IEICE TRANSACTIONS

on Communications

DOI:10.23919/transcom.2023EBP3102

This advance publication article will be replaced by the finalized version after proofreading.

A PUBLICATION OF THE COMMUNICATIONS SOCIETY

The Institute of Electronics, Information and Communication Engineers

Kikai-Shinko-Kaikan Bldg., 5-8, Shibakoen 3chome, Minato-ku, TOKYO, 105-0011 JAPAN

PAPER

SimpleViTFi: A Lightweight Vision Transformer Model for Wi-Fi-based Person Identification

Jichen BIAN[†,††], *Student Member*, **Min ZHENG**[†], **Hong LIU**[†], **Jiahui MAO**[†,††], **Hui LI**[†], *and* **Chong TAN**[†a)], *Nonmembers*

SUMMARY Wi-Fi-based person identification (PI) tasks are performed by analyzing the fluctuating characteristics of the Channel State Information (CSI) data to determine whether the person's identity is legitimate. This technology can be used for intrusion detection and keyless access to restricted areas. However, the related research rarely considers the restricted computing resources and the complexity of real-world environments, resulting in lacking practicality in some scenarios, such as intrusion detection tasks in remote substations without public network coverage. In this paper, we propose a novel neural network model named SimpleViTFi, a lightweight classification model based on Vision Transformer (ViT), which adds a downsampling mechanism, a distinctive patch embedding method and learnable positional embedding to the cropped ViT architecture. We employ the latest IEEE 802.11ac 80MHz CSI dataset provided by [1]. The CSI matrix is abstracted into a special "image" after pre-processing and fed into the trained SimpleViTFi for classification. The experimental results demonstrate that the proposed SimpleViTFi has lower computational resource overhead and better accuracy than traditional classification models, reflecting the robustness on LOS or NLOS CSI data generated by different Tx-Rx devices and acquired by different monitors.

key words: *Wi-Fi sensing, CSI, person identification, lightweight model, vision transformer*

1. INTRODUCTION

With the continuous evolution of Wi-Fi protocols [2], [3] and the exponential growth of Wi-Fi devices, people are no longer solely focused on using Wi-Fi for Internet access. Instead, there is an increasing demand for higher bandwidth, more reliable connections, and improved service quality to accommodate applications such as high-immersive gaming and remote healthcare [4]. This shift has led to the emergence of a more versatile and robust wireless communication infrastructure that not only provides seamless connectivity but also enables novel sensing and interaction capabilities. It is widely recognized that Wi-Fi sensing plays a crucial role in various tasks, including indoor activity recognition, object sensing, and localization [5], [6]. By leveraging the fine-grained channel variations captured in Wi-Fi CSI, researchers can extract meaningful features that correlate with real-world positions, actions, and states [7]. This capability paves the way for an array of novel prospects in the domain of pervasive and context-aware computing applications, including intelligent residential environments, assisted living arrangements, and advanced security systems [5], [8]. However, there exist challenges in achieving efficient Wi-Fi sensing in resource-constrained environments. For instance, remote substations in underdeveloped areas need to deploy the intrusion detection system due to their critical energy supply role and potential security risks. Conventional camera detection is difficult to illuminate at night and to guarantee dead-end coverage, not to mention the large demand for computing resources. Meanwhile, such substations often lack public network coverage because of the remote location, making it hard to access cloud servers for the deployment of highly resource-intensive detection applications [9], [10]. In such scenarios, the lightweight and effective Wi-Fi-based PI method is considered as a reliable alternative, which can operate with local, limited resources [6]. We aim to advance the state-of-the-art of Wi-Fi sensing at the edge and contribute to its broader applicability in challenging environments. This will ultimately enable the deployment of Wi-Fi sensing technologies in a wider range of real-world scenarios, thus improving the efficiency and safety of critical infrastructure management [5].

At present, a multitude of research employs Wi-Fi sensing technology for various tasks. [11] introduces Wisleep, a system that infers sleep duration using passively sensed smartphone network connections from Wi-Fi infrastructure, achieving comparable accuracy to client-side methods. An unavoidable limitation, though, is a reliance on users carrying devices, while current research trends are shifting towards device-free detection methods for greater convenience and user comfort. [12] proposes Temporal Unet, a deep convolutional neural network for sample-level action recognition in the Wi-Fi sensing domain, enabling precise action localization and real-time recognition. Nevertheless, this paper does not address potential issues related to computational complexity and generalizability across diverse environments. [13] presents FewSense, a few-shot learning-based Wi-Fi sensing system capable of recognizing novel classes in unseen domains using limited samples, achieving high accuracy on three public datasets (SignFi, Widar, and Wiar) and improving performance through collaborative sensing while limiting in the large model size, which may render it unsuitable for computationally constrained environments despite its effectiveness in cross-domain scenarios.

Despite a great deal of research being conducted, there is still a lack of studies on Wi-Fi sensing focusing on resource-constrained environments. In this paper, we pro-

[†]Shanghai Institute of Microsystem and Information Technology, Chinese Academy of Sciences, Shanghai, 200050, China

[††]University of Chinese Academy of Sciences, Beijing, 101408, China

a) E-mail: chong.tan@mail.sim.ac.cn

pose a novel neural network model named SimpleViTFi based on ViT. This model performs well on person identification tasks using CSI data generated from Wi-Fi devices. Our developments are inspired by works in [8], [14]. The developments can be concretely described as follows:

(1) Drawing inspiration from the ViT model in the field of Computer Vision (CV), we propose a lightweight ViT model with distinctive patch segmentation, downsampling operation, reduced number of layers, and efficient feature extraction capabilities, termed as SimpleViTFi, specifically designed for PI tasks in the Wi-Fi sensing domain under resource-constrained scenarios.
(2) We conduct a comparative analysis of the impact of two types of position encoding methods - the sin-cos method and learnable embedding - on PI. The results show that the learnable embedding method yields superior performance, and we delve into a discussion attempting to analyze the possible explanations for this outcome.
(3) We benchmark SimpleViTFi against several popular models, including LeNet, ResNet18, and GRU. SimpleViTFi significantly outperforms these models on Wi-Fi-based PI tasks. Furthermore, we introduce an incremental learning approach to further enhance the performance and efficiency of SimpleViTFi, which requires a little extra time and data to achieve robust performance across different CSI datasets generated by various Wi-Fi devices.

The structure of this paper unfolds as follows: Section 2 delves into a comprehensive discussion on related works. Section 3 provides the detail of the proposed SimpleViTFi. Section 4 shows the experimental setup and comparisons of the results with existing works. Section 5 concludes this paper and provides recommendations for some future research topics.

