities across multiple scenarios, and UT-HAR, a smaller dataset collected under more controlled conditions. Together, these datasets provide a strong basis for testing whether diffusion-based augmentation can enhance HAR accuracy and generalization.

A. Data Collection

Wi-Fi-based HAR relies on channel state information (CSI), which records how wireless signals are altered by human movement in the environment. When a person performs an activity, their body disrupts the signals, creating distinctive patterns in the CSI. These measurements are captured using Wi-Fi transceivers and later processed for activity recognition.

For this project, we use two benchmark Wi-Fi datasets. The first, UT-HAR, provides CSI samples of size $1 \times 250 \times 90$ across 7 activity classes: lie down, fall, walk, pickup, run, sit down, and stand up. It includes 3,977 training samples and 996 testing samples. The second, NTU-HAR, contains CSI samples of size $3 \times 114 \times 500$ covering 6 classes: box, circle, clean, fall, run, and walk. Together, these datasets provide a strong basis for evaluating our diffusion-based augmentation approach under both controlled and large-scale scenarios

B. Data Preprocessing

Before training, the CSI datasets required careful preprocessing to ensure compatibility with both the conditional diffusion backbone and the ViT classifier. Since our generative model is conditional, a dedicated condition file was prepared to provide activity labels and other metadata required during the diffusion process. This ensured that generated CSI samples were aligned with the corresponding activity classes.

The two available datasets differed in format, dimensionality, and activity classes, which made integration challenging. Initially, we attempted to resize and reformat the datasets into a unified structure; however, after many trials, we decided to keep the datasets separate. Each dataset contained varying activity classes, and merging them introduced inconsistencies that complicated the conditional diffusion process. By keeping them separate, we preserved dataset integrity and avoided label conflicts that could degrade model performance.

We also faced challenges in linking and loading the datasets due to differences in file organization and metadata structures. Several iterations of cleaning and restructuring were required to ensure compatibility with the training pipeline. Despite these issues, the preprocessing steps successfully produced datasets that could be reliably used by both the diffusion model and the ViT classifier.

C. Model Training

We conducted training both locally and on Google Colab GPUs to balance flexibility and computational efficiency. The diffusion model was trained for 200 epochs with a reduced batch size due to memory constraints. This setup allowed us to generate synthetic CSI signals that were later used to enhance the ViT classifier.

The training was performed using the following hyperparameters:

General parameters:

- max_iter = None (unlimited iterations)
- batch_size = 32
- learning_rate = 1e-3
- max_grad_norm = None
- Inference parameters
- inference_batch_size = 1
- robust_sampling = True

Data parameters:

- sample_rate = 2000 (model expects long signals)
- input_dim = 2000
- extra_dim = [2000]
- cond_dim = 6

Model parameters:

- embed_dim = 256
- hidden_dim = 128
- num_heads = 8
- num_block = 32
- dropout = 0.0
- mlp_ratio = 4
- learn_tfdiff = False

Diffusion parameters:

- signal_diffusion = True
- max_step = 100

This configuration provided a stable training process while ensuring the model could capture both local variations and global patterns in the CSI signals.

.

D. Model Architecture

Our model integrates a diffusion-based generative backbone with a Vision Transformer (ViT) for Human

Activity Recognition (HAR) using Wi-Fi CSI signals. The generative backbone is based on RF-Diffusion, which was originally designed to synthesize RF signals through a hierarchical diffusion transformer. In RF-Diffusion, signals are represented in a time-frequency domain, allowing the model to capture both fine-grained temporal dynamics and global frequency patterns. It also handles complex-valued signals (amplitude and phase) to ensure physically realistic outputs. The model generates synthetic CSI by progressively denoising random noise into structured signal samples, preserving both local variations and global signal structure. This augmented data improves the training of downstream classifiers.

The ViT classifier processes the real and synthetic CSI data. CSI matrices are divided into non-overlapping patches, flattened, and projected into embeddings. A [CLS] token and positional embeddings are added before feeding the data into a transformer encoder. The encoder uses multi-head self-attention and feed-forward layers with residual connections to capture long-range dependencies across patches. Outputs are mean-pooled and passed through a linear layer to predict activity classes.

The two main components of our model build directly on prior work. The diffusion-based generative backbone is adapted from the RF-Diffusion framework, which demonstrated the effectiveness of hierarchical diffusion transformers for synthesizing realistic RF signals. By leveraging its capability to model the time-frequency structure and complex-valued nature of RF data, we ensure that our synthetic CSI samples maintain high fidelity and physical realism.

