TFSemantic: A Time–Frequency Semantic GAN Framework for Imbalanced Classification Using Radio Signals

PENG LIAO, Xidian University, China
XUYU WANG, Florida International University, United States
LINGLING AN, Xidian University, China
SHIWEN MAO, Auburn University, United States
TIANYA ZHAO, Florida International University, United States
CHAO YANG, Xidian University, China

Recently, wireless sensing techniques have been widely used for Internet of Things (IoT) applications. Unlike traditional device-based sensing, wireless sensing is contactless, pervasive, low cost, and non-invasive, making it highly suitable for relevant IoT applications. However, most existing methods are highly dependent on high-quality datasets, and the minority class will not achieve a satisfactory performance when suffering from a class imbalance problem. In this article, we propose a time–frequency semantic generative adversarial network framework (i.e., TFSemantic) to address the imbalanced classification problem in human activity recognition using radio frequency (RF) signals. Specifically, the TFSemantic framework can learn semantic features from the minority classes and then generate high-quality signals to restore data balance. It includes a data pre-processing module, a semantic extraction module, a semantic distribution module, and a data augmenter module. In the data pre-processing module, we process four different RF datasets (i.e., WiFi, RFID, UWB, and mmWave). We also develop Fourier semantic feature convolution and attention semantic feature embedding methods for the semantic extraction module. A discrete wavelet transform is utilized for reconstructed RF samples in the semantic distribution module. In data augmenter module, we design an associated loss function to achieve effective adversarial training. Finally, we validate the effectiveness of the proposed TFSemantic framework using different RF datasets, which outperforms several state-of-the-art methods.

CCS Concepts: • **Human-centered computing** → *Ubiquitous and mobile computing systems and tools;*

Additional Key Words and Phrases: Human activity recognition (HAR), imbalanced classification, wireless sensing, generative adversarial network

ACM Reference format:
Peng Liao, Xuyu Wang, Lingling An, Shiwen Mao, Tianya Zhao, and Chao Yang. 2024. TFSemantic: A Time–Frequency Semantic GAN Framework for Imbalanced Classification Using Radio Signals. *ACM Trans. Sensor Netw.* 20, 4, Article 79 (May 2024), 22 pages.
https://doi.org/10.1145/3614096

Authors' addresses: P. Liao, Guangzhou Institute of Technology, Xidian University, Guangzhou, China; email: Pengl3@stu.xidian.edu.cn; X. Wang (corresponding author) and T. Zhao, Knight Foundation School of Computing and Information Sciences, Florida International University, Miami, Florida, United States; emails: {xuywang, tzhao010}@fiu.edu; L. An (corresponding author), School of Computer Science and Technology, Xidian University, Xi'an, China; email: an.lingling@gmail.com; S. Mao, Department of Electrical and Computer Engineering, Auburn University, Auburn, United States; email: smao@ieee.org. C. Yang, Hangzhou Institute of Technology, Xidian University, Hangzhou, China; email: yangchao@xidian.edu.cn.

1550-4859/2024/05-ART79 $15.00
https://doi.org/10.1145/3614096

1 INTRODUCTION

Human activity recognition (HAR) has received increasing attention from both academia and industry. HAR aims to accurately sense different activities (e.g., gestures, movements, and poses) in the physical space, which can be used for many **Internet of Things (IoT)** applications, such as health monitoring, human–computer interaction, augmented reality, and virtual reality [4, 38]. Generally, there are two types of HAR, i.e., device based and device free. Device-based HAR utilizes wearable devices, such as smartphones or watches, to recognize human gestures or activities [19]. However, device-free HAR does not require wearable devices, where **radio frequency (RF)** signals are used for contactless sensing [1, 24, 26, 27, 39, 50]. Device-free HAR is less intrusive and more suitable for monitoring elders' or babies' activity.

Recently, deep learning-based methods have been proposed for RF-based HAR to achieve high classification accuracy [52]. Various RF technologies, such as millimeter wave (mmWave) radar, **ultra-wideband (UWB)** radar, WiFi, and RFID have been utilized in prior works. For example, Pantomime exploits the spatio-temporal properties of mmWave signals with a deep learning model for contactless gesture recognition [33]. WiFi **channel state information (CSI)**, such as CSI amplitude or phase, have also been used in several HAR systems, For example, a dense **long short-term memory (LSTM)** is to handle WiFi CSI amplitude data for HAR [48]. RFID-based techniques usually leverage the physical layer information (e.g., phase) for, e.g., activity recognition [40] and three-dimensional (3D) pose tracking [44]. Although RF-based HAR with deep learning can achieve pretty good classification performance, it still faces two major challenges: (i) generalization to different RF environments and (ii) high cost of data collection. To address these challenges, a few-shot HAR method is proposed to use deep learning for feature extraction and classification, while model parameters are transferred [42]. Meta-learning has also be used for one-shot RF-based HAR, which can adapt to different environments with minimum efforts [8].

Unlike the above model generalization methods (e.g., few-shot learning and meta-learning), our work is focused on a new research problem of RF-based HAR with imbalanced data. Various wireless devices (e.g., mmWave, WiFi, and RFID) have been used for HAR, which usually generate imbalanced datasets in different domains (e.g., different deployment environments), including considerable environment noise, interference, and missing values. The presence of a "long tail" effect in RF sensing has been identified due to sample selection bias, which can be observed when data are sorted by the frequency of occurrence of different categories from highest to lowest. Figure 1 shows the long-tailed label distribution of four public RF-based HAR datasets [25, 29, 47, 53] with a total of 55 gesture classes. We find that some common gesture classes (e.g., push, pull, slide left, and slide right) have a large number of samples in each dataset, while some designed gestures (e.g., sign language) only occupy much fewer samples. For such imbalanced datasets, the training process naturally favors larger classes (i.e., with more samples) and has lower accuracy for smaller classes (i.e., with fewer samples). Although the class imbalanced problem has been extensively investigated and tackled in computer vision and natural language processing, wireless sensing data includes several unique characteristics (e.g., complex-valued and high-dimensional data), which makes the imbalanced RF data greatly impact the performance of HAR classification tasks.

In this article, we develop a time–frequency semantic **generative adversarial network (GAN)** framework (i.e., **time.frequency semantic (TFSemantic)**) to generate high-quality datasets, aiming to address the imbalance classification problem, as well as enhancing the robustness and environmental adaptability of HAR models. Specifically, we couple GANs with an autoencoder model in TFSemantic to restore the balance of the dataset by generating RF signals in minority classes. The TFSemantic framework includes a data pre-processing module, a semantic extraction module, a semantic distribution module, and a data augmenter module. In the data pre-processing module, we process four different RF datasets, including WiFi, RFID, UWB, and mmWave. In addition,

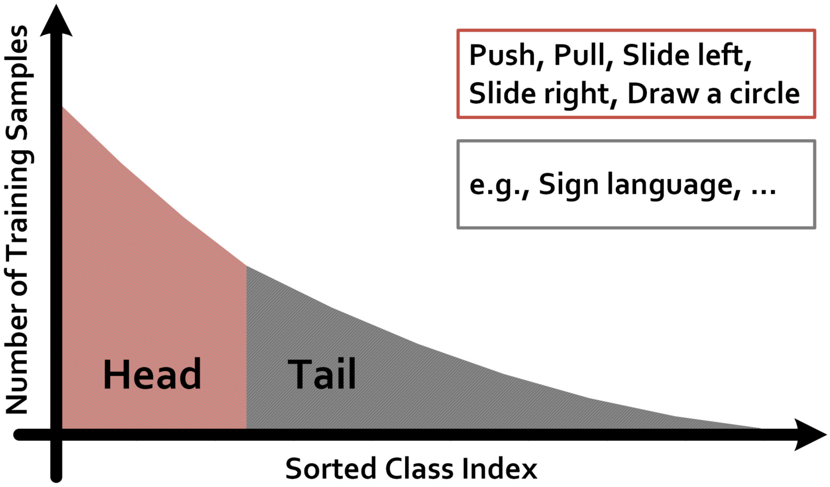


Fig. 1. Label distribution from four public RF-Based HAR datasets [25, 29, 47, 53] with totally 55 gesture classes.

the Fourier **semantic feature convolution (SFC)** and attention **semantic feature embedding (SFE)** approaches are developed in the semantic extraction module. We also propose a **discrete wavelet transform (DWT)** scheme to handle the reconstructed RF samples in the semantic distribution module. Last, we design an associated loss function in the dataset augmenter module to achieve effective training of the GAN model.

To evaluate TFSemantic, we select four public datasets collected by different RF devices under different sensing tasks. We also compare our approach with different baselines, including the state-of-the-art GAN-based augmentation, imbalance-learning, and learning-based sensing. Experimental results show that our enhanced dataset can greatly improve the classification accuracy on imbalanced datasets by about 13% on average. More important, the recognition accuracy of some classes with less sample data (i.e., the minority classes) can be improved by about 24%.

The key contributions of this article are as follows:

- We propose a time–frequency semantic GAN framework, TFSemantic, which, to the best of our knowledge, is the first work that addresses the imbalanced classification problem in RF-based HAR tasks.
- In the TFSemantic framework, we develop three new learning modules, including the semantic extraction module, semantic distribution module, and data augmenter module to augment high-quality RF dataset, which effectively restore the balance of the dataset.
- We evaluate TFSemantic with extensive experiments with several RF datasets in real-world scenarios, including different RF sensor types and experimental settings. The results validate that the proposed method can greatly improve the quality, diversity, and robustness of the RF dataset and lead to superior HAR performance.

The rest of the article is organized as follows. Section 2 discusses related work, and Section 3 introduces the background. Section 4 describes the TFSemantic design. Section 5 presents our experimental study. Discussion is in Section 6, and Section 7 concludes this article.

2 RELATED WORK

This article is focused on RF-based sensing. We review deep learning-based RF sensing with Radar, WiFi CSI, and RFID-based methods, as well as few-shot learning for RF-based HAR in the following.

2.1 Deep Learning–based RF Sensing

RF-based sensing is a technique that utilizes radio frequency signals for detecting the presence of objects, measuring the distance to objects, identifying materials, and detecting changes in temperature, humidity, or other environmental conditions. Besides, deep learning methods have been widely used in RF sensing–based applications.