2. Related Works

In this section, we survey the existing literature on Wi-Fi sensing using CSI data. Research work in the Wi-Fi sensing field bifurcates into two main directions: fundamental model research and application-oriented research. From a methodological perspective, there exists a gradual shift in focus from traditional statistical modeling methods to artificial intelligence (AI) methods.

In terms of fundamental model research, Yang et al. [7] propose an automatic Wi-Fi human sensing learning framework called AutoFi, which can achieve automatic Wi-Fi human sensing with minimal manual annotation. AutoFi can train a robust model from low-quality CSI samples, making it easier to use Wi-Fi sensing technology in new environments. The paper also analyzes the main gaps between existing learning-based methods and practical Wi-Fi sensing, proposing a novel self-supervised learning framework and a new geometric structure loss function to enhance the model's transferability. Extensive experiments are conducted on public datasets and real-world scenarios, demonstrating the high accuracy and robustness of the AutoFi method in automatic Wi-Fi human sensing. In another study, Hernandez and Bulut [15] present WiFederated, a federated learning approach for training machine learning models for Wi-Fi sensing tasks. This method allows for parallel training at the edge, enabling devices to collaboratively learn and share location-independent physical behavior features. The authors demonstrate that their method diminishes the necessity for extensive data collection at each new location, offering a solution that is more accurate and time-efficient compared to both transfer learning and adversarial learning solutions. Liu et al. [16] propose a deep learning-based Wi-Fi sensing approach using a CNN-BiLSTM architecture to identify vigorous activities. This architecture can simultaneously extract sufficient spatiotemporal features of action data and establish the mapping relationship between actions and CSI streams, thereby improving activity recognition accuracy.

In terms of application-oriented research, several mature systems have been developed, showcasing the unique charm of Wi-Fi sensing in various fields. Tong et al. [17] propose FreeSense, a combination of Principal Component Analysis (PCA), Discrete Wavelet Transform (DWT) and Dynamic Time Warping (DTW) techniques, using for CSI waveform-based human identification. The identification accuracy of FreeSense ranges from 94.5% to 88.9% when the number of users changes from 2 to 6. Lin et al. [18] represent WiTL, a contactless authentication system based on Wi-Fi CSI. It is devised using a transfer learning technology, in combination with ResNet and the adversarial network, to extract activity features and learn environment-independent representations. WiTL achieves a great accuracy over 93% and 97% in multi-scenes and multi-activities identity recognition, respectively.

In spite of a few existing studies of Wi-Fi-based PI tasks, they rarely consider the feasibility in resource-constrained environments. Therefore, we would like to combine the latest research based on Wi-Fi sensing and AI methods to make innovations in resource-constrained PI tasks.

3. Methodology

3.1 Channel State Information

Channel State Information (CSI) [19] is a critical component in Wi-Fi sensing systems. It represents the combined effects of the wireless channel's propagation properties, including path loss, shadowing, and multipath fading, which are affected by the environment and the presence of objects or people. CSI can be modeled as channel impulse response (CIR) in the frequency domain as

$$h(\tau) = \sum_{l=1}^{L} \alpha_l e^{j\phi_l} \delta(\tau - \tau_l), \tag{1}$$

where α_l and ϕ_l respectively represent the amplitude and phase of the l_th multipath component, τ_l is the time delay, L indicates the total number of multipath components, and $\delta(\tau)$ denotes the Dirac delta function. CSI has been widely

used in Wi-Fi sensing research to exploit the rich information it contains about the surrounding environment and human activities.

CSI can be obtained from commodity Wi-Fi devices. When a transmitter transmits a signal x, it is received by the receiver as $y = Hx + \eta$, where η represents environmental noise and H represents the CSI complex-valued matrix. Each element in the matrix corresponds to the channel gain between a specific transmitter-receiver antenna pair in a MIMO system. The matrix's dimensions depend on the number of transmitting and receiving antennas. In addition, the CSI matrix is also influenced by the number of Orthogonal Frequency Division Multiplexing (OFDM) subcarriers. The more subcarriers, the finer the frequency resolution, which allows for a more accurate representation of the channel characteristics [20].

The CSI matrix H for a system with N transmitting antennas and M receiving antennas can be represented as:

$$\mathrm{CSI}_{N\times M} = \begin{bmatrix} h_{11} & h_{12} & \dots & h_{1M} \\ h_{21} & h_{22} & \dots & h_{2M} \\ \vdots & \vdots & \ddots & \vdots \\ h_{N1} & h_{N2} & \dots & h_{NM} \end{bmatrix} \tag{2}$$

In this representation, h_{ij} is a complex vector that represents the channel gain between the i-th transmitting antenna and the j-th receiving antenna. The amplitude and phase of each h_{ij} can be calculated as follows:

$$\mathrm{Amp}(h_{ij}) = |h_{ij}| = \sqrt{\mathrm{Re}(h_{ij})^2 + \mathrm{Im}(h_{ij})^2} \tag{3}$$

$$\mathrm{Pha}(h_{ij}) = \angle h_{ij} = \arctan\left(\frac{\mathrm{Im}(h_{ij})}{\mathrm{Re}(h_{ij})}\right) \tag{4}$$

3.2 Vision Transformer

Vision Transformer (ViT) [21], [22] has emerged as a powerful and flexible approach for solving various CV tasks, inspired by the success of Transformers in natural language processing (NLP). ViT is a type of neural network architecture that can process images by dividing them into non-overlapping patches and treating these patches as a sequence of tokens, similar to how Transformers process texts.

The core component of ViT is the self-attention mechanism, which allows the model to learn long-range dependencies between different parts of the image. This mechanism enables ViT to capture both local and global contextual information and adaptively focus on relevant regions in the image.

ViT has demonstrated state-of-the-art performance on a wide range of CV tasks, such as image classification, object detection, and semantic segmentation [23], outperforming traditional convolutional neural networks (CNNs). The flexibility and expressiveness of ViT make them a promising approach for various CV tasks, including those that require fine-grained visual understanding and adaptability to different input modalities [24].

In this paper, we treat the CSI matrix as a multi-channel "image" and attempt to address the CSI-based PI tasks with ViT. From our perspective, CSI images differ from traditional RGB images in two aspects:

(1) The weights in CSI images are evenly distributed across all pixels, unlike conventional images that typically have a focal point and a background. The global receptive field of ViT can better capture the features of CSI images due to this uniform distribution.
(2) CSI images have a temporal dimension, necessitating a focus on the relationships and changes along this dimension. ViT, with its unique sensitivity to positional relationships, is well-suited to this task.

