Vision Transformers for Human Activity Recognition Using WiFi Channel State Information

Fei Luo, Salabat Khan, *Member, IEEE*, Bin Jiang, *Member, IEEE*, and Kaishun Wu, *Fellow, IEEE*

***Abstract*—Wireless sensing and communication evolved separately in the past. However, integrated sensing and communication (ISAC) unlocks a new era of mobile network capabilities, with WiFi emerging as a prime candidate. By leveraging existing WiFi infrastructure and frequencies, ISAC enables powerful services like accurate localization and human activity recognition (HAR). WiFi-based HAR is a prime example powered by the magic of ISAC. WiFi channel state information (CSI) is susceptible to human movement disturbances; the alterations in CSI mirror the dynamic attributes of human activities. Given the intricate relationship between human activities and CSI, numerous deep learning models have been introduced to enhance HAR accuracy. Recently, transformer-based models have achieved excellent performance in various tasks, including speech recognition, natural language processing, and image classification. This has spurred research into incorporating transformer-based models into WiFi sensing applications. However, their application in WiFi-based HAR remains nascent. Vision transformer (ViT) is well-suited for analyzing WiFi CSI signals in the form of spectra, such as the Doppler frequency spectrum frequently utilized in related studies, owing to its data structure mimicking that of images. In this study, we explored five widely used ViT architectures (vanilla ViT, SimpleViT, DeepViT, SwinTransformer, and CaiT) for WiFi CSI-based HAR using two publicly available data sets, UT-HAR and NTU-Fi HAR. Our work aims to assess and compare the performance of diverse ViT architectures for WiFi CSI-based HAR and provide guidelines for WiFi-based HAR modeling and ViT selection, considering accuracy, model size, and computational efficiency.**

***Index Terms*—Human activity recognition (HAR), vision transformer (ViT), WiFi channel state information (CSI), WiFi sensing.**

Manuscript received 29 January 2024; revised 17 February 2024; accepted 3 March 2024. Date of publication 21 March 2024; date of current version 23 August 2024. This work was supported in part by the China NSFC under Grant U2001207, Grant 61872248, Grant 62202308, and Grant 62102264; in part by the Guangdong Provincial Key Laboratory of Integrated Communication, Sensing and Computation for Ubiquitous Internet of Things under Grant 2023B1212010007; in part by the Project of DEGP under Grant 2023KCXTD042; in part by the Young Expert Project of Taishan Scholar under Grant tsqnz20230602; and in part by the Youth Innovation University Team Project in Shandong under Grant 2022KJ062. *(Corresponding author: Kaishun Wu.)*

Fei Luo is with the School of Computing and Information Technology, Great Bay University, Dongguan 523000, China (e-mail: luofei2018@outlook.com).

Salabat Khan is with the School of Computer and Information Engineering, Qilu Institute of Technology, Jinan 250202, Shandong, China (e-mail: salabatwazir@gmail.com).

Bin Jiang is with the College of Oceanography and Space Informatics, University of Petroleum, Qingdao 266580, China (e-mail: jiangbin@upc.edu.cn).

Kaishun Wu is with the Information Hub, The Hong Kong University of Science and Technology, Guangzhou 510000, China (e-mail: wuks@hkust-gz.edu.cn).

Digital Object Identifier 10.1109/JIOT.2024.3375337

I. Introduction

The Internet of Things (IoT) is rapidly expanding, connecting billions of devices and generating a vast amount of data. However, traditional networks often struggle to efficiently manage this ever-growing traffic. Enter integrated sensing and communication (ISAC), a paradigm shift that merges sensing and communication capabilities into a single network. Next-generation mobile access (NGMA) takes this further, envisioning future networks that not only connect but also understand their environment. Here, WiFi emerges as a leading contender, with its ubiquity and untapped sensing potential. By analyzing subtle changes in WiFi signals, we can unlock powerful new applications like real-time localization, anomaly detection, and even human activity recognition (HAR). This transformative capability of WiFi sensing promises to revolutionize the IoT, bringing intelligence and context awareness to our connected world. WiFi sensing plays a significant role in empowering the capabilities of multifunction NGMA. By providing accurate location awareness, rich context information, and enabling novel applications, it paves the way for a more personalized, efficient, and intelligent mobile network experience.

The accurate identification of human activities poses a nontrivial and challenging task. Diversified HAR techniques have been employed, including wearable sensor-based, vision-based, radar-based, and WiFi-based approaches [1]. Wearable sensor-based HAR has gained widespread attention driven by the prevalence of smartphones and smartwatches. Leveraging embedded sensors akin to accelerometers, gyroscopes, and magnetometers, this technique enables the distinction of various activities. Moreover, electromyography (EMG) [2], electrocardiogram (ECG) [3], and photoplethysmography (PPG) [4] sensors have been incorporated into wearable devices for HAR by analyzing physiological characteristics of the human body. Nevertheless, wearable sensors require continuous wearing or carrying, potentially causing user inconvenience. Vision-based HAR emerges as the most extensively investigated and applied technique, as cameras are widely deployed and image/video data closely resemble human visual observation. Despite these advantages, computer vision (CV) faces challenges such as privacy concerns and lighting conditions. Both radar and WiFi modalities utilize radio frequency (RF) technology and operate device-free, offering privacy protection, unobtrusiveness, and resilience to lighting variations. Radar-based techniques generally outperform WiFi-based approaches in terms of sensitivity and spatial resolution [5], but they necessitate users to purchase and install dedicated radar sensors for radar-based HAR systems [6]. WiFi, as an omnipresent backbone of the Internet infrastructure, stands out for its cost-effectiveness and ease of deployment. Therefore, HAR based on commercially available WiFi has emerged as a burgeoning research trend.