2.1.1 Radar.

Radar is widely leveraged in wireless sensing because of the ability to detect a target's position, distance, speed, direction, and other relevant information. Specifically, deep learning-based radar technologies are widely used for HAR to improve activity classification accuracy [52]. For example, Pantomime is a novel mid-air gesture recognition system that utilizes the spatio-temporal properties of mmWave signals with a learning model [33]. RadHAR leverages a neural network with a sliding time window to estimate point clouds in an mmWave radar for accurately detecting different human activities [35]. In addition, commodity Radar systems have been used for vital sign monitoring to mitigate environment interference [49]. To avoid the drawbacks of traditional discrete Fourier transform preprocessing, Reference [51] proposes a learnable preprocessing module, CubeLearn, to extract features directly from the raw radar signal and construct end-to-end deep neural networks for radar-based HAR.

2.1.2 WiFi.

WiFi signals have been exploited to recognize various human activities or gestures from commodity WiFi devices. For example, GaitFi [7] proposes a novel multimodal gait recognition method that leverages WiFi signals and videos for human identification, where WiFi CSI is collected to capture human gaits. More important, the amplitude and phase difference of WiFi CSI are utilized in IoT applications. For example, dense-LSTM and data augmentation methods are used to exploit WiFi CSI amplitude for HAR [48]. In addition to deep learning, model-based approaches have been proposed for robust sensing performance, e.g., to address the position-dependent problem [11].

2.1.3 RFID.

Currently, off-the-shelf RFID devices (e.g., the Impinj R420 reader) have been used for RF sensing. For example, TACT [40] recognizes human activities using commodity RFID but without needing to attach any RFID tags to the subject. A recent work RFID-Pose estimates 3D human poses from RFID phases data using a Deep Kinematic Neural Network trained with vision data [44]. VED [49] demonstrates a promising capability in recovering fine-grained heartbeat waveform from RF-sensing signal by exploiting the universal approximation ability of deep neural networks and the generative potential of variational inference.

2.2 Few-shot Learning for RF-based HAR

The evolution of RF-based HAR includes two distinct phases. In the first phase, the primary objective is to collect RF data and then extract important information for target activity recognition by using a learning model. Subsequently, the second phase focuses on effectively adapting the model to a new environment for enhancing the overall performance of the sensing task.

By exploiting deep learning [5, 52], RF-based sensing has made great progress. Specifically, few-shot learning is to learn a model with a small number of labeled data by transferring knowledge from relevant tasks, which has been widely used in the second phase of RF-based HAR development. For example, RF-Net [8] and Meta-Pose [43] leverage meta-learning to adapt to different environments with minimum effort. OneFi [42] recognizes unseen gestures with only one (or few) labeled samples using transformer networks. Moreover, TOSS [54] uses a domain adaptation method for WiFi-based HAR with labeled and unlabeled target samples. FreeAuth [17], WiHF [21], and TARF [45] leverage adversarial learning to improve the robustness of HAR applications in different environments and with different RF technologies.

Summary. Unlike the related work, we propose new data enhancement techniques to address *the data imbalanced classification problem* by generating high-quality RF datasets. The robustness and environmental adaptability of RF sensing models can also be greatly enhanced by using the augmented data.

3 BACKGROUND AND MOTIVATION

In this section, we first introduce the class imbalance problem and then present the motivation study for the GAN-based data augmentation methods in this article.

3.1 Class Imbalance Problem

In a multi-class classification problem, class imbalance occurs when one or more classes (i.e., the minority class) contain(s) fewer samples than other classes. The class with the most samples is termed the majority class. When an imbalanced dataset is used, the training process naturally favors the majority class, resulting in poor accuracy for the minority classes. To address this problem in **computer vision (CV)**, a common strategy is to resample the image dataset (e.g., by oversampling in the minority group) [22]. The other approach is to use cost-sensitive learning (i.e., by adjusting the classifier to decouple the learning of the recognition model from that of the classification module) [55].

Generally, class imbalance in RF signal-based datasets is more complex than that in CV data. First, despite the fact that rare classes in RF sensing tasks are executed less often, each execution tends to be misclassified. Second, while additional collection is a way to obtain more rare class data, it will lead to a significant labor overhead. Third, some rare categories in public RF datasets may be critical in practical applications, such as rare cases in medical diagnosis. Therefore, to mitigate the impact of the category imbalance problem on the model, the importance of all categories need to be fully considered when training the model.

For example, the datasets collected in one propagation environment will be hard to transfer to other environments. Due to different deployment environments (e.g., sensors and layouts) or user characteristics (e.g., genders and preferred hands) in real RF sensing applications, the generated user-side training data will usually be imbalanced. To study the limitations of directly applying the above methods for RF datasets, we choose a public RF dataset (i.e., Widar 3.0 [53]) and compare it with a CV dataset [28]. In addition, a public toolbox (i.e., imbalanced-learning [20]) is used to generate both imbalanced datasets based on the class imbalance ratio (see its definition in Section IVA).

Table 1 shows the learning results on a class-imbalanced CV dataset and an RF dataset. Both of them follow the long-tailed distribution, which are divided into minority and majority classes. We choose training accuracy and **Index of Balanced Accuracy (IBA)** [12] to evaluate the learning process on the imbalanced datasets. IBA can be defined by $IBA = (TPR + TNR)/2$, where **True Positive Rate (TPR)** refers to the ratio of the number of samples correctly identified by the classifier as positive cases to the total number of positive cases, and **True Negative Rate (TNR)** refers to the ratio of the number of samples correctly identified by the classifier as negative cases to the total number of negative cases. To intuitively describe the influence of minority class to the classifier, we only consider the balance accuracy of positive and negative two categories, i.e., minority and majority classes.

For both datasets, as the class imbalance ratio grows, the training accuracy also increases, while the IBA decreases. Specifically, with a ratio of 20, the training accuracy on the RF dataset is 85.10%, while the IBA is down to 50.88% (50% is the lowest limit of the IBA), which means the classifier cannot effectively predict the minority classes.

Remark: Both methods [22, 55] usually ignore the fact that the oversampled minority classes may have *poor generalization* or *low diversity*. We explore a more effective way to overcome the problem

Table 1. Learning Results from Class-imbalanced Datasets

Dataset	CV [28]			RF [53]		
Ratio	1	10	20	1	10	20
Training Acc (%)	87.11	90.22	95.04	79.30	83.53	85.10
IBA (%)	86.27	79.81	61.14	78.71	64.03	50.88

of imbalanced RF data in different environments by focusing on fully utilizing the features in RF signals and leveraging GANs to augment the HAR dataset.

3.2 Motivational Study

Unlike simple data augmentation methods (e.g., image rotation, flipping, and introducing Gaussian noise), GAN can generate more realistic images by adversarial learning [13]. GAN are made up of two components: a generator and a discriminator. The generator generates samples, while the discriminator attempts to distinguish these samples from real data. Through this competition, the two components can improve each other, leading to the generation of high-quality synthetic data. For example, RF-based gesture recognition systems (e.g., WiGAN [9] and SS-GAN [36]) have used GAN to enhance gesture features. Unlike the related works that only augment the original data, we propose new data synthesis techniques to address the problem of *imbalanced RF data*, which also help to adapt to different deployment environments and devices. Using GAN for RF-based HAR, two major problems shall be addressed as follows.

3.2.1 How Does GAN Learn Time–Frequency and Semantic Information in RF Signals? In RF sensing applications, different wireless devices have been leveraged for HAR in which the measured wireless data stream in the time-domain is fundamentally abstract and not easy to interpret. Therefore, using **short-time Fourier transform (STFT)** [23], the time–frequency RF spectrum can be obtained to extract the semantic information (e.g., activity features) of HAR, thus boosting the classification performance. To increase the time–frequency domain resolution, we design new mapping approaches (i.e., using an autoencoder) in the GAN framework.

3.2.2 How Can GAN Generate Minority Class Data in Imbalanced RF Datasets? Generally, it is challenging for GAN to generate high-quality samples with a small training set from a minority class. The BAGAN model [30] is proposed to learn useful features from the majority class and to synthesize images for minority classes. However, the generated results maybe unstable when different classes become similar, while the class similarity is usually high in RF sensing applications (e.g., pushing and pulling in gesture recognition). In this article, we couple GANs with an autoencoder and develop a data augmenting module with an associated loss function to explicitly improve the diversity of synthesized data.

4 PROPOSED METHODOLOGY

In this section, we first formulate the imbalance problem and then provide an overview of TFSemantic, as well as its key strategies and hyperparameters in detail.

4.1 Problem Formulation

In this article, the aim of TFSemantic, parameterized by Ω, is to restore the balance of RF data. First, we develop a Semantic Extraction Module, parameterized by $\mathcal{S}_\Theta$, to extract semantic features from input observations (i.e., an RF signal matrix). In addition, a Semantic Distribution Module, parameterized by $\mathcal{S}_\Phi$, is designed for achieving reliable data generation. To generate diverse data,

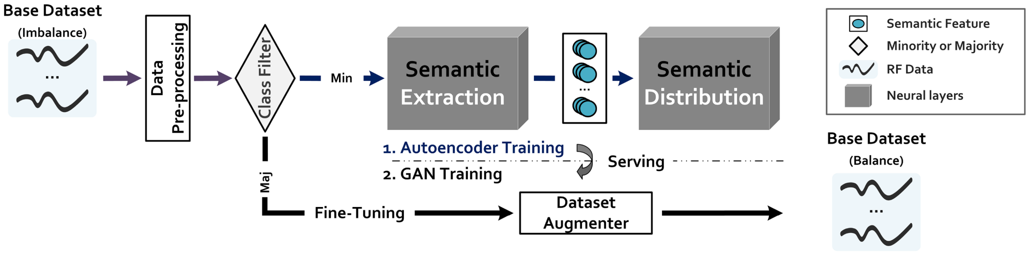


Fig. 2. An overview of the TFSemantic framework.

we also adopt new metrics (e.g., ensemble margin [10]) in a Dataset Augmenter Module to evaluate the discriminability and diversity of the synthetic samples in the Semantic Distribution Module.