Therefore, this paper aims to explore the potential of ViT in the realm of CSI-based classification, hoping to uncover the unique capabilities of this technology in handling such tasks.

3.3 SimpleViTFi

As shown in Fig 1, we propose SimpleViTFi, which is designed for processing CSI images with a focus on efficient feature extraction and classification. SimpleViTFi is inspired by the ViT and incorporates several key components with data flow as shown by the bold red arrows. SimpleViTFi comprises the following main components:

Patch Embedding: The input CSI matrix $\mathbf{X} \in \mathbb{R}^{B\times A\times S\times T}$ is first downsampled and divided into non-overlapping patches along the temporal dimension, where the dimensions represent the number of antennas(A), subcarriers(S), and the time sequence(T) respectively. Then the patches are linearly embedded into a higher-dimensional feature space. A Layer Normalization operation is applied to the embedded patches. Unlike traditional image patch segmentation methods, we do not partition the data along the subcarrier dimension, as we prefer the model to focus on the temporal dimension.

Position Encoding: Learnable positional embeddings $\mathbf{P} \in \mathbb{R}^{S\times T}$ are added to the patch embeddings to capture the spatial relationships between the patches in SimpleViTFi. There are two main types of positional embeddings:

(1) Fixed Positional Embbdings follow the original method in [25], which are initialized with a sinusoidal function.
(2) Learnable Positional Embeddings are initialized randomly and then updated through backpropagation during the training process.

The CSI dataset involves complicated spatial and temporal relationships across different antennas and subcarriers. This multi-dimensional complexity could pose challenges to traditional sinusoidal position encodings such as the sin-cos method used in the Transformer model, which provides a fixed encoding based on the position of data points in the sequence. In contrast, learnable positional embeddings, added to the patch embeddings to capture the spatial relationships between time sequences, offer a more flexible approach. By allowing the model to learn the position embeddings from the data itself, it could enable the discovery of more intricate or subtle patterns in the sequence order, thereby improving

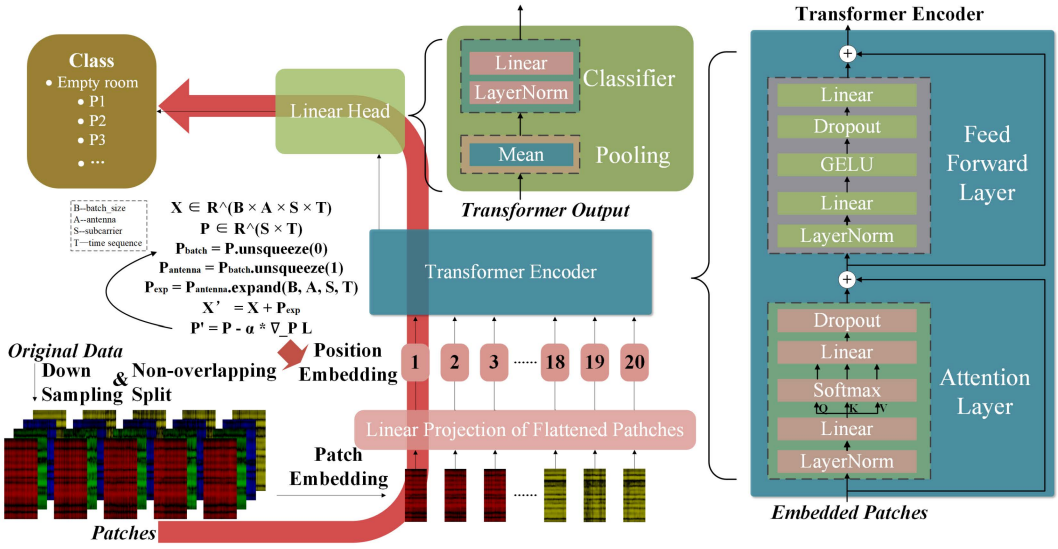
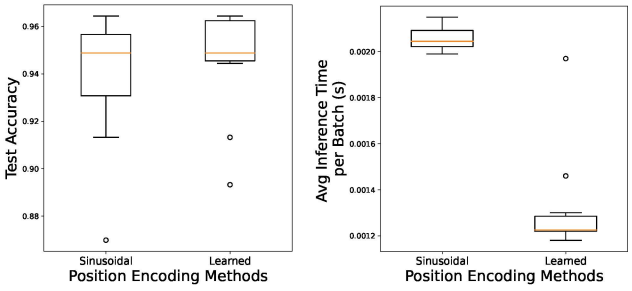


Fig. 1: SimpleViTFi Model

its ability to identify individuals.

We compare two methods mentioned above: the sin-cos method and learnable embedding. Fig 2a shows that the learnable embedding achieves a more consistent high rate of accuracy within 20 replicate experiments, as it enables the model to adapt to the specific patterns present in the CSI data. Although using learnable embedding increases the number of parameters and requires additional optimization during training, it results in a shorter inference time compared to the other as shown in Fig 2b. This is attributable to the learnable embedding being computed in parallel, whereas the sin-cos method requires sequential computation. The combined embeddings can be represented as $\mathbf{X}' = \mathbf{X} + \mathbf{P}\text{exp}$, where $\mathbf{P}\text{exp} \in \mathbb{R}^{B\times A\times S\times T}$ is the expanded version of $\mathbf{P}$.



(a) Test Accuracy with Two Position Methods

(b) Inference Time with Two Position Methods

Fig. 2: Test Accuracy and Inference Time of Two Position Methods

Transformer Encoder: The combined patch and positional embeddings are fed into a Transformer encoder, which consists of multiple layers of multi-head self-attention and feedforward neural networks. In the experiments that follow, we employ 2 layers of self-attention and feedforward networks.

Pooling: Following the Transformer encoder, a global average pooling operation is performed to aggregate the features across the sequence dimension. This operation reduces the dimensionality of the output and prepares it for the classification head. The pooled features can be represented as $\mathbf{Z} = \text{mean}(\mathbf{X}', 1)$.

Classifier Head: The pooled features $\mathbf{Z}$ are then passed through a LayerNormalization layer, which can be represented as:

$$\mathbf{Z}norm = \frac{\mathbf{Z} - \text{E}[\mathbf{Z}]}{\sqrt{\text{Var}[\mathbf{Z}] + \epsilon}}, \tag{5}$$

where E[.] is the expectation operation, Var[.] is the variance operation, and ϵ is a small constant for numerical stability. The normalized features $\mathbf{Z}norm$ are subsequently processed by a Linear layer that maps the features to the desired number of output classes. This can be represented as:

$$\mathbf{Y} = \mathbf{W}\left(\frac{\mathbf{Z} - \text{E}[\mathbf{Z}]}{\sqrt{\text{Var}[\mathbf{Z}] + \epsilon}}\right) + \mathbf{b}, \tag{6}$$

where W is the weight matrix and b is the bias vector of the Linear layer.