WiFi signals can extend their coverage for tens of meters indoors, and the wireless connections among electronic devices generate a wealth of reflected echoes. Human presence and behaviors significantly affect WiFi signals, causing substantial changes in both the phase and amplitude of the received signals. These alterations can be exploited to categorize distinct human activities [7]. WiFi-based techniques typically utilize two metrics, received signal strength indication (RSSI) and channel state information (CSI), to infer human activity. RSSI refers to the path loss of wireless signals relative to a particular distance. The presence of a human body within WiFi coverage triggers signal attenuation, resulting in variations in RSSI measurements. RSSI, readily accessible on most WiFi devices, is apt for coarse-grained indoor localization and activity recognition. In comparison to RSSI, CSI serves as a more detailed metric, providing a channel estimation for each subcarrier of each transmission link. This encompasses the phase and amplitude information on each subcarrier in the frequency domain [8]. WiFi CSI employs complex values to represent the phase shift and amplitude attenuation of multiple paths, enabling the generation of unique patterns of human motions. Some commercially available WiFi devices (e.g., Atheros 9580 NIC [9] and Intel 5300 NIC [10]) grant the ability to retrieve CSI values, facilitating researchers to execute more refined HAR, including gesture recognition [11], lip reading [12], and vital signs monitoring [13], by analyzing CSI alteration patterns in both amplitude and frequency domains. Numerous previous research contributions have demonstrated that CSI outperforms RSSI in both HAR and indoor localization. WiFi CSI-based HAR has emerged as the cutting-edge trend in device-free HAR.

A WiFi CSI-based HAR system typically consists of several key components: signal collection, data processing, feature extraction, and activity prediction. Each component plays a crucial role in the whole system's functionality, enabling the detection and classification of human activities based on the analysis of WiFi signals. Data collection is the initial step in any machine learning or signal processing system. In a WiFi CSI-based HAR system, data collection involves capturing the CSI data from WiFi signals. There are several options available for collecting CSI data, including Atheros CSI Tool [9], Linux 802.11n CSI Tool [10], ESP32-CSI-Tool [14], and Nexmon CSI Extractor [15]. Signal processing is crucial for preprocessing the collected CSI data. It involves many signal processing methods to enhance the quality and accuracy of the data. Typical signal processing techniques include noise reduction, synchronization, and channel equalization. These techniques help in removing noise and distortions from the collected CSI data, ensuring accurate analysis and classification of human activities. Feature extraction aims to identify and extract the most informative aspects of the preprocessed CSI data. These features capture the essential characteristics of the WiFi signals that are useful for activity classification. Common features include amplitude, frequency, and time-domain characteristics of the CSI data. Feature extraction plays a crucial role in determining the accuracy and performance of the activity classification algorithm. Activity classification is the final component of a WiFi CSI-based HAR system. Most researchers focus on improving the accuracy of activity classification. Activity recognition typically relies on machine learning algorithms. Common machine learning methods used for activity classification contain random forests, support vector machines (SVMs), and deep learning models such as convolutional neural networks (CNNs) and RNNs. These algorithms learn to associate specific patterns or features in the CSI data with different human activities. With the emergence of deep learning, feature extraction can be omitted. Deep learning's hierarchical structure facilitates the automatic extraction of features from raw WiFi signals. Currently, deep learning is the most prominent technique for CSI-based HAR.

Over the recent years, a diverse array of deep learning techniques, including CNNs, RNNs, and long short-term memory (LSTM) networks, have been employed for WiFi CSI-based HAR. CNNs demonstrate remarkable proficiency in extracting intricate local spatial or spectral patterns from WiFi CSI spectra. RNNs, on the other hand, excel at modeling intricate temporal dependencies within time-series CSI signals. However, CNNs face challenges in uncovering global feature correlations, while RNNs encounter limitations in capturing long-range dependencies due to their memory constraints [16]. While hybrid models, including the CRNN [1], CNN-LSTM [17], and CNN-GRU [18], have been proposed to integrate local spectral features and temporal dependencies, they still fail to adequately address the shortcomings of long-range dependency and global feature extraction. The transformer has ascended to prominence in the realm of deep learning architectures, particularly within natural language processing (NLP) and CV. Its architecture leverages the self-attention mechanism and position encoding to efficiently obtain long-range temporal and spatial information simultaneously. Self-attention empowers the model to simultaneously analyze all positions while computing the output of a particular position, significantly improving computational efficiency. Moreover, it facilitates the establishment of connections between all positions simultaneously, enabling effortless long-range dependency detection.

Variant transformers such as SwinTransformer [19], DeepViT [20], CaiT [21], etc., have been implemented to enhance the performance of various tasks. Recently, some researchers began to adopt transformers for WiFi CSI-based HAR. As illustrated in Fig. 1, spectrograms generated from CSI data can be patched and input into a transformer for human activity classification. For example, Yang et al. [16] proposed WiTransformer to classify body-coordinate velocity profiles (BVPs) generated from WiFi CSI data for gesture recognition. Zhou et al. [22] developed a novel WPFormer architecture to effectively map WiFi CSI signals to human pose landmarks, thereby accurately estimating the spatial context of human posture. However, most of the current work simply

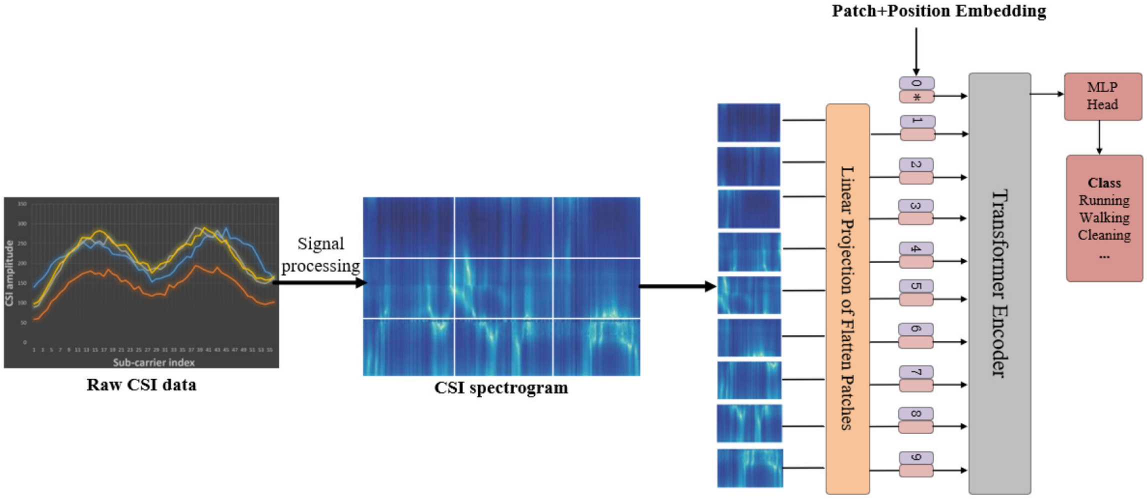


Fig. 1. ViT for CIS-based activity recognition.