Suppose the dataset $\mathcal{D} = \{\mathbb{R}^1, \mathbb{R}^2, \dots, \mathbb{R}^N\}$ contains both majority and minority classes, where $\mathbb{R}^i$ represents the set of samples in class i and N is the total number of classes in the dataset. The ratio of the number of samples in the majority class to the number of samples in the minority class (i.e., the class imbalance ratio) is defined as $\mathcal{R} = \frac{\text{Majority}(\mathcal{D})}{\text{Minority}(\mathcal{D})}$. TFSemantic will synthesize data for the minority classes until such classes are balanced with the majority class (i.e., when $\mathcal{R}$ equals to 1). The ensemble margin after training instances effectively indicates the performance of a classification algorithm. Generally, imbalanced datasets will lead to high ensemble margin values for the majority class and low values for minority class. Therefore, we use ensemble margin to evaluate the impact of $\mathcal{R}$ and the augmented dataset (e.g., when synthesized data are included). Essentially, an objective function can be formulated as

$$\Omega^*(\mathcal{D}, \mathcal{R}) = \arg\max_{\Omega} \left[\frac{\mathcal{V}_{\mathrm{y}} - \max_{c \neq y}(\mathcal{V}_c)}{\sum_{c=1}^{N}(\mathcal{V}_c)} \right]_{\text{margin}}, \tag{1}$$

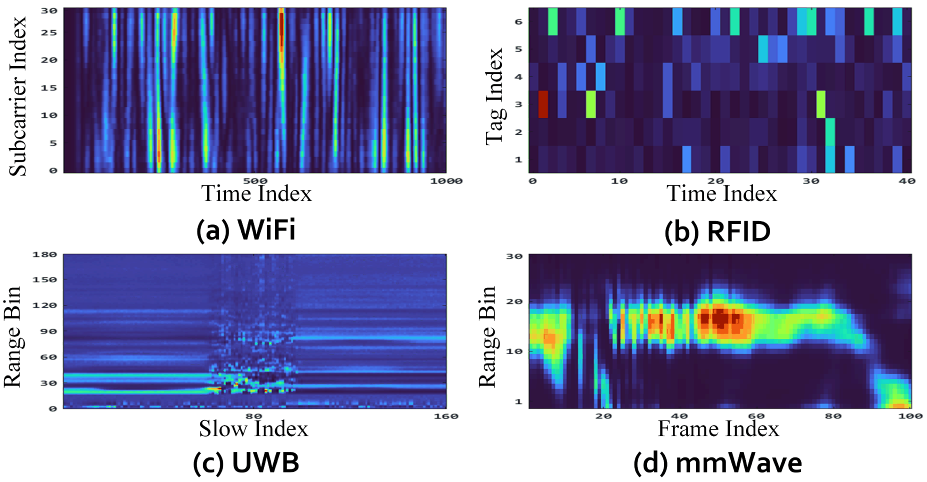
where $\mathcal{V}_{\mathrm{y}}$ is the number of votes (i.e., classified results) for the true class y, $\mathcal{V}_c$ is the amount of votes in the other class c, and $[.]_{\text{margin}}$ is the ensemble margin for each class in $\mathcal{D}$, which is in the range $[-1, +1]$. Generally, a positive ensemble margin corresponds to correctly classified examples. To improve the discriminability of data augmentation for minority classes, our goal is to maximize the ensemble margin.

4.2 Overview of TFSemantic

We design an effective GAN-based dataset augmentation framework for RF sensing datasets. Specifically, we couple GANs with an autoencoder in TFSemantic to recover the balance of the dataset by synthesizing signals in minority classes. As shown in Figure 2, our proposed TFSemantic framework consists of four modules, which is trained in two stages. First, TFSemantic takes the imbalanced base dataset and then performs data pre-processing over four different RF devices. And the pre-processed base dataset is then fed into the class filter module to discriminate the minority classes and majority class. In the first stage, the GAN learns and connects the semantic features (e.g., activity features) of the RF signals by training the autoencoder, where two modules (i.e., semantic extraction and semantic distribution) are designed and incorporated. In the second stage, the GAN uses the trained autoencoder to generate samples for the minority class. We design the associated loss function in the dataset augmenter module to achieve effective training of the GAN.

4.3 Key Design Strategies

In this section, we discuss the key strategies used in each module of our TFSemantic framework.



(a) WiFi

(b) RFID

(c) UWB

(d) mmWave

Fig. 3. Illustration of the RF signals from different sampling devices.

4.3.1 Data Pre-processing Module. Due to different frequency bands, RF communication protocols, and hardware designs, different RF platforms (i.e., WiFi, UWB, RFID, and mmWave) have their unique features in specific deployment environments. To tackle the different timestamps in different sampling batches, we synchronize sampled data with a fix-sized sliding window over the entire time-domain sequence. As shown in Figure 3, the RF data collected by four different RF devices for the same activity will be converted into different matrices. We then perform different data pre-processing operations on the different RF datasets.

WiFi CSI data are collected for HAR. However, the WiFi CSI data contains considerable phase noises caused by asynchronous transceivers and hardware defects. To filter out uncorrelated noises, only the dominant measurement data will be extracted. We use the conjugate multiplication of CSI values between two antennas to remove phase noise [53]. For UWB devices, we obtain the fast-slow RF images and then normalize the UWB image data [34]. For the RFID dataset, we use received signal strength, which are denoised by a Gaussian-weighted averaging method [47]. For the mmWave dataset [25], we remove the correlated noise and environmental interference from the original signals using 3D **Fast Fourier transform (FFT)**. RF spectrum features of the four datasets become more clear after static removal, conjugate multiplication, and noise removal.

4.3.2 Semantic Extraction Module $\mathcal{S}_{\Theta}$ [Stage I]. This module is the encoder part of the autoencoder. To supplement the data of minority classes, the base dataset $\mathcal{D}$ after preprocessing is input to the class filter module to determine the minority and majority classes for data augmentation.

Motivated by the STFT module in the classification of time-series data [23], we use multi-sensory channels to design the Semantic Extraction Module through integrating the channel attention mechanism [37] with STFNet-convolution [46] to strengthen the generalization of our deep model. It exploits multiple branching channels to sense changes in RF spectrum in the time–frequency domain and thus enables the effective extraction of semantic features. As shown in Figure 4, we design a new time–frequency semantic feature learning layer with two key strategies, i.e., Fourier SFC and attention SFE, to obtain a fine-grained semantic feature representation in the time chunks from the output of each channel.

Fourier SFC: In STFT, the time–frequency resolution is controlled by a sliding window of a fixed length. The fine-grained frequency representation is limited by the window length, with a

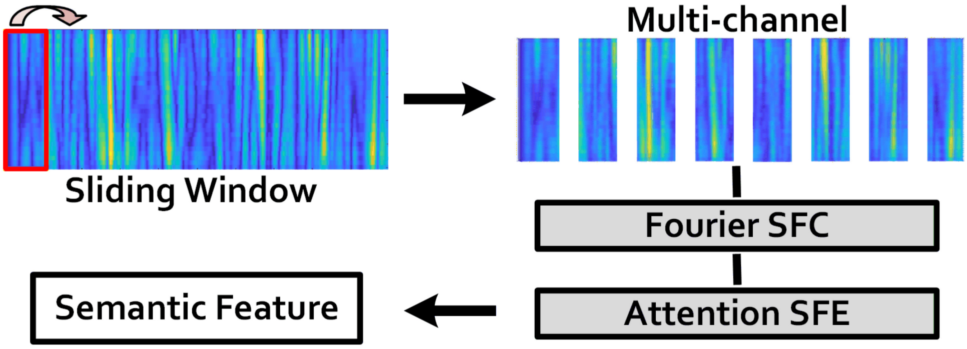


Fig. 4. Learning process of the semantic extraction module.

tradeoff between high frequency resolution and high time resolution. Specifically, we initialize the size of the sliding window based on the sampling rate of the RF data, and then the sliding window divides the timing data equally into p small blocks of time. To ensure flexibility and generalization for different RF sensing tasks, we integrate the frequency domain transform into the deep learning pipeline, which can be optimized by standard methods as in Section 4.4.

Denote the input to the pipeline as $\mathrm{X} \in \mathbb{R}_{\mathrm{d}\times\mathrm{m}\times\mathrm{n}}$, where $\mathbb{R}$ contains n signal matrices, and the size of X is $d \times m$. The width of the window $\mathcal{W}(\tau)$ is denoted by τ. The STFT operation in a single pipeline can be formulated as

$$STFT(X, \mathcal{W}) = \sum_{i=1}^{d} X_{[i,\tau]} \cdot \mathcal{W}(i - s \cdot m) \cdot K_{[i,\tau]}, \tag{2}$$

where s is the step of sliding and K is the convolution kernel of the channel. In this article, Fourier SFC is a multi-channel parallel weighting module and we will perform (2) with different resolutions of X. First, we segment each input signal matrix X into p frames with the same size along the time dimension and then feed it into the corresponding channel of Fourier SFC.

Because of the limit time–frequency resolution of window $\mathcal{W}(\tau)$, multiple transform channels with different window widths will be applied in the Fourier SFC. The output for the jth segmented frame X^j with window $\mathcal{W}^{[j]}(\tau)$ is given by

$$X^j{}_{SFC} = STFT\{X^j, \mathcal{W}^{[j]}(\tau)\} + Bias(X), j = 1, 2, \ldots, p, \tag{3}$$

where $Bias(X)$ is the bias for specific input data. The multi-channel output from Fourier SFC is denoted by $X_{SFC} = [X^1_{SFC}, X^2_{SFC}, \ldots, X^p_{SFC}]$. To adapt to different RF sensing tasks, the parameters of Fourier SFC should also be adjusted appropriately.

Attention SFE: In the attention SFE model, the main goal is to reinforce the multi-channel output X_{SFC} obtained from Fourier SFC. To detect the correlation of channels, the attention mechanism is applied. Specifically, we capture the local cross-channel interaction by considering adjacent channels. We define the underlying mapping as $g(c) = \mu^{(g' * c + b)}$, where $(g' * c + b)$ is a linear function, c is the target sample, and b and g' are the linear parameters of the mapping space.