The SimpleViTFi architecture is designed to be lightweight and efficient while maintaining high performance on the task of processing and classifying CSI matrices. By leveraging the strengths of both Vision Transformers and learned positional embeddings, the SimpleViTFi model demonstrates the robustness and adaptability to various CSI data patterns.

4. Experiment

4.1 80MHz CSI Dataset of IEEE 802.11ac

The datasets mentioned in [1], [14] consisting of three types

of datasets applicable to activity recognition(AR), person identification(PI), and people counting(PC), are produced by the University of Padova. Our focus is on the subset dedicated to PI in this paper.

Dataset Experiment Setup: As shown in Fig 3, the experiments are set within a meeting room. Two pairs of devices are strategically positioned. Specifically:

- Tx1 communicates with Rx1, establishing a line-of-sight (LOS) condition.
- Tx2 communicates with Rx2, resulting in a non-line-of-sight (NLOS) condition.

Additionally, two monitors, M1 and M2, are positioned to sniff and calculate the CSI data from both communication links. Consequently, each monitor stores two distinct sets of CSI data, named PI-1 – PI-4 shown in Table 1.

CSI Collection Method: An iPerf3 session is established between each pair of Tx and Rx, transmitting at a consistent rate of 173 packets per second. This rate corresponds to time intervals of approximately 6ms between each packet. The monitors configure the Nexmon-CSI extraction tool [26] to sniff packets continuously. The dataset involves 10 participants, each of whom moves individually and randomly within the colored areas in Fig 3.

Table 1: Measurement Conditions of the Dataset

	PI-1	PI-2	PI-3	PI-4
w×l×h	7m×7.5m×3.5m			
obst.	✗	✓	✗	✓
devices pos.	M1-Tx1-Rx1	M1-Tx2-Rx2	M2-Tx1-Rx1	M2-Tx2-Rx2
Tx	Netgear	Netgear	Netgear	Netgear
Rx	Netgear	TP-Link	Netgear	TP-Link
furniture	7 desks, chairs			

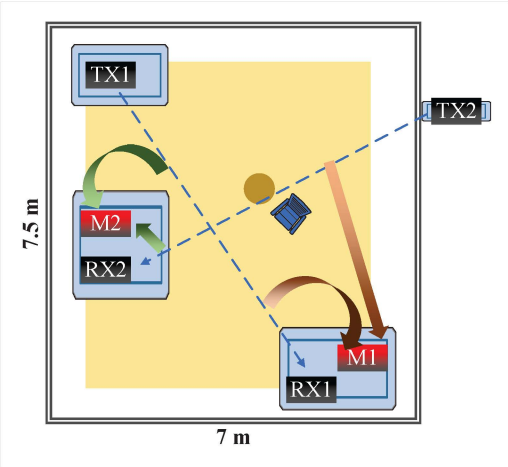


Fig. 3: Devices and Users' Positions in the Meetingroom

4.2 Data Preprocessing

Taking PI2_p03 as an example, this file represents the CSI data of Participant-3 created by Tx2 and Rx2, which is monitored by M1 in NLOS condition. It is a complex matrix of size 187264×256, where 256 represents the number of OFDM subcarriers under the 80MHz bandwidth, and 187264 represents the CSI indices of 46816 packets obtained separately by the four antennas. We preprocess this data file as follows:

(1) Load raw data and apply a Fast Fourier Transform shift operation.
(2) Remove invalid subcarriers and zero-sum rows from the CSI matrix, retaining 242 subcarriers.
(3) Calculate the number of complete groups of 4-antenna CSI data.
(4) Due to hardware artifact, negate the data from the 64th column onwards in each group.
(5) Convert the original complex values to amplitude values by taking the modulus.
(6) Divide the matrix into submatrices of size (4, 242, 2000) using a boundary of 2000 packets, facilitating subsequent analysis.

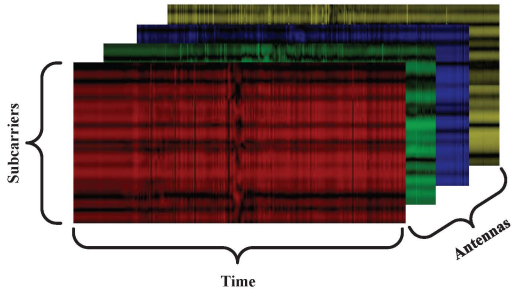


Fig. 4: CSI Amplitude Matrix

4.3 Experiment Setup

To demonstrate the effectiveness of the proposed method, we use the dataset mentioned in 4.1, and implement the SimpleViTFi based on Pytorch. Then, we conduct extensive experiments to evaluate the performance of SimpleViTFi concerning classification accuracy, model parameters and inference time of PI task.

System Design: The edge server in resource-constrained scenarios is simulated by the PC equipped with one NVIDIA RTX 3060 GPU. To fully evaluate the performance of SimpleViTFi and the others, we attempt to set up multiple experiments comprising different data sets. Four sets of experiments are set up as shown in Table 2. Specifically:

(1) **Experiment 1**: Utilizing $\frac{2}{3}$ of the PI-1 dataset as the training set and the remaining $\frac{1}{3}$ as the test set, this experiment aims to validate the model's classification ability in handling CSI data generated from LOS condition.

(2) **Experiment 2**: By employing $\frac{2}{3}$ of the PI-4 dataset for training and the rest for testing, this experiment is designed to assess the model's classification ability with CSI data stemming from NLOS condition.

(3) **Experiment 3**: This experiment combines $\frac{2}{3}$ of the PI-1 dataset with $\frac{1}{3}$ of the PI-3 dataset to form the training set, while the remaining data serves as the test set. Both PI-1 and PI-3 generate CSI data using Tx1 and Rx1 communication link but utilize different monitors. The primary objective is to evaluate the model's robustness to variations in devices' locations.

(4) **Experiment 4**: Incorporating a mixed dataset from PI-1 to PI-4, with $\frac{2}{3}$ used for training and the remainder for testing, this experiment seeks to gauge the model's resilience under the complexities of different devices and different monitors.