adopts the vanilla transformer or integrates the transformer with CNN layers on the WiFi CSI data set. Since there are many different transformers, it is still unclear and waiting to explore which transformer structure is suitable for WiFi CSI-based HAR. Since most work processes WiFi CSI data to spectrograms, vision transformers (ViTs) are more suitable for analyzing this data structure. In this article, we investigated the five most used ViT structures (vanilla ViT, SimpleViT [23], DeepViT, SwinTransformer, and CaiT) and applied them to two public WiFi CSI data sets including UT-HAR [24] and NTU-Fi HAR [25]. Our work is the first to investigate the impact of different transformer architectures for WiFi CSI-based HAR. The contribution of our work can be summarized as follows.

1) We applied five different ViTs upon two WiFi CSI data sets for HAR. These five transformers are the most used in related work and they achieved state-of-the-art (SOTA) in different tasks. We fine-tuned their hyperparameters to optimize their performance for WiFi CSI-based HAR.
2) We evaluated and compared these five transformers not just from accuracy but also from model size and computation efficiency. It is found that the model size and computation complexity of ViT cannot guarantee accuracy; and CaiT has achieved the best overall performance with the consideration of accuracy, model size, and computation efficiency.
3) We compared our work with the related work. The CaiT implemented in our work has achieved SOTA on the UT-HAR data set. Although its performance on the NTU-Fi HAR data set is not superior, it keeps the advantages in model size and computation efficiency.

This article is structured as follows. Section II provides a concise overview of WiFi CSI-based HAR, Deep learning for WiFi sensing, and transformer. Section III unveils the methodologies and techniques employed for WiFi CSI-based HAR, along with detailed descriptions of the architectures of five transformers. Section IV delves into the evaluation of these five transformers using three metrics, followed by a comprehensive analysis of the results. The final section summarizes the findings of this study.

II. Related Work

As our work focuses on WiFi CSI-based HAR using transformers, this section reviews related works addressing WiFi CSI-based HAR, deep learning for WiFi sensing, and transformers.

A. WiFi CSI-Based HAR

WiFi has become an indispensable backbone for the Internet, handling over half of its traffic and powering a plethora of applications. It is an Internet infrastructure deployed in almost every living and working house. The global value of WiFi is expected to increase to $4.9 trillion by 2025 [26]. Wireless sensing using WiFi is also increasingly getting attention from academia and industry since it is a device-free and cost-effective technique. With widely deployed WiFi routers, it does not need to install any extra device. WiFi sensing also has the following advantages: contactless, privacy-preserving, unintrusive, and light-insensitive. WiFi sensing can be classified as RSSI-based and CSI-based techniques, which use two signal indicators. CSI is more stable and information-intensive than RSSI [27]. CSI-based techniques can provide more robust and fine-grained WiFi sensing. WiFi CSI-based sensing has a wide range of emerging applications, which can be categorized into five categories: activity recognition, intrusion/occupancy detection, vital sign monitoring, localization, and person identification. In this work, we pay attention to WiFi CSI-based HAR.

WiFi CSI-based HAR aims to distinguish performing activities by examining the unique effects of each activity on the received WiFi CSI signals. Different human activities induce different shifts in the received CSI signals. Besides human activity, the scenario layout, environmental obstacles, and vibrations also have impacts on WiFi signals. To eliminate environmental noise and signal distortion, CSI sanitization is required, which generally comprises denoising, outlier removal, and phase calibration [28]. Common denoising methods contain the Butterworth low-pass filter, principal component analysis (PCAs), wavelet filters, etc. Outliers Removing is to eliminate some anomalous measurements. Hampel filter is a frequently used outlier removing method [29], [30]. The phase of received CSI signals is often plagued by various impairments, including carrier frequency offset (CFO), the phase-locked-loop (PLL) initial phase, packet detection delay (PDD), and sampling frequency offset (SFO) [28]. The methods for phase calibration reply on the application. The raw phase information can be refined through a linear transformation, thereby effectively eliminating the predominant portion of random phase offsets [31]. WiFi CSI-based HAR can be broadly classified into two prominent approaches: model-based and learning-based [32]. The model-based approach relies on mathematical representations to map the association between the received CSI data and the performing activity. Wang et al. [33] devised a novel CSI-activity model, aptly named CRAM, which meticulously quantifies the relationship between human speeds and their corresponding activities. IndoTrack [34] proposed a Doppler-AoA model and applied MUSIC method to track a target. The Fresnel zone model meticulously captures the nuanced link between the movements of a target and the associated signal variations, illuminating the interplay of distance and signal behavior [35]. Learning-based methods enable activity recognition by employing learning algorithms to map CSI signals and corresponding activities. Learning-based methods depend on the data collected from human activity and usually achieve higher recognition accuracy. Activity recognition can be accomplished using a variety of learning algorithms, ranging from conventional machine learning methods to advanced deep learning architectures. The performance of conventional machine learning also depends on the quality of the extracted features of CSI samples. Deep learning can perform feature extraction automatically but usually needs a large volume of data.

WiFi CSI-based HAR has broader applications in healthcare, human-computer interaction, and smart homes. Many researchers have built various WiFi CSI-based HAR systems for these areas. Wi-Sleep, a novel sleep monitoring system proposed by Liu et al. [29], utilizes the unique features of CSI signals to effectively extract rhythmic patterns linked to respiration and sudden alterations arising from body movements, enabling the accurate assessment of sleep patterns and identifying potential sleep disturbances. WiFall [36], a pioneering CSI-based fall detection system, utilizes a SVM classifier to effectively distinguish falls from common human activities, such as sitting, standing up, and walking. Wi-Key [37], a novel keystroke recognition system, employs the unique characteristics of CSI signals to accurately identify the typed keys based on the distinctive pattern changes observed at the WiFi signal receiver end. WiG [38], a groundbreaking CSI-based gesture recognition system, exhibits remarkable performance, achieving an impressive recognition accuracy of 92% in line-of-sight (LOS) scenarios and 88% in non-LOS scenarios. WiFinger, proposed by Li et al. [39], has demonstrated remarkable capabilities in recognizing nine finger gestures of sign language, achieving an astounding accuracy of 90.4%. As wireless technology and deep learning are continuously improved, more novel and valuable applications in various commercial areas will be created.