Due to the limited relationship represented by the linear function, we set the channel dimension μ as a power function to construct the nonlinear mapping relationship. Accordingly, the correlation weights between channels in X_{SFC} can be computed by

$$\omega = \vartheta(g(X_{SFC}) \cdot f(X_{SFC})), \tag{4}$$

Table 2. Performance of the Reconstructed Signal Spectrum

Training batch	Accuracy	Cosine similarity	Pairwise distance
Epoch-100	84.71%	35.54%	15.6073
Epoch-200	88.29%	45.34%	13.5136

where $f(X_{SFC})$ is the channelwise global average pooling and ϑ is the Leaky ReLU function. An adjusted X_{SFC} can be obtained by multiplying the output channel weights ω by the channel output values X_{SFC}, where the dimension of ω should be expanded to be the same as X_{SFC}. Therefore, dimensional reduction can be avoided and cross-channel interactions can be effectively captured. This can further enhance the time–frequency resolution of X_{SFC}.

In summary, the sensed signal matrices serve as input to the Semantic Extraction Module. We obtain a set of enhanced multi-channel outputs through using the key strategies (i.e., Fourier SFC and attention SFE), which will be filtered and stitched by a convolutional filter W_{est}. We denote F^* as the final output (i.e., the semantic features of the input signals), which can be written as

$$F^* = W_{est} \cdot \{\mathcal{S}_\Theta(X)\}^T, \tag{5}$$

where $\mathcal{S}_\Theta(X)$ represents the Fourier SFC and attention SFE operations on the input data X with parameter Θ.

4.3.3 Semantic Distribution Module $\mathcal{S}_\Phi$ [Stage I]. This is the decoder part of the autoencoder. To use the semantic features produced by the previous module, we design a generative neural network, which reconstructs input samples by predicting the conditional semantic distributions and reassembling the extracted semantic features. However, the distribution of the collected RF signals from different RF platforms may vary considerably in different environments. Therefore, we use a DWT method to complement the details of different frequency bands in the reconstructed samples, which allows our generated data to approximately represent real samples that are collected from different perceptual environments.

The high-resolution semantic features F^* can be directly used to improve the accuracy of classification, because they carry the key part of information in the signal. However, it is challenging to recover the input data with the obtained semantic features, because the detailed frequency distribution in RF spectrum is smoothed. To validate this observation, we train a semantic extractor for a specified batch and then test the classification and reconstruction performance using CNNs. In addition, we use the cosine similarity and pairwise distance [6] to measure the similarity of two signals.

As shown in Table 2, the classification accuracy is between 84.71% and 88.29% by using semantic features F^*. In addition, we use a fully connected layer to reconstruct signals. We find that the quality of reconstructed signal is poor, since the cosine similarity is only about 40% and the pairwise distance is around 14%. Thus, to restore the complete spectrum information of the original signal, we utilize semantic **skeleton prediction (SP)** as a key strategy as follows.

Semantic SP: The semantic SP in $\mathcal{S}_\Phi$ aims to predict the original signal skeleton, which measures the location distribution of F^* in the original signal spectrogram. Generally, the main features of RF signal are preserved in low-frequency components while the high-frequency components reflect the subtle detail information. One way to perform signal enhancement using DWT is to decompose the signal into its frequency components and then perform different types of processing or filtering on different frequency bands. Specifically, in this article, we want to generate a signal that is as rich in high-latitude features, i.e., high-frequency features and low-frequency features, as the real signal. This can improve the overall quality of the generated signal and make it easier to be analyzed or understand. To preserve most frequency information, we exploit DWT to decompose signal into different frequency bands and select different frequency bands for signal supplementation according to practical needs.

The low-frequency region intensity $\Xi_1(x, y)$ and high-frequency region intensity $\Xi_2(x, y)$ in semantic SP can be written as

$$\Xi_1(x, y) = \sum_{i,j \in M} w(i, j) \left|A(x + i, y + j)\right|, \tag{6}$$

$$\Xi_2(x, y) = \sum_{i,j \in M} w(i, j)(H(x + i, y + j))^2, \tag{7}$$

where $A(x, y)$ is the low-band wavelet coefficient at location (x, y), $H(x, y)$ is the high-frequency-band wavelet coefficient, and w is the mask operator with a window size of $M \times M$. Different from the noise reduction using DWT, we aim to restore the original distribution of the signal (e.g., complementary noise). Therefore, we will use the coefficients of both frequency bands to guide the fully connected layer (i.e., semantic mapping of the signal).

In summary, both X and F^* serve as the input to the Semantic Distribution Module. Also, semantic SP is used to predict the signal skeleton C_{dis} from X. Then, by combining the semantic features F^* and the predicted signal skeleton C_{dis}, the output signal X_{dis} can be reconstructed by $X_{dis} = C_{dis} \cdot F^*$, which is similar to the original signal. Specifically, the process of acquiring C_{dis} can be represented as

$$C_{dis} = W_{dis} \cdot \{\mathcal{S}_\Phi(\Xi_1, \Xi_2)\}^T, \tag{8}$$

where W_{dis} is a convolutional filter similar to W_{est}, which can model the spectral distribution of the semantic features. Note that to distinguish the high-frequency bands in various directions corresponding to $\Xi_2(x, y)$ in the signal (i.e., the time–frequency matrix), we only retain the horizontal direction to obtain frequency domain multi-channel data.

Remark: In Stage I, our main goal is for the autoencoder to learn how to extract the semantic features from the real signal and how to use the semantic features to restore the real signal. During the training of the autoencoder, we use Kullback–Leibler divergence [15] to constrain the entire extraction and distribution operations to optimize the difference between the distribution of the output X_{dis} and that of the true signal X in each epoch of the training process, which is defined as

$$\mathcal{L}_{e \cdot d} = \mathcal{L}_{kl}\{\mathcal{N}(X) \| \mathcal{N}(X_{dis})\}, \tag{9}$$

where $\mathcal{L}_{kl}$ is the calculated function of Kullback–Leibler divergence and $\mathcal{N}(\cdot)$ is the softmax function to obtain the probability distribution.

4.3.4 Dataset Augmenter Module [Stage II]. In the second stage of GAN training, we need to integrate the Semantic Extraction and Semantic Distribution into the Dataset Augmenter. To achieve the reliability and efficiency of the augmentation process for the entire base dataset $\mathcal{D}$, we design a light-weight two-stage fine-tuning scheme, i.e., separately training for the extraction layers and distribution layers. This way, we use randomly sampled data as input and then generate a large amount of synthesized data, which is similar but yet different from the real samples used in training. The dataset augmenter will output an augmented base dataset consisting of the minority classes that are augmented with synthesized data and the fine-tuned majority class.

By training the Semantic Extraction Module and Semantic Distribution Module in Stage I, the autoencoder has already leaned how to extract semantic features and their distribution pattern from the signal matrices, which can be used to provide a potential semantic mapping space for the GAN in our Dataset Augmenter Module.

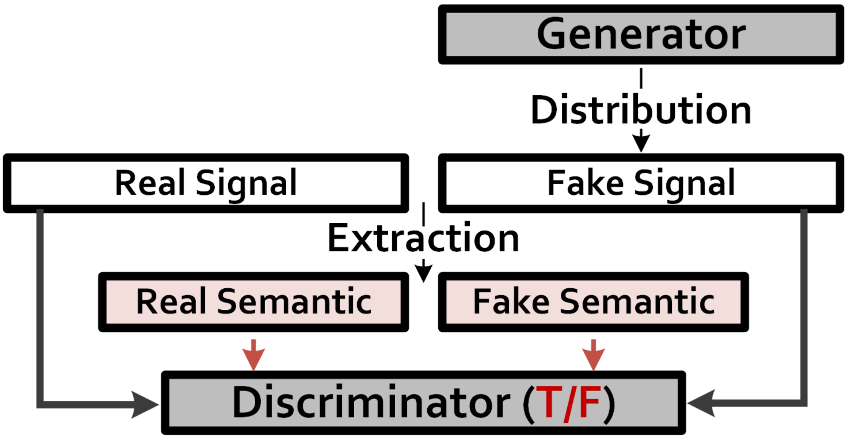


Fig. 5. The training process of the GAN model.

As shown in Figure 5, using the generator, we obtain the initial signals by using randomly sampled data, which do not contain meaningful physical characteristics. To extract the same properties from real-world perception signals ($X_{real} \in \mathbb{R}$), the well-trained Semantic Distribution Module can generate fake signals X_{fake}, which are similar but different from the input training signals. Then we utilize the Semantic Extraction Module to extract the semantic feature of X_{real} and X_{fake} and feed the results into the discriminator. The process of the adversarial network is that the generator aims to minimize the difference between the input X_{real} and X_{fake} and the discriminator aims to maximize the difference between X_{real} and X_{fake}. In particular, we design several efficient loss functions to perform specific tasks at each step as follows.

Reconstruction Loss $\mathcal{L}_{rec}$: In the Semantic Extraction Module and Semantic Distribution Module, the loss function $\mathcal{L}_{e\cdot d}$ is used to restore the original signal. Different from the loss, we need to generate signals that are similar but not having identical characteristics as the real signal. Therefore, the new loss function $\mathcal{L}_{rec}$ is defined by

$$\mathcal{L}_{rec} = \sum_{\propto} \mathcal{L}_{kl} \left\{ \mathcal{N}(X^{\propto}_{real}) \| \mathcal{N}(X^{\propto}_{fake}) \right\}, \tag{10}$$

where $X^{\propto}_{fake}$ and $X^{\propto}_{real}$ are the adjusted time–frequency resolutions of X_{fake} and X_{real}, respectively, and $\sum_{\propto}$ indicates the sum of all returned values for each pair of $X^{\propto}$. In addition, the batch gradient descent is used to minimize the loss function.

Adversarial Loss $\mathcal{L}_{adv}$: A generative adversarial loss $\mathcal{L}_{adv}$ is used to simulate the distribution of the real signal. The discriminator is trained to distinguish X_{fake} and X_{real}, which is formulated as

$$\mathcal{L}_{adv}(X_{real}, X_{fake}) = \int_{X_{real}} p_r(X_{real}) \log(D(X_{real})) d_{X_{real}} + \int_{X_{fake}} p_f(X_{fake}) \log(1 - D(X_{fake})) d_{X_{fake}}, \tag{11}$$

where p_r and p_f are the equalization parameters of the generator and the discriminator, respectively, and $D(*)$ is the maximum likelihood function for the discriminator.