Table 2: Experiment Setup

Experiment Index	TrainSet	TestSet	Method
1	PI-1(2/3)	PI-1(1/3)	LeNet ResNet18 GRU SimpleViTFi
2	PI-4(2/3)	PI-4(1/3)	
3	PI-1(2/3)	PI-1(1/3)	
	PI-3(1/3)	PI-3(2/3)	
4	PI-1(2/3)	PI-1(1/3)	
	PI-2(2/3)	PI-2(1/3)	
	PI-3(2/3)	PI-3(1/3)	
	PI-4(2/3)	PI-4(1/3)	

Network Implementation: The network design has been shown in Table 3. Note that Transformer Encoder is a sequence of 2 attention and feed-forward layers. The attention layer uses the scaled dot-product attention mechanism with 8 heads, and the feed-forward layer is a two-layer fully connected network with a hidden dimension of 2048 and a GELU activation function in between. The model is trained with the Adam optimizer with a learning rate of 0.0001 and a weight decay of 0.1. The loss function used is CrossEntropyLoss. The model employs an early stopping mechanism during training, which halts the training process if there is no improvement in validation loss for 8 consecutive epochs, preventing overfitting and ensuring better generalization.

Criterion: In our experiments, we evaluate and compare the models based on three key metrics: the number of training parameters, inference time, and identification accuracy. The identification accuracy is denoted as the ratio of true predicted samples and all testing samples.

Baselines: We compare our method with three traditional methods. LeNet, as one of the earliest convolutional neural networks, has made significant contributions to the field of image classification, setting the foundation for future advancements [27]. ResNet18, with its innovative residual learning framework, has further improved the performance of deep neural networks in image classification tasks, notably reducing the training error [28]. On the other hand, GRU (Gated Recurrent Unit) has shown exceptional performance in time series prediction due to its efficient gating mechanisms, which handle the vanishing gradient problem

Table 3: Network Design of SimpleViTFi

Layer_Index	Components	details
input	CSI_amp: $4 \times 242 \times 2000$ (antenna pairs × subcarriers × time sequence)	
1	Patch Embedding	1) downsampling: $4 \times 121 \times 500$ 2) 100 patches: $100 \times 4 \times 121 \times 5$
2	Position Encoding	learnable embedding
3	Transformer Encoder	dim: 64 depth: 2 heads: 8 mlp_dim: 2048 dropout_rate: 0.3 learning_rate: 0.0001 weight_decay: 0.1 loss_function: CrossEntropyLoss
4	Pooling	average pooling
5	Classifier Head	$\mathbf{Y} = \mathbf{W}\left(\frac{\mathbf{Z}-\mathrm{E}[\mathbf{Z}]}{\sqrt{\mathrm{Var}[\mathbf{Z}]+\epsilon}}\right) + \mathbf{b}$
output	Classificaition Results	

and allow for long-term dependencies [29]. In light of our approach where we interpret the Channel State Information (CSI) matrix as an image, and considering the substantial temporal correlations this 'image' embodies, we deem it appropriate to draw comparisons with the aforementioned methods.

4.4 Evaluation

The proposed SimpleViTFi is compared with baselines. Fig 5 illustrates the efficiency of SimpleViTFi in comparison to the others. Notably, SimpleViTFi demonstrates the shortest average inference time clocking in at 1.338 ms and requires the least number of parameters with a total of 1,079,923, which makes it consume the fewest computational complexity and memory usage with high efficiency for real-time tasks.

Following this, we examine the performance of SimpleViTFi on PI-1 (**Experiment 1**). In addition to the amplitude-based results shown in Fig 6, we also incorporate phase-based results shown in Fig 7. However, the phase-based results are not as anticipated. For all four models, the accuracy barely surpasses 25%, indicating that the models are virtually non-functional with the phase value. We believe that the potential reasons for this could be the inherent instability and sensitivity of phase to environment. Under complex multipath effects, the phase undergoes multiple cumulative changes, making it highly unstable. This heightened sensitivity can lead the model to overfit, making it challenging to capture essential features.

Returning to the amplitude-based results, as presented in Fig 6, SimpleViTFi outperforms the others, achieving the highest accuracy on the test set. The box plot visualizes the range and distribution of accuracy scores achieved by Sim-

pleViTFi and the others across multiple runs. The central line in the box plot represents the median accuracy, which for SimpleViTFi is an impressive 0.9566, about at least 10% higher than the others such as 0.8525 for ResNet18. The box itself spans from the first quartile (Q1) to the third quartile (Q3), representing the interquartile range (IQR). For SimpleViTFi, Q1 is 0.91037 and Q3 is 0.9566. This range captures the middle 50% of accuracy scores, providing a sense of the model's consistency. This consistency, coupled with the high median accuracy, underscores the robustness of SimpleViTFi, indicating that it consistently delivers high performance under various conditions.

In **Experiment 2** shown in Fig 8, similar trends are observed. The two experiments utilize CSI data generated from two distinct sets of devices. After training on their respective train sets, the model achieved commendable results on their test sets, with classification accuracies exceeding 95%. This indicates that SimpleViTFi is adept at adapting to both LOS and NLOS scenarios. Furthermore, the results from the NLOS condition in **Experiment 2** even surpass those from the LOS condition in **Experiment 1**. This suggests that the model might be benefiting from the distinct noise characteristics introduced by different devices.

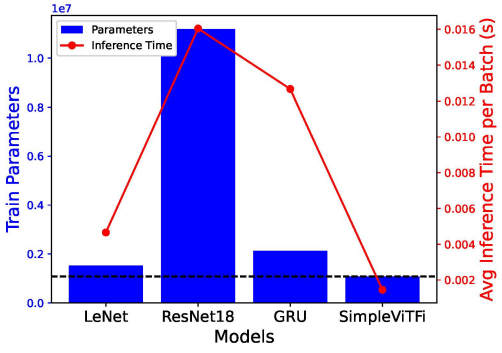


Fig. 5: Train Parameters and Inference Time per Batch

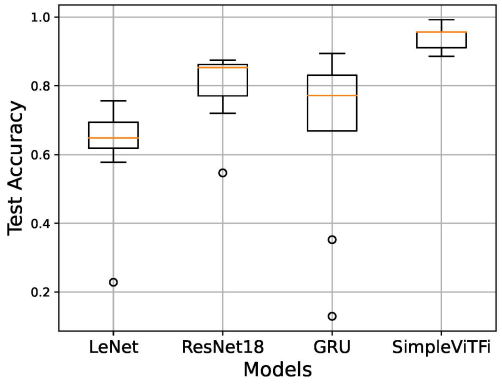


Fig. 6: Test Accuracy of **Experiment 1** with amplitude values. TrainSet and TestSet consist of PI-1.