B. Deep Learning for WiFi Sensing

Deep learning has demonstrated its remarkable prowess in a diverse range of tasks, including image classification, object detection, and speech recognition. Compared to conventional machine learning, deep learning takes advantage of a large volume of complex data, and can efficiently extract features and achieve accurate results [40]. It is a challenging task to map WiFi signals to human activities since WiFi signals are complex and affected by many surrounding objects. Model-based approaches face limitations in comprehensively capturing the intricate interplay between WiFi signals and human behaviors, hindering their ability to fully mathematically represent this complex relationship. Deep learning can model these complex correlations and perform complex human activity classification tasks. In this section, we will introduce some classic deep learning architectures used in WiFi CSI-based HAR.

CNN has rapidly gained recognition as a versatile tool for wireless sensing, particularly when leveraging the rich information embedded in CSI data. Natural visual perception is the main driver behind CNN and has achieved tremendous success in CV. CNN has three characteristics: sparse connectivity, parameter sharing, and equivariant representations [41]. It excels in feature extraction by employing stacked convolutional kernels and spatial pooling operations. In the context of WiFi sensing, these kernels can analyze either 2-D patches of CSI data, capturing spatial relationships, or 1-D patches of each subcarrier, extracting temporal dynamics. This ability to extract spatial-temporal features is a key advantage of CNNs for WiFi sensing, particularly considering their reduced training parameters and the preservation of CSI data's subcarrier and time dimension [25]. Nakamura et al. [42] used a CNN to classify the spectrogram generated by WiFi CSI data for fall detection and it achieved 90% accuracy in experiments. WiPass [43] is a keystroke recognition system that employs 1D-CNN to detect and classify keystrokes through the fluctuation of WiFi CSI. BeAware [44] converted WiFi CSI data into time-series heatmap images and used CNN to recognize the corresponding behaviors.

RNN is frequently used for sequential data modeling, including speech recognition, weather forecasts, and time-series prediction. RNNs excel in handling sequential data by utilizing recurrent connections, enabling them to maintain a memory of past inputs, allowing them to memorize arbitrary-length sequences and learn meaningful relationships between them. Standard RNNs are susceptible to the challenges of vanishing and exploding gradients, hindering their ability to effectively capture long-term dependencies in sequential data [45]. LSTM networks overcome the limitations of conventional RNNs by introducing memory cells and three gating methods: input, forget, and output gates. The three gating methods regulate the flow of information, allowing LSTMs to efficiently store and maintain long-term dependencies. In contrast, gated recurrent units (GRUs) simplify the gating process by merging the input and forget gates into a single update gate. This simplification makes GRUs more computationally efficient than LSTMs while still maintaining the ability to capture long-term dependencies. Since human activity is dynamically changing over time and WiFi CSI data is time-series, it can be considered a sequential problem. While CNN is suitable for spatial feature representation. Hence, RNN is more suitable to model temporal dependencies. Zhang et al. [46] implemented an LSTM on the synthetic data through eight CSI transformation methods to recognize five activities and achieved around 90% of accuracy. Ming et al. [47] proposed a human identification system that uses an LSTM to extract temporal characteristics of human gait and perform human identification. Ding and Wang [48] employed a combination of statistical features, involving time-frequency analysis (TFA) and channel power variation (CPV), to build an RNN model capable of recognizing human activities.

Many research combine different basic deep learning structures in a hybrid model to improve the performance of HAR. The most popular hybrid model is the convolutional recurrent model, which stacks convolutions and recurrent blocks sequentially to combine the spatial pattern extraction of CNNs and the temporal dependency modeling of RNNs. Guo et al. [49] proposed an innovative hybrid model called LCED, which combines the strengths of LSTMs and CNNs to mitigate the accuracy variations among individuals in CSI-based HAR. Zhang et al. [50] proposed a HAR system called AF-ACT, which can fuse the semantic features extracted by CNN and temporal features extracted by GRU, and achieved an accuracy of 91.23% in recognizing eight activities.

There are many other deep learning architectures including adversarial networks, autoencoders, and transformers. The former two networks have more complex architectures but usually still consist of basic neural network components, such as CNN and RNN. Transformers recently raised lots of attention due to their excellent performance in many tasks. We will introduce the transformer in the next section.

C. Transformers

Transformers leverage the attention mechanism to significantly boost feature representation by aligning queries with essential salient features. This facilitates the extraction of long-range dependencies across both spatial and temporal domains. CNNs are limited to identifying local features, while RNNs tend to lose track of earlier inputs in lengthy sequences. Attention constructs a context vector and utilizes it in conjunction with the previous state to determine the following state of the decoder. Transformers harness three distinct forms of attention: self-attention, masked self-attention, and cross-attention. Various research has applied transformers for HAR task. Li et al. [51] developed a multiagent transformer network (MATN), which is a multiagent attention-based deep learning algorithm for addressing multimodal feature extracting and spatial-temporal relationship modeling in multimodal HAR. Sowmiya and Menaka [52] employed a transformer-based model to classify human activities from data collected using IMU. This work leverages self-attention mechanisms in transformers to explore complex patterns from time-series data. Wensel et al. [53] proposed and designed a neural network composed of two transformers, a recurrent transformer (ReT) and a ViT to improve the speed and scalability of activity recognition.