Similarly, the Vision Transformer (ViT) classifier in our pipeline draws inspiration from SensFi, a recent approach that successfully applied transformer-based architectures to Wi-Fi sensing tasks. SensFi showed that dividing CSI into patch embeddings and applying self-attention can effectively capture long-range dependencies in signal patterns. The Vision Transformer (ViT) was configured with patch sizes of 9 × 25, resulting in an embedding dimension of 225 for each patch. A single input channel was used, corresponding to the CSI matrix, which was reshaped to a fixed size of 342 × 500 before patching. The transformer encoder operated with a depth of 1 layer, embedding dimension of 225, and 5 attention heads, with a dropout probability of 0.5 applied within the encoder blocks. The feed-forward network expanded embeddings by a factor of 4 before projecting them back to the original dimension. For classification, the model employed mean pooling across token embeddings, followed by a LayerNorm and a linear projection to the target number of activity classes.
By adopting this design, our model benefits from a proven classification backbone while combining it with synthetic data generation for improved robustness.

III. Results

We evaluated the RF-Diffusion and ViT models, as proposed in the original studies, on the small CSI datasets available to us. The goal was to assess how well these pre-trained models generalize to datasets with limited size and varying activity classes.

For the diffusion-generated signals, the SSIM (Structural Similarity Index) ranged between 80 and 83 for most samples, with some high-quality examples reaching 95, indicating that the generated CSI retained strong structural similarity to real signals. The FID (Fréchet Inception Distance) was measured at 12, reflecting a relatively close alignment between the distribution of generated and real samples.

However, the limited size of our datasets posed a challenge. With fewer samples and reduced variability, the model's outputs may not fully reflect the diversity seen in larger datasets, which could affect the reliability of SSIM and FID scores. Despite this limitation, the results demonstrate that the RF-Diffusion model and ViT classifier are able to process small CSI datasets and generate structurally meaningful signals that can support downstream human activity recognition tasks.

The Vision Transformer (ViT) classifier was evaluated on the available CSI datasets to assess the impact of synthetic data generated by the RF-Diffusion model. When trained only on the original, non-augmented datasets, the ViT achieved an accuracy of **86%** on NTU dataset samples, reflecting baseline performance with limited real data. After augmenting the training data with synthetic CSI samples produced by the diffusion model, the classifier's performance improved on the NTU dataset, reaching **95% accuracy**. This demonstrates that the diffusion-generated signals effectively enhanced the model's ability to generalize and recognize activity patterns. On the UT dataset, however, the improvement was less pronounced, with accuracy remaining lower, highlighting the challenges posed by smaller datasets with varying activity classes. These results indicate that synthetic augmentation can significantly benefit model performance, especially when the original dataset has sufficient coverage of activity classes, but its effectiveness may vary depending on dataset size and diversity.

IV. Conclusion

In this study, we explored the application of RF-Diffusion and Vision Transformer (ViT) models for Human Activity Recognition using Wi-Fi CSI signals. The RF-Diffusion model demonstrated its ability to generate realistic synthetic CSI samples that preserve both temporal and frequency patterns, as reflected in SSIM scores ranging from 80 to 83 (with some samples reaching 95) and an FID score of 12. The synthetic data proved valuable for augmenting small CSI datasets, helping the ViT classifier improve its performance on the NTU dataset from 86% accuracy without augmentation to 95% with diffusion-augmented data.

It is worth noting that both the RF-Diffusion and ViT models are computationally heavy, requiring significant resources for training and inference, which can limit their deployment in resource-constrained environments. Additionally, limitations in dataset size and diversity affected model generalization, particularly on the UT dataset, and may have influenced the reliability of SSIM and FID metrics.

Despite these constraints, the results indicate that combining generative diffusion models with transformer-based classifiers can enhance human activity recognition, especially in scenarios where real data is scarce. Future work could focus on optimizing these architectures for efficiency, incorporating larger and more diverse datasets, and exploring advanced augmentation and fusion strategies to further improve performance and robustness.

REFERENCES

[1] *G. Chi, Z. Yang, C. Wu, J. Xu, Y. Gao, Y. Liu, and T. X. Han, "RF-Diffusion: Radio Signal Generation via Time-Frequency Diffusion," 2024*

[2] *F. Luo, S. Khan, B. Jiang, and K. Wu, "Vision Transformers for Human Activity Recognition Using WiFi Channel State Information," 2024.*

[3] *D. Wang, J. Yang, W. Cui, L. Xie, and S. Sun, "Multimodal CSI-Based Human Activity Recognition Using GANs," 2022.*

[4] *A. Y. Radwan, M. Yildirim, N. Hasanzadeh, H. Tabassum, and S. Valaee, "A Tutorial-cum-Survey on Self-Supervised Learning for Wi-Fi Sensing: Trends, Challenges, and Outlook," 2025.*

[5] *D. Varga, "Exposing Data Leakage in Wi-Fi CSI-Based Human Action Recognition: A Critical Analysis," 2023.*