To make the augmented datasets generalizable across different tasks, we use the boundary distribution of the discriminator, and control the variation of the generated conditional distribution to ensure their duality. Let L_{cos} denote the cosine distance between F_{real} and F_{fake} to control the similarity or difference, where F_{real} and F_{fake} are the semantic features of X_{real} and X_{fake}, respectively. The objective is defined as

$$\mathcal{L}_{cos}(F_{real}, F_{fake}) = \frac{F_{real} \cdot F_{fake}}{\max(\|F_{real}\|_2 \cdot \left\|F_{fake}\right\|_2, \upsilon)}, \tag{12}$$

where υ is the weight. The combination of Equations (11) and (12) constitutes a coupled process of adversarial generation training, which is used to guide the generation of data. The data are batched through the generator and the discriminator, and the model weights are fine-tuned to optimize the loss function during adversarial training.

Augmenter Loss $\mathcal{L}_{aug}$: As elaborated in the problem formulation section, the purpose of TFSemantic is to restore the balance of the dataset. After augmenting the minority signal classes with synthesized data by the GAN, we need to evaluate the quality of the synthesized data (i.e., the effectiveness of our solution on tackling the imbalance problem). Therefore, the augmenter loss function $\mathcal{L}_{aug}$ is defined as

$$\mathcal{L}_{aug}(\mathcal{D}, \mathcal{D}^*) = \Omega^*(\mathcal{D}, \mathcal{R}) - \Omega^*(\mathcal{D}^*, \mathcal{R}), \tag{13}$$

where $\mathcal{D}^*$ denotes the re-balanced dataset, and it will be used in the weight function Ω^* to calculate the ensemble margin. Finally, we design the joint GAN module optimization objective as follows:

$$\mathcal{L}_{sum}(\mathcal{D}, \mathcal{D}^*) = \alpha\mathcal{L}_{rec} + \beta\mathcal{L}_{\cos} + \lambda\mathcal{L}_{aug} + \mathcal{L}_{adv}, \tag{14}$$

where α, β, and λ are the hyper-parameters. When the real signal and synthesized signal are too similar, we have $\beta > 0$; otherwise, we have $\beta < 0$.

Remark: In the second stage, we use $\mathcal{L}_{sum}(\mathcal{D}, \mathcal{D}^*)$ to evaluate both the majority and minority classes in $\mathcal{D}$ and $\mathcal{D}^*$. In addition, the gradient descent loss values are used to initialize the generator, which allows TFSemantic to start adversarial training from a more stable point, which can alleviate the convergence problem of traditional GAN training.

4.4 Hyper-parameters

To achieve better generalization for all RF-based datasets, a necessary step before applying TFSemantic to a new dataset is to determine the proper values for the set of hyperparameters. We use an abstraction-based cost function based on coarse-grained space abstractions [2] to find suitable values for the hyper-parameters, which can effectively adapt to different RF sensing devices and scenarios for data augmentation, and support applications with timeliness requirements.

5 EVALUATION

In this section, we evaluate TFSemantic over different RF datasets and compare with several state-of-the-art methods.

5.1 Experimental Setup

5.1.1 Dataset. To cover a broad range of RF sensing tasks in our evaluation, we choose four public RF datasets. Their relevant information is summarized in Table 3.

WiFi: The WiFi datasets are usually collected by the Linux 802.11n CSI Tool, which contains the amplitude and phase of CSI data with the three antennas. In the dataset in Reference [53], 25 subjects were involved for the data collection in five different environments. Six receivers were used in each activity, where each antenna transmits 30 subcarriers, with a sampling frequency of 1, 000 Hz. There are 22 activities (e.g., Push, Pull, Sweep, Clap, Side, and Draw).

UWB: The UWB dataset was acquired with XeThru X4M03. The UWB radar system was mounted on a wall. In the dataset in Reference [34], participants were lying in the middle of the bed at a supine position, where line of sight of the radar is required. The dataset contains six subtle

Table 3. Statistics of the Datasets Used in Evaluation

Sensor	Activity (types)	Sample	Sampling Rate
WiFi [53]	Gesture (22)	5,400	1000 Hz
UWB [34]	Activity (6)	960	10 Hz
RFID [47]	Gesture (6)	900	13.56 MHz
mmWave [25]	Gesture (13)	24,050	4 MHz

activities: from supine to left lateral position, from left lateral to supine position, from supine to right lateral position, from right lateral to supine position, from supine to prone position, and from prone to supine position. Each data sample is a 16-s UWB radar signal output in a two-dimensional array, where the x-axis represents slow time and the y-axis is the distance box (i.e., fast time).

RFID: In the RFID dataset [47], the signal was sent and received using H47 RFID tags and directional antennas powered by an Impinj R420 RFID reader. Six tags were used in the acquisition process and three polarized antennas were used to interrogate these tags to ensure that each RFID tag was covered by at least one antenna. Different from other datasets, the RFID dataset was collected by using a robotic arm to simulate the activity.

mmWave: The mmWave dataset [25] was obtained based on a TI AWR1843 mmWave radar and a DCA1000 real-time data acquisition board. Six predefined gestures (push, pull, slide left, slide right, rotate clockwise, and rotate counterclockwise) and seven negative samples (lift right arm, lift left arm, sit, stand up, wave, turn, walk). The difference between the predefined activities and negative samples will be discussed in the experiment. The original captured signals were processed into dynamic range angle image sequences after 3D-FFT and denoising.

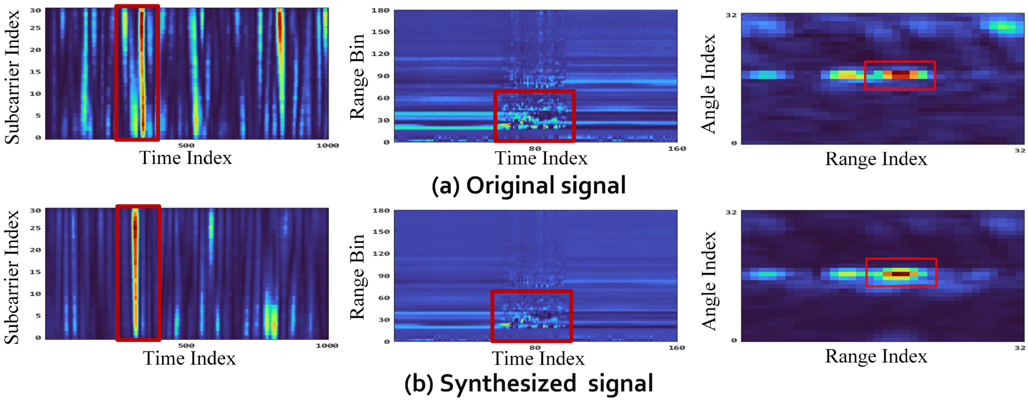
5.1.2 Metrics. Generally, the overall accuracy of the imbalanced dataset is not adequate to show whether the imbalance problem has been effectively solved. Therefore, we quantify the TFSemantic performance using the recognition accuracy of minority classes and G-mean [14]. The formula for G-mean is expressed by $sqrt(TPR \cdot TNR)$. The G-mean is a metric that measures the overall performance of a classifier on an imbalanced dataset, which is calculated as the geometric mean of the recall values for each class and is often used as an alternative to the overall accuracy in imbalanced datasets. We repeat each experiment with a different random seed and report the average value of these metrics.

5.1.3 Baselines. We focus on three different tasks to evaluate the performance of TFSemantic, i.e., GAN-based data augmentation, imbalanced learning, and learning-based sensing. For GAN-based data augmentation, we compare the proposed TFSemantic system with GAN [13], ACGAN [32], and BAGAN [30] to evaluate the quality of synthesized data. For solving the imbalanced learning problem, we compare TFSemantic with SMOTE [22] in their sampling methods and CSSVM [16] with respect to their cost-sensitivity. Last, the augmented data by TFSemantic will be used for training different **deep neural networks (DNN)**, including AlexNet [18], DSN [3], OneFi [42], and RFNet [8], to evaluate the impact on learning-based RF sensing.

We implement TFSemantic with PyTorch and the above baseline schemes are validated with their open source codes. Training of TFSemantic is performed on a server with two GPUs (RTX 3080Ti). The training time over 100 epochs for a single task is about 1.5 hours and the average test time is 10 s.

5.2 Augmented Datasets

Compared with traditional data augmentation approaches (e.g., rotation as in OneFi [42]), our proposed method uses the time–frequency semantic GAN model to generate diverse RF data. Figure 6



(a) Original signal

(b) Synthesized signal

Fig. 6. Comparison of the original signal and synthesized signal by TFSemantic.

Table 4. Average SSIM Values for Four Different Wireless Devices

Methods	WiFi	RFID	UWB	mmWave
GANs [13]	86.30%	84.40%	83.89%	87.07%
ACGAN [32]	60.92%	67.82%	59.91%	70.45%
BAGAN [30]	49.51%	50.26%	50.83%	54.72%
TFSemantic	36.15%	37.90%	35.21%	40.39%

shows the raw signal and our synthesized signal, respectively, where the key regions are marked by red boxes. We find that the critical regions can be well generated and the noisy data in the irrelevant regions are effectively suppressed.

To verify the diversity of augmented data, we use the **Structural Similarity Index (SSIM)** metric [41] to measure the similarity between the generated signal and the real signal. The SSIM is a method for measuring the similarity between two images. It is commonly used to evaluate the quality of generated images in computer vision, because it is based on the idea that the human visual system is highly adapted to perceive structural information in images. SSIM is calculated by comparing the local patterns in the two images, taking into account the luminance, contrast, and structure of the images. The resulting value is a number between −1 and 1, where a value of 1 indicates that the two images are the same, and a value of −1 indicates that they are completely different. We calculate the average SSIM by repeating the data augmentation process randomly for each type of RF data. The SSIM metric is defined by

$$\mathrm{SSIM}(X,Y) = [\mathcal{A}(X,Y)\cdot\mathcal{B}(X,Y)\cdot C(X,Y)] = \frac{(2\mu_x\,\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)}, \tag{15}$$

which is measured based on three elements between two samples (i.e., X and Y), namely luminance, contrast and structure, denoted as $\mathcal{A}(.)$, $\mathcal{B}(.)$, $C(.)$. μ_x and μ_y are the means of X and Y, respectively, and σ_x and σ_x are the variances of X and Y, respectively. σ_{xy} is the covariance of X and Y, and c_1 and c_2 are two constants calculated based on the range of pixel values, respectively. For Equation (15), a fixed size window is taken from the image, and then the window is continuously slid to perform the calculation, and, finally, the average value is considered as the SSIM. Table 4 shows that the average SSIM values of TFSemantic are all considerably smaller than the baseline GAN methods for all the four devices, which means that TFSemantic can achieve the best diversity in data augmentation.