As transformers maintain a consistent width across their layers, a learnable linear transformation is employed to map each vectorized path to the model dimension d, producing patch embeddings. ViT architectures can connect with every CSI patch, enabling them to model long-range temporal dependencies and global spatial features within CSI data. Numerous ViT variants have been developed over the past years, and these variants have also been applied to WiFi CSI-based HAR. Li et al. [54] proposed a novel deep learning architecture called THAT that effectively extracts range-based patterns from CSI data for HAR. This model incorporates a multiscale convolution module and a residual-connected multihead attention (MHA) mechanism to capture both local and global features, demonstrating SOTA performance in terms of both accuracy and efficiency on four benchmark HAR data sets. Abdel-Basset et al. [55] devised an H2HI-NET that harnesses the strengths of residual learning and transformer networks to effectively extract intricate spatial features associated with human activities. Yang et al. [16] proposed WiTransformer, a novel tactic based on pure ViTs, to classify the BVP derived from WiFi CSI for HAR. Yang et al. [25] compared the performance of the ViT with MLP, CNN-5, and RNN in WiFi-based HAR. Yao et al. [56] proposed a hybrid deep learning model that combines a CNN and a ViT for WiFi CSI-based gesture recognition with the consideration of spatial localization characteristics and long-distance dependence. These studies demonstrate that ViTs hold immense promise for CSI-based HAR applications.

Although more and more researchers adopt transformers and ViTs for WiFi CSI-based HAR, there is still no work to compare the performance of different ViTs. Our work investigated five ViTs and compared their performance on two WiFi CSI data sets. Our work aims to offer guidance for the selection of ViTs in CSI-based HAR.

III. Methodology

Since our work investigates WiFi CSI-based HAR using different transformer architectures, we will introduce the mathematical theories of CSI and the architectures of five commonly used transformers in this section.

A. Channel State Information

The CSI embedded within the WiFi preamble is obtained from the training symbols, which are modulated using orthogonal frequency-division multiplexing (OFDM). OFDM transceivers divide a wideband channel into W orthogonal subchannels. In OFDM systems, groups of transmit bits are mapped onto a smaller number of OFDM symbols, denoted by W', where $W' < W$ symbols, Each OFDM symbol is then employed to modulate a unique subcarrier in the OFDM signal using standard modulation techniques like QAM or PSK [57]. The k^{th} OFDM symbol x_k is transmitted within a time interval $t \in [kT, (k+1)T]$ with duration T, can be represented as

$$x_k(t) = \sum_{w=1}^{W} a_{w,k} \exp\left[j2\pi \left(f_c + \frac{f_w}{T}\right) t\right] \tag{1}$$

where $a_{w,k}$ represents the constellation point modulating the w^{th} subcarrier of the k^{th} symbol, f_w denotes the base-band frequency of the w^{th} subcarrier, and f_c represents the central frequency of the WiFi channel.

Let's examine an OFDM communication system that employs W subcarriers. For a single OFDM symbol, (1) can be decomposed into a product of a modulating vector $x = [a_1, \ldots, a_W]$ and a symbol-independent vector of complex exponentials. As a result, the connection between the transmitted signal $x \in \mathbb{C}^W$ and the received signal $y \in \mathbb{C}^W$ is expressed as

$$\mathrm{y} = H \circ x \tag{2}$$

where $H \in \mathbb{C}^W$ represents the frequency response of the wideband wireless channel, and $\circ$ denotes the Hadamard product, which is equivalent to element-wise multiplication.

As the OFDM channel typically exhibits frequency selectivity, the receiver must accurately gauge its impact to adequately equalize the received signal. The CSI $\mathcal{H} \in \mathbb{C}^W$ serves as a quantized estimation of the channel response H calculated by the receiver using the HE-LTF symbols from the WiFi preamble. Utilizing the CSI, the receiver can invert (2) to equalize the received signal and ultimately restore the original transmitted signal

$$x \simeq \mathcal{H} \circ \mathrm{y}. \tag{3}$$

In scenarios where the receiver has multiple antennas, $N > 1$, it can generalize (2) and (3) by computing y_i for each antenna i from 1 to N. This enables simultaneous acquisition of N distinct CSI $\mathcal{H}_i$ from a single transmission.

The characteristics of the wireless channel, including its bandwidth, antenna configuration, and subcarrier spacing, significantly impact the obtained CSI. Utilizing wider bandwidths and finer frequency resolution is beneficial for CSI-based sensing applications. The CSI exhibits a unique interference pattern that arises from the accumulation of multiple signal copies at the receiver. The frequency domain representation of the wireless channel's properties, as measured through the CSI, is analogous to its time domain representation through the channel's impulse response. Given a signal $x(t)$ as a function of time t and its Fourier transform $X(f)$ as a function of frequency f, the relationship is expressed as

$$x(t-\gamma) x\mathcal{F}X(f) \cdot \exp(-j2\pi f\tau) \tag{4}$$

where $\mathcal{F}$ represents the Fourier transform operator and τ signifies a time delay. As demonstrated by (4), receiving multiple copies of the WiFi signal at distinct τ's (e.g., due to multipath propagation) results in a combination of complex exponentials in the frequency domain. Broader CSI bandwidth assists in distinguishing lower-frequency components, while higher spectral resolution facilitates the detection of higher-frequency components.

B. Architectures of ViTs

The ViT was proposed in 2021 [58]. It represents images as sequences of patches, where each patch is flattened into a single vector by concatenating the pixel values of all channels and then further transformed into the desired input dimension through a linear projection. Compared to CNNs, ViTs can extract both global and local features from the early layers; they can retain more spatial information than residual networks (ResNets); and ViTs can efficiently acquire high-quality intermediate representations by training on massive data sets. Due to their similarity to images, ViTs are also well-suited for CSI-based sensing tasks. We will introduce five different ViT models applied for WiFi CSI-based HAR In this section.

1) Vanilla ViT: The vanilla ViT was introduced in 2021 [58]. While the standard Transformer operates on a sequence of token embeddings, ViTs handle images by treating them as sequences of image patches. Each patch is transformed into a single vector representation by combining the pixel values from all channels within the patch and then further transformed linearly into the desired input dimension. Given an image $x \in \mathbb{R}^{H \times W \times C}$, it can be divided into a sequence of 2-D patches $x_p \in \mathbb{R}^{N \times (P^2 \cdot C)}$, where (H, W) represents the image resolution, C denotes the number of channels, (P, P) defines the patch size, and $N = HW/P^2$ indicates the total number of patches. The patch-based representation determines the actual length of the input sequence for the transformer. The transformer employs a fixed latent vector size D throughout its layers, so the flattened patches are projected to D-dimensional vectors using a learnable linear projection. These patch embeddings are further augmented with position embeddings to retain spatial information. The resulting sequence of embedding vectors is then fed into the transformer encoder. As illustrated in Fig. 2, the transformer encoder's structure alternates between MLP blocks and multiheaded self-attention layers.