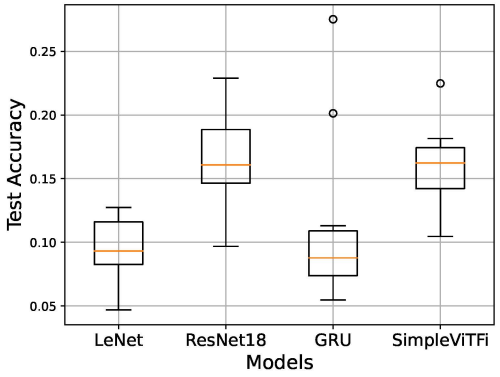


Fig. 7: Test Accuracy of **Experiment 1** with phase values. TrainSet and TestSet consist of PI-1.

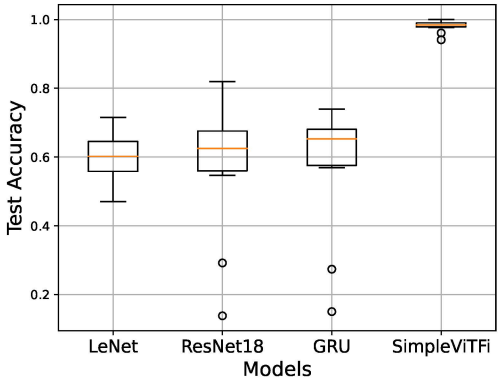


Fig. 8: Test Accuracy of **Experiment 2** with amplitude values. TrainSet and TestSet consist of PI-4.

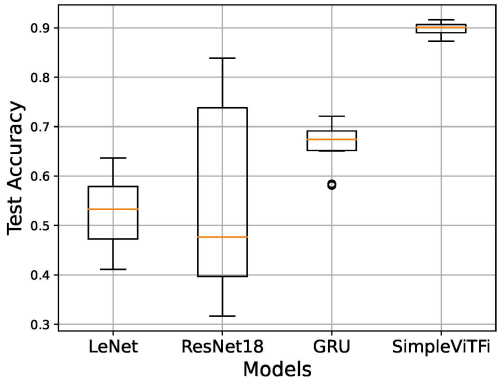


Fig. 9: Test Accuracy of **Experiment 3** with amplitude values. TrainSet and TestSet consist of PI-1 & PI-3.

We get similar results through **Experiment 3** and **4**. Through analyzing the box plots from Fig 6 to Fig 10, it is obvious that SimpleViTFi not only gets a high median accuracy but also demonstrates consistent performance, as indicated by the relatively small IQR, either on individual or mixed data sets generated by different devices or acquired by different monitors.

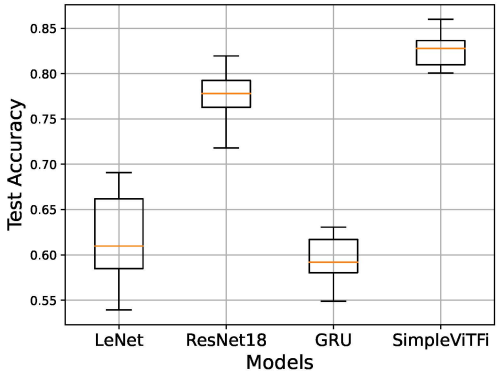


Fig. 10: Test Accuracy of **Experiment 4** with amplitude values. TrainSet and TestSet consist of PI-1 & PI-2 & PI-3 & PI-4.

In conclusion, our experiments showcase the superior performance of the SimpleViTFi model in terms of both user identity recognition accuracy and inference time. By outperforming traditional methods, the SimpleViTFi model demonstrates its robustness and adaptability to various CSI data patterns.

4.5 Insights and Analysis

In the preceding subsections, we detail the architecture, Implementation, and evaluation of SimpleViTFi. Although the quantitative results indicate the model's efficacy, it is essential to dive deeper into the underlying mechanisms that contribute to its performance. In this subsection, we try to elucidate some of the key factors that are pivotal for the observed results.

(1) **Model Architecture**: SimpleViTFi employs a ViT-based architecture, which fundamentally differs from traditional convolutional (such as LeNet and ResNet18) and recurrent (such as GRU) neural networks. SimpleViTFi utilizes self-attention mechanisms to process input data. The self-attention mechanism is computationally expressed as:

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}}\right)\mathbf{V}, \qquad (7)$$

where $\mathbf{Q}$, $\mathbf{K}$, and $\mathbf{V}$ are the query, key, and value matrices, respectively, and d_k is the dimension of the key. The self-attention mechanism allows each element in the input sequence to focus on other parts, governed by the weight calculated in the softmax term.
The self-attention mechanism's ability to weigh and capture relationships between different parts of the input is particularly crucial for tasks involving WiFi CSI. In the context of CSI "images" classification, these relationships can be both spatial, as in different antenna pairs, and temporal, as in different time slots. Therefore, the self-attention mechanism, defined by the formula above, enables SimpleViTFi to capture these complex relationships efficiently.
On one hand, convolutional models struggle to capture the long or short-range dependencies inherent in time series data. On the other hand, while GRU can capture these temporal features, it computes in a time-step manner. In contrast, the self-attention mechanism stands out with its ability to address these challenges, offering both flexibility and parallelized computation. This makes SimpleViTFi highly effective and efficient in handling tasks that involve both spatial and sequential data.

(2) **Feature Representation Capability**: In traditional CNN architectures, the receptive field is generally localized, focusing primarily on capturing local features such as edges and textures. In contrast, SimpleViTFi leverages self-attention mechanisms to offer a dynamic receptive field, which allows the model to adaptively adjust its focus and capture features at various scales and complexities. The dynamic nature of its receptive field enables SimpleViTFi to integrate both local and global information more effectively, thereby providing an extra layer of flexibility and power in representing features.

(3) **Training and Implementation Efficiency**: A significant advantage of SimpleViTFi lies in its efficiency. By utilizing only two transformer layers, the model inherently has fewer parameters as shown in Fig 5. This streamlined architecture not only expedites the training process but also ensures a swift inference time. Furthermore, the inherent parallel computation capability of the architecture further boosts the inference speed. As a result, SimpleViTFi boasts the shortest inference time among the four models, making it highly suitable for real-time applications.

(4) **Robustness to Noise and Deformation**: SimpleViTFi incorporates dropout layers in both the FeedForward and Attention modules. Dropout is a regularization technique that helps prevent overfitting, especially when the model might be exposed to sharp noise features in the data. Meanwhile, self-attention mechanism offers a more adaptive response to noise compared to other methods. Furthermore, the parallel processing capability ensures that SimpleViTFi remains resilient even when faced with temporal distortions in the data.