Table 5. Performance of Augmented Datasets on the Three Tasks

Task	GAN-based Augmentation								Imbalanced Learning					
Methods	GAN		ACGAN		BAGAN		TFSemantic		SMOTE		CSSVM		TFSemantic	
Metrics	Acc	G-mean	Acc	G-mean	Acc	G-mean	Acc	G-mean	Acc	G-mean	Acc	G-mean	Acc	G-mean
D40%	81.40	75.29	81.76	76.32	84.29	78.94	**89.03**	**84.62**	78.83	75.85	**84.16**	82.06	83.35	**82.41**
D60%	78.47	70.31	76.24	69.88	84.53	80.95	**90.62**	**85.58**	80.69	71.39	79.20	75.96	**85.12**	**82.34**
D80%	68.35	60.76	63.38	60.05	81.89	73.29	**84.10**	**81.07**	72.90	62.79	69.42	63.44	**84.91**	**81.22**
D90%	69.14	61.78	62.50	60.39	81.25	73.02	**82.09**	**75.48**	56.20	50.77	63.60	62.08	**80.73**	**73.72**
D95%	67.04	58.75	49.21	46.23	**80.20**	74.01	78.93	**75.34**	49.14	47.93	60.27	58.22	**69.20**	**66.07**

Task	Learning-based Sensing							
Methods	AlexNet		DSN		OneFi		RFNet	
Metrics	Acc	G-mean	Acc	G-mean	Acc	G-mean	Acc	G-mean
D40%	7.90(+)	8.43(+)	6.14(+)	8.24(+)	6.85(+)	7.45(+)	5.18(+)	6.83(+)
D60%	17.20(+)	16.46(+)	10.83(+)	14.45(+)	5.86(+)	10.98(+)	8.84(+)	12.89(+)
D80%	16.54(+)	18.48(+)	14.32(+)	17.53(+)	7.84(+)	12.07(+)	9.08(+)	10.43(+)
D90%	18.60(+)	20.21(+)	17.88(+)	21.34(+)	13.41(+)	17.47(+)	13.17(+)	16.91(+)
D95%	24.61(+)	27.90(+)	21.40(+)	25.47(+)	15.55(+)	17.19(+)	17.11(+)	18.05(+)

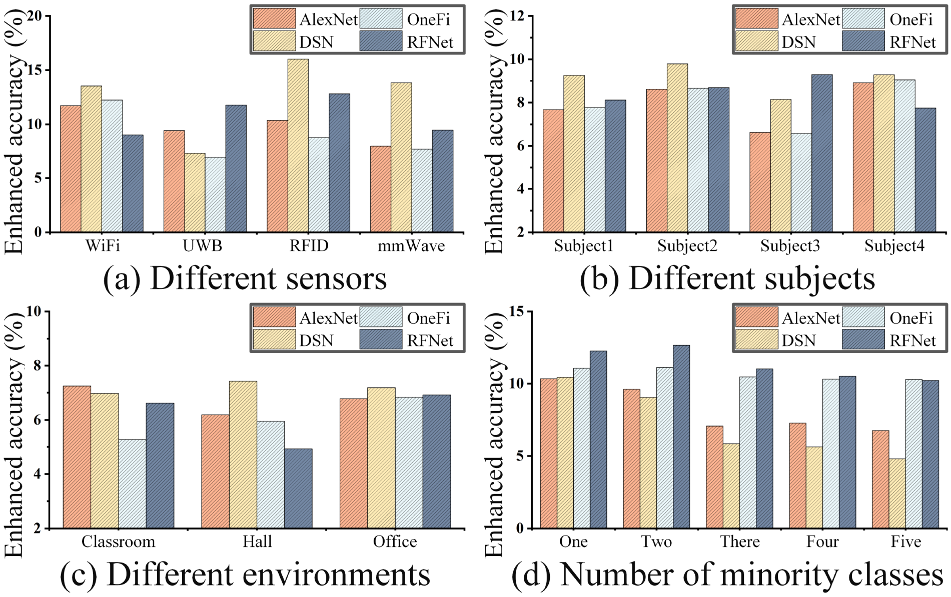
Note: (+) Represents the improvement in accuracy and G-mean; $Dx\%$ is the percentage of data removed from the original dataset to make the minority classes.

5.3 Results on the Three Tasks

For each dataset, we randomly select two classes of activities and remove the data by a specified ratio (denoted as $Dx\%$), which are regarded as minority classes. Considering the different sample sizes of the four datasets, we do not additionally set the ratio of the training set to the test set but use the removed data from the minority classes as test set. Table 5 shows the performance of TFSementic on the three tasks. Note that we use bold numbers and the symbol (+) to mark the best results and the improved accuracy, respectively.

5.3.1 GAN-based Data Augmentation. TFSemantic and three GAN-based methods are used to restore the balance of the WiFi dataset. For fairness, we use the same CNN model to calculate the recognition accuracy of the minority classes. It can be seen that TFSemantic achieves the best performance, and the highest recognition accuracy is 90.62% when the samples of minority classes are reduced by 60%. Meanwhile, the TFSemantic can reach the interval of 75% to 85% in the experimental results of the G-mean assessment index. Besides, we can see that TFSemantic has the largest G-mean value when Dx% is from 40% to 90% and also obtains the second G-mean value with Dx% with 95% when the number of samples is insufficient. Therefore, TFSeamtic demonstrates the high distinguishability of our synthesized signals.

5.3.2 Imbalanced Learning. For imbalanced learning, we use all four datasets to evaluate the learning performance. For fairness, we set their activity classifiers as fully connected layers. TFSemantic can greatly improve the average accuracy compared to the baselines. The highest accuracy of TFSemantic is 85.12%. When $Dx\% = 95\%$, the average accuracy of TFSemantic becomes 69.20%, which is still higher than that of SMOTE and CSSVM. In addition, we find the accuracy of SMOTE decreases to 49.14% when Dx% becomes 95%. This is due to the blind generalization in generating minority examples of SMOTE in the highly skewed case. Based on the G-mean metric, TFSemantic demonstrates the feasibility of solving the imbalance problem compared with other baseline models, which can improve the recognition accuracy of minority classes, while avoiding a decrease in the accuracy of majority classes.



(a) Different sensors (b) Different subjects

(c) Different environments (d) Number of minority classes

Fig. 7. Impact of different system parameters.

5.3.3 Learning-based Sensing. For learning-based sensing, we also use all four datasets. Different from the above dataset construction, we evaluate the full RF dataset. The constructed imbalanced dataset will be first used for the baseline model to perform the recognition calculations for minority classes. We then input the dataset to TFSemantic for dataset augmentation. Finally, the recognition calculation is performed once to obtain the enhancement of TFSemantic. Note that the sample size varies from dataset to dataset, so the amount of data in each category in a dataset is randomly cropped to a specified amount for ensuring that the sample size of the dataset used for training is consistent. The performance of TFSemantic is evaluated by comparing the recognition accuracy of the minority classes of the four DNN modes using both imbalanced and augmented datasets created by TFSemantic. The augmented dataset improves the recognition accuracy by 5% to 24% and the G-mean by 6% to 27%. The improvements of OneFi and RFNet are relatively smaller, both of which are few-shot learning models. These models can achieve a higher recognition accuracy with only a few training samples during a fine-tuning phase.

5.4 Robustness Study

We choose the enhanced accuracy as the y-axis in Figure 7, which is the magnitude of the accuracy improvement when TFSemantic is tested on the enhanced dataset, comparing with the original dataset. Figure 7(a) shows the accuracy on the datasets collected by different sensors. For RFID data using DSN, the enhanced accuracy is close to 15%. Figure 7(b) shows the accuracy on different subjects. We find that the average accuracy is similar over different subjects, which demonstrates the better adaptation of TFSemantic for different subjects. Figure 7(c) shows the accuracy in different environments. It is noted that the average improvement is 6% in the lobby case, which is higher than that of the classroom and office cases, because the lobby is more cluttered. Figure 7(d) shows that TFSemantic can well adapt to different numbers of minority classes.

For domain adaptation of RF sensing applications using the above four datasets, we consider that the dataset from the source domain is balanced, while the dataset in the target domain is imbalanced. Here a domain means a deployment environment. Specifically, the data of minority classes

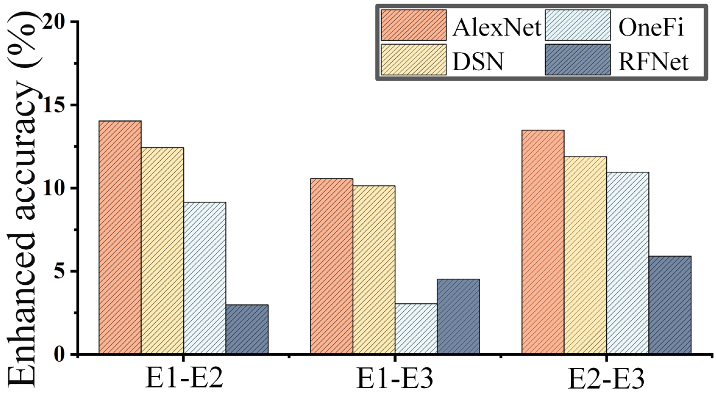


Fig. 8. Performance of adaptation to different environments: E1, E2, and E3 represent three different environments.

Table 6. Ablation Study of TFSemantic

Strategy I	Strategy II	Strategy III	Strategy IV	Accuracy (%)
				57.62
✓				73.88
✓	✓			82.79
✓	✓	✓		86.84
✓	✓	✓	✓	89.72

from the target domain will be combined with the data in the source domain (in a ratio of 1:20), which is used for dataset augmentation by TFSemantic. As shown in Figure 8, the average accuracy has been increased by about 4% to 12% over different learning methods by leveraging TFSemantic. Moreover, OneFi and RFNet have a smaller accuracy improvement, because their original accuracy is already high.