2) SimpleViT: SimpleViT [23] builds upon the vanilla ViT architecture with a few subtle adjustments. It employs a smaller batch size of 1024 compared to the original 4096, utilizes global average pooling (GAP) to aggregate feature information instead of relying on a class token, adopts fixed 2-D sine-cosine position embeddings, and incorporates a moderate level of RandAugment [59] and Mixup [60] augmentations (level 10 and probability 0.2, respectively). These seemingly minor modifications result in substantial performance gains compared to the original ViT. Furthermore, SimpleViT incorporates additional architectural changes, regularization techniques such as dropout or stochastic depth [61], advanced optimization algorithms like SAM [62], supplementary augmentations like CutMix [63], repeated augmentations [64], and blurring effects. While these modifications may not be groundbreaking individually, they collectively lead to significant performance improvements.

3) DeepViT: ViTs consist of three fundamental building blocks: a patch embedding layer, a transformer block stack for feature extraction, and a linear layer for classification. However, stacking transformer blocks can exacerbate the

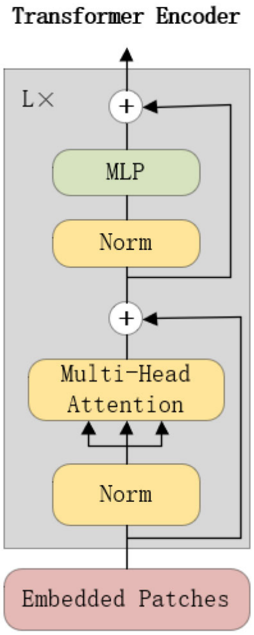


Fig. 2. Vanilla transformer encoder.

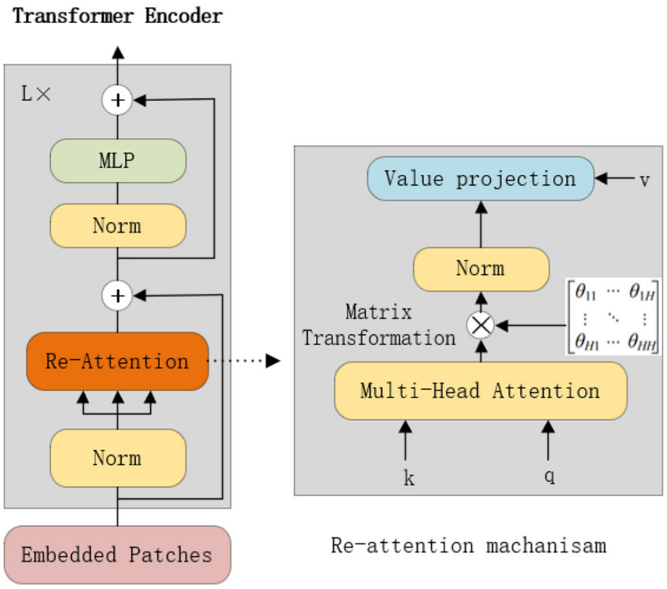


Fig. 3. Reattention mechanism in DeepViT.

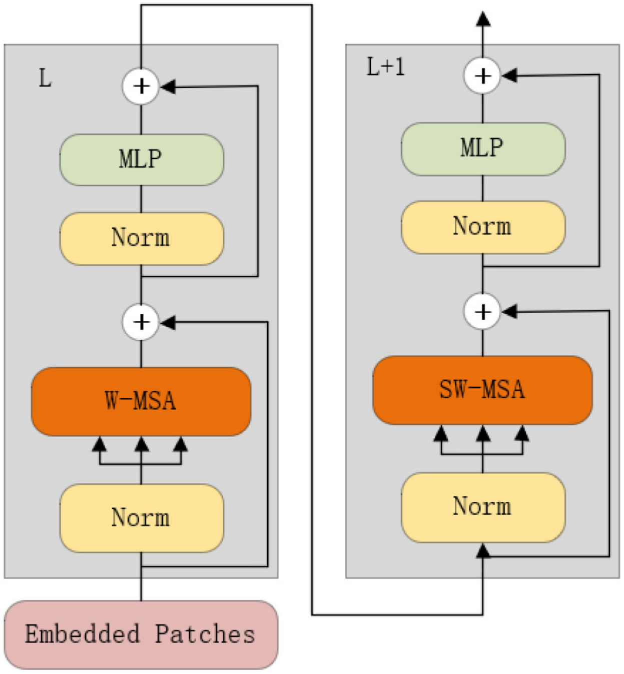


Fig. 4. Two successive swin transformer blocks.

problem of model rank degeneration [20] and limit the model's learning capacity. Attention collapse is a major challenge in scaling up ViTs to deeper architectures. DeepViT addresses this issue by introducing a novel reattention mechanism. As illustrated in Fig. 3, the mechanism facilitates cross-head information exchange, enabling the regeneration of attention maps and empowering deeper ViTs to achieve superior performance. DeepViT utilizes the attention maps from the heads as the foundation and dynamically aggregates them to create a new set of attention maps. A trainable transformation matrix $\Theta \in \mathbb{R}^{H \times H}$ is applied to mix the MHA maps into regenerated new ones, which are then multiplied by V. The Reattention is implemented as follows:

$$\text{Re-Atention}(Q, K, V) = \text{Norm}\left(\Theta^{\top}(\text{Softmax}(\frac{QK^{\top}}{\sqrt{d}}))\right)V \tag{5}$$

where transformation matrix Θ is multiplied to the self-attention map $\mathbf{A}$ along the head dimension. The normalization function *Norm* is used to reduce the layer-wise variance. Θ is learnable throughout the training process. The Reattention mechanism effectively exploits the interaction between different attention heads to gather their complementary information and enhance the diversity of attention maps.