4.6 Incremental Learning

Based on the SimpleViTFi model trained in **Experiment 3**, we implement incremental learning [30]–[32] by training with a small amount of data from PI-4. As presented in Fig 11, the loss curve of the incremental learning model converges faster than the normal one. Meanwhile, the accuracy of the incremental learning model is higher under the same training conditions.

5. Conclusion

In this paper, we introduce a novel Wi-Fi sensing method,

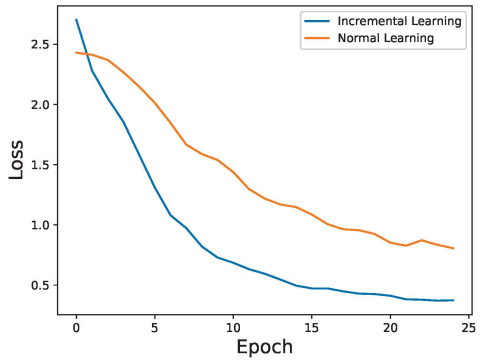


Fig. 11: Loss Curve of Incremental SimpleViTFi and Normal SimpleViTFi

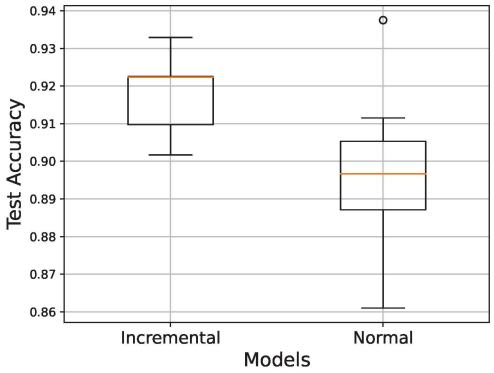


Fig. 12: Test Accuracy of Incremental SimpleViTFi and Normal SimpleViTFi

SimpleViTFi, designed for Wi-Fi-based PI in cross-device sensing scenarios. To address the limitations of existing algorithms, we develop a lightweight neural network model based on ViT with learnable embedding. The original CSI data are generated by 2 pairs of Netgear and TP-Link Wi-Fi devices, which enable a single antenna to enforce the communication over a single spatial stream. The packets transmitted over-the-air by the Tx are monitored by 2 Asus routers equipped with 4 antennas and then form 4 folders containing both LOS and NLOS scenarios. Subsequently, we train the proposed SimpleViTFi under 4 experimental conditions, utilizing data generated by different devices or acquired by different monitors. Extensive experiments demonstrate that SimpleViTFi achieves state-of-the-art performance in test accuracy, inference time and model parameters compared to baseline methods (LeNet, ResNet18 and GRU). Finally, we experiment with incremental learning to obtain a new model at a low cost. Here, a SimpleViTFi model initially trained on one set of devices is subjected to incremental training on another set of devices with a small amount of additional data. The results show that better accuracy and faster convergence are gained compared to training directly with data from another set of devices.

In the future, we have several avenues of exploration to further enhance our research. Firstly, we plan to propose a new method of position encoding that is better adapted to the CSI-based classification. Our experiments have underscored the significant impact of this aspect on the results. Furthermore, we aim to delve deeper into the potential of utilizing various CSI parameters, such as phase values, Doppler shifts and AoA, to improve the model's performance. In addition, we intend to test our model on Wi-Fi devices based on OpenWrt and then conduct pilot tasks in substations within the State Grid of China. By pursuing these avenues, we hope to further refine our model and broaden its applicability, ultimately contributing to the advancement of Wi-Fi sensing technologies.

Acknowledgment

This research was supported by the National Key R&D Program of China (No. 2020YFB2103300) and Science and Technology Project of State Grid Corporation of China (5108-202218280A-2-201-XG). The authors would like to thank the team behind [1], [14] for generously providing the dataset and assistance.

References

[1] F. Meneghello, N. Dal Fabbro, D. Garlisi, I. Tinnirello, and M. Rossi, "A CSI Dataset for Wireless Human Sensing on 80 MHz Wi-Fi Channels," IEEE Communications Magazine, 2023.

[2] "Ieee standard for information technology–telecommunications and information exchange between systems - local and metropolitan area networks–specific requirements - part 11: Wireless lan medium access control (mac) and physical layer (phy) specifications - redline," IEEE Std 802.11-2020 (Revision of IEEE Std 802.11-2016) - Redline, pp.1–7524, 2021.

[3] F. Restuccia, "Ieee 802.11bf: Toward ubiquitous wi-fi sensing," 2021.

[4] W. Jiang, H. Xue, C. Miao, S. Wang, S. Lin, C. Tian, S. Murali, H. Hu, Z. Sun, and L. Su, "Towards 3d human pose construction using wifi," pp.295–308, Assoc Comp Machinery, 2020. 26th Annual International Conference on Mobile Computing and Networking (MobiCom), ELECTR NETWORK, SEP 21-25, 2020.

[5] D. Wu, Y. Zeng, F. Zhang, and D. Zhang, "Wifi csi-based device-free sensing: from fresnel zone model to csi-ratio model," CCF TRANSACTIONS ON PERVASIVE COMPUTING AND INTERACTION, vol.4, no.1, pp.88–102, MAR 2022.

[6] K. Niu, X. Wang, F. Zhang, R. Zheng, Z. Yao, and D. Zhang, "Rethinking doppler effect for accurate velocity estimation with commodity wifi devices," IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS, vol.40, no.7, pp.2164–2178, JUL 2022.

[7] J. Yang, X. Chen, H. Zou, D. Wang, and L. Xie, "Autofi: Towards automatic wifi human sensing via geometric self-supervised learning," IEEE Internet of Things Journal, pp.1–1, 2022.

[8] J. Yang, X. Chen, H. Zou, D. Wang, Q. Xu, and L. Xie, "Efficientfi: Toward large-scale lightweight wifi sensing via csi compression," IEEE Internet of Things Journal, vol.9, no.15, pp.13086–13095, 2022.

[9] Y. Yang, K. McLaughlin, L. Gao, S. Sezer, Y. Yuan, and Y. Gong, "Intrusion detection system for iec 61850 based smart substations," 2016 IEEE Power and Energy Society General Meeting (PESGM), pp.1–5, 2016.