5.5 Ablation Study

To measure the impact of semantic extraction and semantic distribution on the TFSemantic performance, we conduct an ablation study on the WiFi dataset. Specifically, we study the influence of the four key strategies in TFSemantic, namely Data Preprocessing (Strategy I), Fourier SFC (Strategy II), Attention SFE (Strategy III), and Semantic SP (Strategy IV). Each strategy will be replaced by the simplest fully connected or convolutional layer to ensure the subsequent steps. Table 6 shows the impact of each key strategy on the overall system performance. As can be seen from the table, Data Preprocessing, Fourier SFC, and Attention SFE make larger contributions to TFSemantic.

5.6 Behavior Analysis

Using the WiFi dataset, we examine the relationship between the recognition accuracy of the minority classes and the parameters set in Section 4.1. As shown in Figure 9, different $\mathcal{R}$ values (i.e., 1, 2, 4, and 8) are set. We can see that as the increase of the ensemble margin, the accuracy for different $\mathcal{R}$ values also increases steadily. In addition, it is noted that the accuracy can reach up to 92.8% when $\mathcal{R} = 1$, because the classes in the dataset have already been balanced. The results indicate the superiority of the behavior of TFSemantic, which validates our optimization objective.

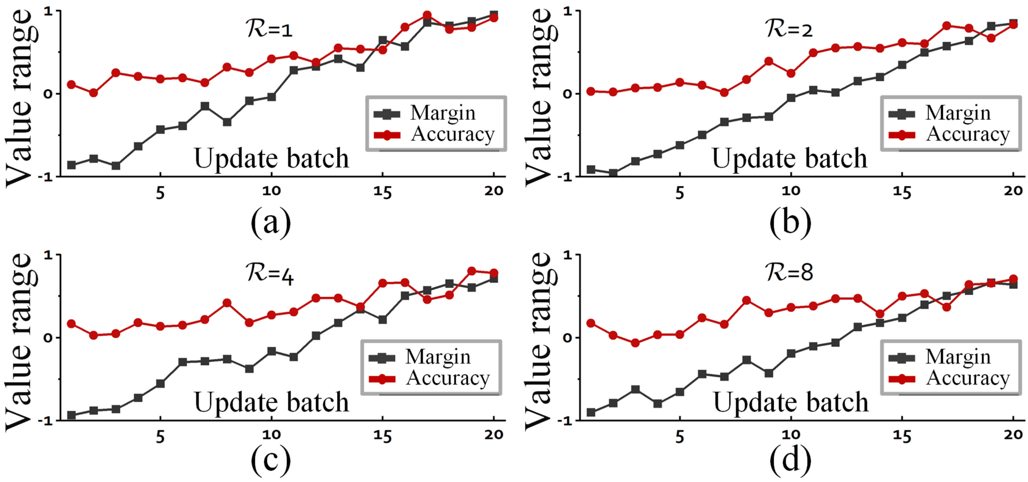


Fig. 9. Behavior analysis in TFSemantic for different $\mathcal{R}$ values.

6 DISCUSSIONS

In this section, we discuss the limitations of the proposed method and future research directions.

6.1 Computational Overhead

The utilization of generated data through dataset augmentation within the TFSemantic framework serves to mitigate the issue of imbalanced datasets, while also reducing the need for extensive data collection. The implementation of TFSemantic involves two phases, including dataset augmentation and training on the augmented dataset. While this approach leads to additional computational workloads, they are acceptable in light of the reduction in data collection efforts. In the future, we will focus on optimizing the system through the techniques such as dynamic planning and parallel computing to improve the computational efficiency of TFsemantic. Besides, a few-shot or zero-shot learning will be exploited as a potential solution to optimize the computational overhead of TFSemantic. Also, it can be used to improve the generalization and learning ability of the model during the model training phase.

6.2 Semantic Control

TFSemantic in our work is a GAN model redesigned based on the time–frequency semantic information of RF signals, where a generator network is trained to generate new RF data similar to the given dataset and a discriminator network is trained to distinguish the real data from the generated data. We generate minority classes of data for restoring the balance of the dataset. Therefore, when an extreme imbalance occurs, the minority class data for learning is very limited, which will lead to an unsatisfactory performance of TFSemantic. In a follow-up work, we consider extending semantic learning in TFSemantic to semantic control, where **conditional GAN (cGAN)** [31] will be exploited. The generator is tuned by additional inputs based on class labels, titles, or other forms. TFSemantic combined with cGAN can generate new examples of a particular class, e.g., signals that match a particular content description, or convert input signals from one style to another. This enables the above data-limited problem to be solved and will also enable TFSemantic to serve more relevant tasks (e.g., elderly fall detection, sports rehabilitation training, and gait recognition). Additionally, radio signals are characterized by parameters such as signal-to-noise ratio, bandwidth, and frequency, which are directly related to spectrum image quality. To generate data more close to signal in real world, we will also consider exploring the way to inverting synthetic spectrum data achieved by TFSemantic back to radio signals.

7 CONCLUSIONS

In this article, we proposed a time–frequency semantic GAN framework, TFSemantic, to address the imbalanced classification problem in RF-based sensing tasks. First, we discussed the related work and the class imbalance problem. Then we formulated the problem and provided the system framework including the data pre-processing module, the semantic extraction module, the semantic distribution module, and the dataset augmenter module. In addition, hyper-parameters were also examined. Finally, we validated the proposed TFSemantic framework using different RF datasets and comparison with several state-of-the-art methods. The results validated that our proposed method can effectively address the data imbalance problem and is effective in improving the performance of RF-based classification tasks.

REFERENCES

[1] Harshit Ambalkar, Xuyu Wang, and Shiwen Mao. 2021. Adversarial human activity recognition using Wi-Fi CSI. In *Proceedings of the IEEE Canadian Conference on Electrical and Computer Engineering (CCECE'21)*. 1–5.

[2] Sergiy Bogomolov, Alexandre Donzé, Goran Frehse, Radu Grosu, Taylor T. Johnson, Hamed Ladan, Andreas Podelski, and Martin Wehrle. 2016. Guided search for hybrid systems based on coarse-grained space abstractions. *Int. J. Softw. Tools Technol. Transf.* 18, 4 (Aug. 2016), 449–467.

[3] Konstantinos Bousmalis, George Trigeorgis, Nathan Silberman, Dilip Krishnan, and Dumitru Erhan. 2016. Domain separation networks. In *Advances in Neural Information Processing Systems*, D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett (Eds.), Vol. 29.

[4] Azhar Chara, Tianya Zhao, Xuyu Wang, and Shiwen Mao. 2023. Respiratory biofeedback using acoustic sensing with smartphones. *Smart Health* 28 (2023), 100387. https://www.sciencedirect.com/science/article/abs/pii/S2352648323000156

[5] Zhe Chen, Chao Cai, Tianyue Zheng, Jun Luo, Jie Xiong, and Xin Wang. 2021. RF-based human activity recognition using signal adapted convolutional neural network. *IEEE Trans. Mobile Comput.* 22, 1 (2021), 487–499.

[6] Fabrizio Costa and Kurt De Grave. 2010. Fast neighborhood subgraph pairwise distance kernel. In *Proceedings of the International Conference on Machine Learning (ICML'10)*. 255–262.

[7] Lang Deng, Jianfei Yang, Shenghai Yuan, Han Zou, Chris Xiaoxuan Lu, and Lihua Xie. 2022. Gaitfi: Robust device-free human identification via WiFi and vision multimodal learning. *IEEE IoT J.* 10, 1 (2022), 625–636.

[8] Shuya Ding, Zhe Chen, Tianyue Zheng, and Jun Luo. 2020. RF-net: A unified meta-learning framework for RF-enabled one-shot human activity recognition. In *Proceedings of the ACM Conference on Embedded Networked Sensor Systems (SenSys'20)*. 517–530.

[9] Baris Erol, Sevgi Z. Gurbuz, and Moeness G. Amin. 2019. GAN-based synthetic radar micro-doppler augmentations for improved human activity recognition. In *Proceedings of the IEEE Radar Conference*. 1–5.

[10] Wei Feng, Wenjiang Huang, and Jinchang Ren. 2018. Class imbalance ensemble learning based on the margin theory. *MDPI Appl. Sci.* 8, 5 (May 2018), 815.

[11] Ruiyang Gao, Mi Zhang, Jie Zhang, Yang Li, Enze Yi, Dan Wu, Leye Wang, and Daqing Zhang. 2021. Towards position-independent sensing for gesture recognition with Wi-Fi. *Proc. ACM Interact. Mob. Wear. Ubiq. Technol.* 5, 2 (Jun. 2021), 1–28.

[12] Vicente García, Ramón Alberto Mollineda, and José Salvador Sánchez. 2009. Index of balanced accuracy: A performance measure for skewed class distributions. In *Proceedings of the Iberian Conference on Pattern Recognition and Image Analysis*. 441–448.

[13] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Proceedings of the Conference and Workshop on Neural Information Processing Systems (NIPS'14)*. 1–9.

[14] Huaping Guo, Hongbing Liu, Changan Wu, Weimei Zhi, Yan Xiao, and Wei She. 2016. Logistic discrimination based on G-mean and f-measure for imbalanced problem. *J. Intell. Fuzzy Syst.* 31, 3 (2016), 1155–1166.

[15] John R Hershey and Peder A Olsen. 2007. Approximating the kullback leibler divergence between gaussian mixture models. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'07)*. IV-317–IV-320.

[16] Arya Iranmehr, Hamed Masnadi-Shirazi, and Nuno Vasconcelos. 2019. Cost-sensitive support vector machines. *Neurocomputing* 343 (May 2019), 50–64.

[17] Hao Kong, Li Lu, Jiadi Yu, Yanmin Zhu, Feilong Tang, Yi-Chao Chen, Linghe Kong, and Feng Lyu. 2022. Push the limit of WiFi-based user authentication towards undefined gestures. In *Proceedings of the IEEE International Conference on Computer Communications (INFOCOM'22)*. 410–419.