4) SwinTransformer: ViT's vanilla implementation employs global self-attention, which captures long-range dependencies within the input sequence. This global computation scales quadratically with the number of tokens, which makes it impractical for tasks that need a large number of tokens for dense prediction or high-resolution image representation. SwinTransformer, on the other hand, performs self-attention within local windows for efficient modeling. The windows partition the image into nonoverlapping segments. Additionally, SwinTransformer strikes a delicate balance between maintaining cross-window connections and preserving the efficiency of nonoverlapping windows through its shifted window partitioning approach. The shifted window arrangement alternates between two partitioning configurations in consecutive SwinTransformer blocks.

As presented in Fig. 4, SwinTransformer replaces the standard multihead self-attention (MSA) module with a shifted window-based module in a Transformer block, while retaining other layers. Using the shifted window partitioning approach, consecutive SwinTransformer blocks can be formulated as follows:

$$\hat{\mathbf{z}}^{l} = \text{W-MSA}\left(\text{LN}(\hat{\mathbf{z}}^{l-1})\right) + \hat{\mathbf{z}}^{l-1} \tag{6}$$

$$\mathbf{z}^l = \text{MLP}\Big(\text{LN}(\hat{\mathbf{z}}^l)\Big) + \hat{\mathbf{z}}^l \tag{7}$$

$$\hat{\mathbf{z}}^{l+1} = \text{SW-MSA}\Big(\text{LN}(\mathbf{z}^l)\Big) + \mathbf{z}^l \tag{8}$$

$$\mathbf{z}^{l+1} = \text{MLP}\Big(\text{LN}(\hat{\mathbf{z}}^{l+1})\Big) + \hat{\mathbf{z}}^{l+1} \tag{9}$$

where $\hat{\mathbf{z}}^l$ and $\mathbf{z}^l$ represent the output of the SwinTransformer's (S)W-MSA module and the MLP module for the block l, respectively; W-MSA and SW-MSA correspond to window-based MSA that leverages regular and shifted window partitioning strategies, respectively. The shifted window partitioning approach effectively connects neighboring nonoverlapping windows in the previous layer and has been demonstrated to be beneficial for tasks such as object detection, image classification, and instance segmentation.

5) CaiT Transformer: Fig. 5 shows the architecture of the CaiT transformer. It employs two distinct processing phases: the self-attention stage and the class-attention stage (CLS). The former replicates the vanilla ViT transformer structure but with the introduction of CLS. The latter stage encompasses a sequence of layers that aggregate patch embeddings into a class embedding CLS, which is then passed to a linear classifier. The CLS cycles between multihead class-attention (CA) and feed-forward network (FFN) layers. Only the class embedding is updated during this stage. Similar to the class token used in ViT and DeepViT, it is a learnable vector. However, unlike these architectures, CaiT avoids direct information transfer from the class embedding to the patch embeddings during the forward pass. Instead of directly updating the class embedding with patch embeddings, CaiT relies on the residual connections within the CA and FFN layers of the CLS to dynamically modify the class embedding. The CA layer focuses on extracting information from the processed patches, operating similarly to a SA layer, but reliant on attention between the class embedding x_{class} (initialized at CLS in the first CA) and itself along with the frozen patch embeddings x_{patches}.

Given a network with h attention heads and p patches, with an embedding size of d, the multihead CA module can be parameterized by a set of projection matrices, $W_q, W_k, W_v, W_o \in \mathbf{R}^{d\times d}$, along with their corresponding biases $b_q, b_k, b_v, b_o \in \mathbf{R}^d$. Hence, the CA residual block operates as follows. Initially, the patch embeddings are concatenated to form a combined representation $z = [x_{\text{class}}, x_{\text{patches}}]$. Subsequently, the projections are executed as follows:

$$Q = W_q x_{\text{class}} + b_q \tag{10}$$

$$K = W_k z + b_k \tag{11}$$

$$V = W_v z + b_v. \tag{12}$$

The CA weights are calculated as

$$A = \text{Softmax}\Big(Q.K^T/\sqrt{d/h}\Big) \tag{13}$$

where $Q.K^T \in \mathbf{R}^{h\times 1\times p}$. This attention is then used to compute the weighted sum $A \times V$, producing the residual output vector

$$\text{out}_{CA} = W_o A V + b_o \tag{14}$$

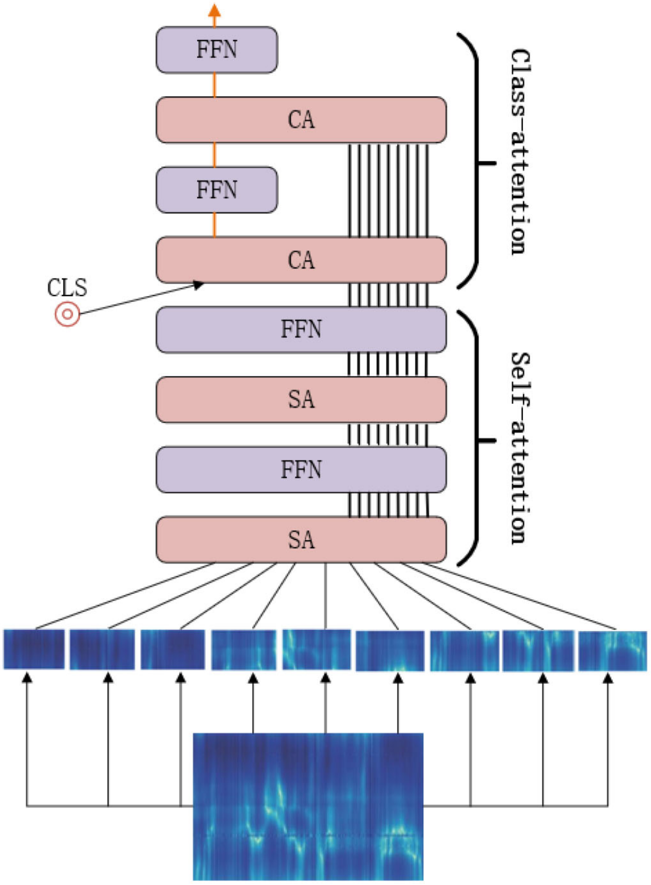


Fig. 5. CaiT architecture.

which is then combined with the class embedding X_{class} for further processing. Relevant information from the patch embeddings is effectively extracted by the CA layers and transferred to the class embedding.