[10] S. Qiao, Q. Zheng, W. Li, S. Yang, and H. Zhang, "Intrusion detection of intelligent substation video surveillance based on average background interframe difference method," 2023 IEEE 2nd International Conference on Electrical Engineering, Big Data and Algorithms

(EEBDA), pp.1023–1026, 2023.
[11] P.M. Mammen, C. Zakaria, T. Molom-Ochir, A. Trivedi, P. Shenoy, and R. Balan, "Wisleep: Inferring sleep duration at scale using passive wifi sensing," 2022.
[12] F. Wang, Y. Song, J. Zhang, J. Han, and D. Huang, "Temporal unet: Sample level human action recognition using wifi," 2019.
[13] G. Yin, J. Zhang, G. Shen, and Y. Chen, "Fewsense, towards a scalable and cross-domain wi-fi sensing system using few-shot learning," IEEE Transactions on Mobile Computing, 2022.
[14] F. Meneghello, D. Garlisi, N.D. Fabbro, I. Tinnirello, and M. Rossi, "Sharp: Environment and person independent activity recognition with commodity ieee 802.11 access points," IEEE Transactions on Mobile Computing, pp.1–16, 2022.
[15] S.M. Hernandez and E. Bulut, "Wifederated: Scalable wifi sensing using edge-based federated learning," IEEE Internet of Things Journal, vol.9, no.14, pp.12628–12640, 2022.
[16] Y. Liu, S. Li, J. Yu, A. Dong, L. Zhang, C. Zhang, and Y. Cao, "Wifi sensing for drastic activity recognition with cnn-bilstm architecture," 2022 IEEE International Conference on Systems, Man, and Cybernetics (SMC), IEEE.
[17] T. Xin, B. Guo, Z. Wang, M. Li, Z. Yu, and X. Zhou, "Freesense: Indoor human identification with wi-fi signals," 2016 IEEE Global Communications Conference (GLOBECOM), pp.1–7, 2016.
[18] C. Lin, P. Wang, C. Ji, M.S. Obaidat, L. Wang, G. Wu, and Q. Zhang, "A contactless authentication system based on wifi csi," ACM TRANSACTIONS ON SENSOR NETWORKS, vol.19, no.2, MAY 2023.
[19] W. Li, M.J. Bocus, C. Tang, S. Vishwakarma, R.J. Piechocki, K. Woodbridge, and K. Chetty, "A taxonomy of wifi sensing: Csi vs passive wifi radar," 2020 IEEE Globecom Workshops (GC Wkshps), pp.1–6, 2020.
[20] M.A. Khalighi, S. Long, S. Bourennane, and Z. Ghassemlooy, "Pam-and cap-based transmission schemes for visible-light communications," IEEE Access, vol.5, pp.27002–27013, 2017.
[21] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," 2021.
[22] Y. Xu, H. Wei, M. Lin, Y. Deng, K. Sheng, M. Zhang, F. Tang, W. Dong, F. Huang, and C. Xu, "Transformers in computational visual media: A survey," COMPUTATIONAL VISUAL MEDIA, vol.8, no.1, pp.33–62, MAR 2022.
[23] O. Moutik, H. Sekkat, S. Tigani, A. Chehri, R. Saadane, T.A. Tchakoucht, and A. Paul, "Convolutional neural networks or vision transformers: Who will win the race for action recognitions in visual data?," SENSORS, vol.23, no.2, JAN 2023.
[24] Y. Zhao, J. Li, X. Chen, and Y. Tian, "Part-guided relational transformers for fine-grained visual recognition," IEEE TRANSACTIONS ON IMAGE PROCESSING, vol.30, pp.9470–9481, 2021.
[25] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," arXiv, 2017.
[26] F. Gringoli, M. Schulz, J. Link, and M. Hollick, "Free your csi: A channel state information extraction platform for modern wi-fi chipsets," WiNTECH '19, New York, NY, USA, p.21–28, Association for Computing Machinery, 2019.
[27] W.C. Yeh, Y.P. Lin, Y.C. Liang, and C.M. Lai, "Convolution neural network hyperparameter optimization using simplified swarm optimization," 2021.
[28] M. Kashif, "Urdu handwritten text recognition using resnet18," 2021.
[29] X. Li, C. Wang, X. Huang, and Y. Nie, "A gru-based mixture density network for data-driven dynamic stochastic programming," 2020.
[30] Z. Wu, C. Baek, C. You, and Y. Ma, "Incremental learning via rate reduction," IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2021.
[31] S. Mittal, S. Galesso, and T. Brox, "Essentials for class incremental learning," arXiv preprint arXiv:2202.00002, 2022.
[32] G.M. van de Ven, T. Tuytelaars, and A.S. Tolias, "Three types of incremental learning," Nature Machine Intelligence, vol.4, pp.1185–1197, 2022.

Jichen Bian received the B.S. degree in Electronic Information Science and Technology from Zhengzhou University, Zhengzhou, China, in 2019. He is currently pursuing the Ph.D. degree in Commnuication and Information System with Shanghai Institute of Microsystem and Information Technology, CAS. His research interests include Wi-Fi-based integrated sensing and communications, heterogeneous network resource allocation technology, and wireless sensor networks.

Min Zheng received the Ph.D. degree in Commnuication and Information System from Beijing University of Posts and Telecommunications, Beijing, China, in 2006. He is currently a Professor at the Laboratory of Broadband Wireless Technology, Shanghai Institute of Microsystem and Information Technology, CAS. His research interests include Automatic configuration technology for self-organizing networks and emergency applications of broadband wireless communication systems.

Hong Liu received the B.S. and Ph.D. degrees in 2003 and 2011, respectively, from the School of Information Science and Technology, Peking University, Beijing, China. He is currently a Professor at the Laboratory of Broadband Wireless Technology, Shanghai Institute of Microsystem and Information Technology, CAS. His research interests include broadband self-assembling networks, 4G/5G communications and communications in complex electromagnetic/terrain environments.

Jiahui Mao received the B.S. degree in Electronic Information Engineering from Zhejiang University of Technology, Hangzhou, China, in 2020. He is currently pursuing the Ph.D. degree in Commnuication and Information System with Shanghai Institute of Microsystem and Information Technology, CAS. His research interests include resource management for wireless networks , mobile edge computing, and wireless sensor networks.

Hui Li received the Ph.D. degree in 2012 from the School of Communications and Information Engineering, Tongji University, Shanghai, China. She is currently a Professor at the Laboratory of Broadband Wireless Technology, Shanghai Institute of Microsystem and Information Technology, CAS. Her research interests include wireless ad-hoc network, artificial intelligence and communication.

Chong Tan received the M.S. and Ph.D. degrees in 2009 and 2013, respectively, from the School of Communications and Information Engineering, Shanghai University, Shanghai, China. She is currently a Professor at the Laboratory of Broadband Wireless Technology, Shanghai Institute of Microsystem and Information Technology, CAS. Her research interests include Internet of Things, sensing and positioning, edge computing, etc.