[18] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, F. Pereira, C. J. Burges, L. Bottou, and K. Q. Weinberger (Eds.), Vol. 25. Lake Tahoe, USA.

[19] Oscar D. Lara and Miguel A. Labrador. 2012. A survey on human activity recognition using wearable sensors. *IEEE Commun. Surv. Tutor.* 15, 3 (Third Quarter 2012), 1192–1209.

[20] Guillaume Lemaître, Fernando Nogueira, and Christos K Aridas. 2017. Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. *J. Mach. Learn. Res.* 18, 1 (Jan. 2017), 559–563.

[21] Chenning Li, Manni Liu, and Zhichao Cao. 2020. WiHF: Enable user identified gesture recognition with WiFi. In *Proceedings of the IEEE International Conference on Computer Communications (INFOCOM'20)*. 586–595.

[22] Junnan Li, Qingsheng Zhu, Quanwang Wu, and Zhu Fan. 2021. A novel oversampling technique for class-imbalanced learning based on SMOTE and natural neighbors. *Inf. Sci.* 565 (2021), 438–455.

[23] Shuheng Li, Ranak Roy Chowdhury, Jingbo Shang, Rajesh K Gupta, and Dezhi Hong. 2021. UniTS: Short-time fourier inspired neural networks for sensory time series classification. In *Proceedings of the ACM Conference on Embedded Networked Sensor Systems (SenSys'21)*. 234–247.

[24] Youpeng Li, Xuyu Wang, and Lingling An. 2023. Hierarchical clustering-based personalized federated learning for robust and fair human activity recognition. *Proc. ACM Interact. Mob. Wear. Ubiq. Technol.* 7, 1 (2023), 1–38.

[25] Yadong Li, Dongheng Zhang, Jinbo Chen, Jinwei Wan, Dong Zhang, Yang Hu, Qibin Sun, and Yan Chen. 2021. Towards domain-independent and real-time gesture recognition using mmWave signal. arXiv:2111.06195. https://arxiv.org/abs/2111.06195

[26] Chi Lin, Pengfei Wang, Chuanying Ji, Mohammad S Obaidat, Lei Wang, Guowei Wu, and Qiang Zhang. 2023. A contactless authentication system based on WiFi CSI. *ACM Trans. Sens. Netw.* 19, 2 (2023), 1–20.

[27] Jian Liu, Hongbo Liu, Yingying Chen, Yan Wang, and Chen Wang. 2019. Wireless sensing for human activity: A survey. *IEEE Commun. Surv. Tutor.* 22, 3 (Third Quarter 2019), 1629–1645.

[28] Ziwei Liu, Zhongqi Miao, Xiaohang Zhan, Jiayun Wang, Boqing Gong, and Stella X Yu. 2019. Large-scale long-tailed recognition in an open world. In *Proceedings of the IEEE/CVF Computer Vision and Pattern Recognition Conference (CVPR'19)*. 2537–2546.

[29] Yongsen Ma, Gang Zhou, Shuangquan Wang, Hongyang Zhao, and Woosub Jung. 2018. SignFi: Sign language recognition using WiFi. *Proc. ACM Interact. Mob. Wear. Ubiq. Technol.* 2, 1 (Mar. 2018), 1–21.

[30] Giovanni Mariani, Florian Scheidegger, Roxana Istrate, Costas Bekas, and Cristiano Malossi. 2018. BAGAN: Data augmentation with balancing GAN. arXiv:1803.09655. Retrieved from https://arxiv.org/abs/1803.09655

[31] Mehdi Mirza and Simon Osindero. 2014. Conditional generative adversarial nets. arXiv:1411.1784. Retrieved from https://arxiv.org/abs/1411.1784

[32] Augustus Odena, Christopher Olah, and Jonathon Shlens. 2017. Conditional image synthesis with auxiliary classifier GANs. In *Proceedings of the International Conference on Machine Learning (ICML'17)*. 2642–2651.

[33] Sameera Palipana, Dariush Salami, Luis A Leiva, and Stephan Sigg. 2021. Pantomime: Mid-air gesture recognition with sparse millimeter-wave radar point clouds. *Proc. ACM Interact. Mob. Wear. Ubiq. Technol.* 5, 1 (Mar. 2021), 1–27.

[34] Maytus Piriyajitakonkij et al. 2020. SleepPoseNet: Multi-view learning for sleep postural transition recognition using UWB. *IEEE J. Biomed. Health Inf.* 25, 4 (Apr. 2020), 1305–1314.

[35] Akash Deep Singh, Sandeep Singh Sandha, Luis Garcia, and Mani Srivastava. 2019. Radhar: Human activity recognition from point clouds generated through a millimeter-wave radar. In *Proc. ACM Workshop on Millimeter-Wave and Terahertz Networks and Sensing Systems (mmNets'19)*. 51–56.

[36] Jie Wang, Liang Zhang, Changcheng Wang, Xiaorui Ma, Qinghua Gao, and Bin Lin. 2020. Device-free human gesture recognition with generative adversarial networks. *IEEE IoT J.* 7, 8 (Apr. 2020), 7678–7688.

[37] Q. Wang, B. Wu, P. Zhu, P. Li, and Q. Hu. 2020. ECA-Net: Efficient channel attention for deep convolutional neural networks. In *Proceedings of the IEEE/CVF Computer Vision and Pattern Recognition Conference (CVPR'20)*. 11534–11542.

[38] Xuyu Wang, Xiangyu Wang, and Shiwen Mao. 2018. RF sensing in the internet of things: A general deep learning framework. *IEEE Commun. Mag.* 56, 9 (2018), 62–67.

[39] Xuyu Wang, Chao Yang, and Shiwen Mao. 2020. On CSI-based vital sign monitoring using commodity WiFi. *ACM Trans. Comput. Healthc.* 1, 3 (2020), 1–27.

[40] Yanwen Wang and Yuanqing Zheng. 2018. Modeling RFID signal reflection for contact-free activity recognition. *Proc. ACM Interact. Mob. Wear. Ubiq. Technol.* 2, 4 (Dec. 2018), 1–22.

[41] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE Trans. Image Process.* 13, 4 (Apr. 2004), 600–612.

[42] Rui Xiao, Jianwei Liu, Jinsong Han, and Kui Ren. 2021. OneFi: One-shot recognition for unseen gesture via COTS WiFi. In *Proceedings of the ACM Conference on Embedded Networked Sensor Systems (SenSys'21)*. 206–219.

[43] Chao Yang, Lingxiao Wang, Xuyu Wang, and Shiwen Mao. 2022. Environment adaptive RFID based 3D human pose tracking with a meta-learning approach. *IEEE J. Radio Freq. Ident.* 6, 1 (Jan. 2022), 413–425.

[44] Chao Yang, Xuyu Wang, and Shiwen Mao. 2020. RFID-Pose: Vision-aided three-dimensional human pose estimation with radio-frequency identification. *IEEE Trans. Reliabil.* 70, 3 (Sep. 2020), 1218–1231.

[45] Chao Yang, Xuyu Wang, and Shiwen Mao. 2023. TARF: Technology-agnostic RF sensing for human activity recognition. *IEEE J. Biomed. Health Inf.* 27, 2 (Feb. 2023), 636–647. https://doi.org/10.1109/JBHI.2022.3175912

[46] Shuochao Yao et al. 2019. STFNets: Learning sensing signals from the time-frequency perspective with short-time fourier neural networks. In *Proceedings of the ACM International World Wide Web Conference (WWW'19)*. 2192–2202.

[47] Shuangjiao Zhai, Zhanyong Tang, Petteri Nurmi, Dingyi Fang, Xiaojiang Chen, and Zheng Wang. 2022. RISE: Robust wireless sensing using probabilistic and statistical assessments. In *Proceedings of the International Conference on Mobile Computing and Networking (MobiCom'21)*. 309–322.

[48] Jin Zhang, Fuxiang Wu, Bo Wei, Qieshi Zhang, Hui Huang, Syed W Shah, and Jun Cheng. 2020. Data augmentation and dense-LSTM for human activity recognition using WiFi signal. *IEEE IoT J.* 8, 6 (Mar. 2020), 4628–4641.

[49] Shujie Zhang, Tianyue Zheng, Zhe Chen, and Jun Luo. 2022. Can we obtain fine-grained heartbeat waveform via contact-free RF-sensing? In *Proceedings of the IEEE International Conference on Computer Communications (INFOCOM'22)*. 1759–1768.

[50] Yi Zhang, Zheng Yang, Guidong Zhang, Chenshu Wu, and Li Zhang. 2021. XGest: Enabling cross-label gesture recognition with RF signals. *ACM Trans. Sens. Netw.* 17, 4 (2021), 1–23.

[51] Peijun Zhao, Chris Xiaoxuan Lu, Bing Wang, Niki Trigoni, and Andrew Markham. 2023. CubeLearn: End-to-end learning for human motion recognition from raw mmWave radar signals. *IEEE IoT J.* 10, 12 (2023), 10236–10249.

[52] Tianyue Zheng, Zhe Chen, Shuya Ding, and Jun Luo. 2021. Enhancing RF sensing with deep learning: A layered approach. *IEEE Commun. Mag.* 59, 2 (Feb. 2021), 70–76.

[53] Yue Zheng, Yi Zhang, Kun Qian, Guidong Zhang, Yunhao Liu, Chenshu Wu, and Zheng Yang. 2019. Zero-effort cross-domain gesture recognition with Wi-Fi. In *Proceedings of the ACM International Conference on Mobile Systems, Applications, and Services (MobiSys'19)*. 313–325.

[54] Zhipeng Zhou, Feng Wang, Jihong Yu, Ju Ren, Zhi Wang, and Wei Gong. 2022. Target-oriented semi-supervised domain adaptation for WiFi-based HAR. In *Proceedings of the IEEE International Conference on Computer Communications (INFOCOM'22)*. 420–429.

[55] Xiaofeng Zhu, Jianye Yang, Chengyuan Zhang, and Shichao Zhang. 2019. Efficient utilization of missing data in cost-sensitive learning. *IEEE Trans. Knowl. Data Eng.* 33, 6 (2019), 2425–2436.

Received 5 February 2023; revised 6 June 2023; accepted 25 July 2023