IV. Experiments, Results, and Analysis

The implementation and evaluation of five ViT architectures for HAR on two WiFi CSI data sets are presented in this section. The performance of these models is assessed across three dimensions: accuracy, parameter count, and computational efficiency. Accuracy is determined using fivefold cross-validation and presented as average accuracy ± standard deviation. Parameter count reflects model size, while computational efficiency is measured in terms of multiply-accumulate operations (MACs), which indicates the model's computational complexity. These metrics are evaluated using the ptflops tool [65]. In this section, we conducted CSI-based HAR on two benchmark data sets: UT-HAR and NTU-Fi HAR. These data sets are chosen due to their diverse collection of activities performed in realistic environments, and their widespread use in various CSI-based HAR research endeavors.

A. Data Set Description

1) UT-HAR Data Set: The UT-HAR is the inaugural publicly used CSI data set for HAR. It encapsulates seven activities performed by six individuals: Lay Down, Pick up

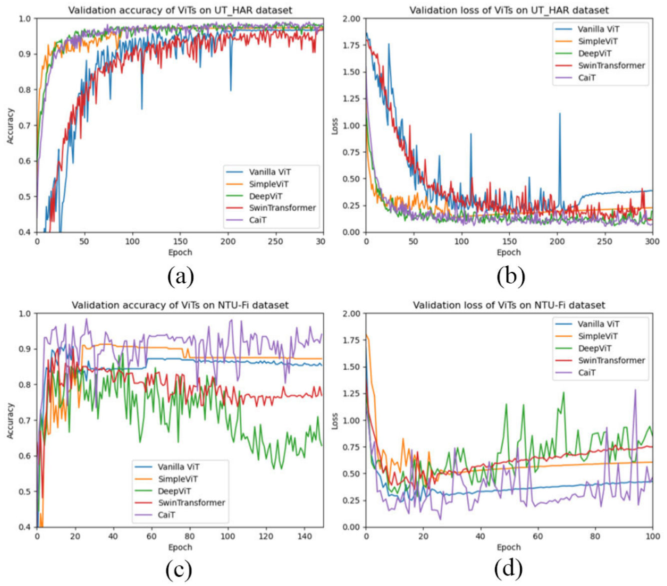
Fall, Sit Down, Run, Walk, and Stand Up, each with 20 trials. Collected within an indoor office setting, the data was acquired using a WiFi antenna system with the Tx and Rx placed 3 m separated in LOS conditions. The Rx employs a commercial Intel 5300 NIC, operating at a sampling rate of 1 kHz. Each activity is recorded for an interval of 20 s under LOS conditions. The raw CSI signals are segmented by using a sliding window of 250. The number of subcarriers and the streams for each sample are 30 and 2000. So the shape of one CSI amplitude sequence sample is $30 \times 3 \times 2000$ [66]. The data preprocessing involves PCA, followed by transformation into frequency spectrograms every 100 ms using STFT.

2) NTU-Fi Data Set: The NTU-Fi data set was acquired using the Atheros CSI Tool, implemented on a pair of TP-Link N750 APs, one serving as the transmitter and the other as the receiver. These APs feature three antenna duos working at 5GHz with 40MHz bandwidth, enabling the extraction of 114 subcarriers of CSI data per timestamp [67]. The CSI data was collected at a sampling rate of 500 Hz. The data set encompasses both Human identity (Human ID) and HAR tasks, comprising fourteen gait patterns and six activities of humans. The HAR data set was captured across three distinct environments. Seven female and thirteen male subjects performed six activities: running, boxing, cleaning floor, walking, falling down, and circling arms. Each subject repeated each activity 20 times, resulting in 400 samples per category. Each sample is recorded for 1 s. Thus, the shape of each sample is $3 \times 114 \times 500$. Since the related work directly implemented HAR models on the raw CSI signal, we did not apply preprocessing on this CSI data set for the comparison of model performance.

B. Evaluation

To effectively train and evaluate our models, we divided each data set into two portions: 80% for training and 20% for testing and evaluation. In this section, we trained all five ViTs on the training data sets and investigated their performance on the test data sets. For the UT-HAR data set, since the data has been processed into spectrograms, we further reshape the spectrogram to the size of (250, 90) as the input of the ViTs. The hyperparameters of ViTs comprise the *patch-size*, the dimensionality of patch embeddings, the number of encoder blocks (*depth*) along with self-attention (*dim*), the number of attention heads (*heads*), and the hidden dimension of MLP blocks (*mlp-dim*). We tuned the hyperparameters of five ViTs. The hyperparameters of vanilla ViT contains $patch - size = [18, 50]$, $depth = 1$, $dim = 900$, and $heads = 8$. DeepViT and SimpleViT have the same hyperparameters, which are $patch - size = [18, 50]$, $depth = 1$, $dim = 800$, $heads = 16$, and $mlp - dim = 2047$. The hyperparameters of CaiT are set as $patch - size = [18, 50]$, $depth = 1$, $dim = 300$, $heads = 1$, $mlp - dim = 600$, it also has one more hyperparameter, the depth of cross-attention of CLS tokens to patch (cls-depth), which is set to 1 in our work. The SwinTransformer has the hyperparameters of $patch - size = (25, 9)$, $depth = 1$, $heads = 2$, $mlp - dim = 800$, $embed_dim = 80$, and $window - size = 5$.

For the NTU-Fi data set, the input shape is (342, 500) since it is collected from 3 antenna pairs with 114 subcarriers and the sampling rate is 500 Hz. The hyperparameters of the ViT are tuned as $patch - size = (25, 9)$, $depth = 1$, $dim = 225$, $heads = 5$, $mlp - dim = 1024$. The hyperparameters of DeepViT and SimpleViT are tuned as $patch - size = (25, 18)$, $depth = 1$, $dim = 800$, $heads = 16$, $mlp - dim = 2048$. The hyperparameters of CaiT are set as $patch - size = (50, 38)$, $depth = 1$, $dim = 400$, $heads = 1$, $mlp - dim = 800$. And the SwinTransformer has the hyperparameters of $patch - size = (57, 30)$, $depth = 1$, $heads = 2$, $mlp - dim = 800$, and $window - size = 5$.



(a) (b) (c) (d)

Fig. 6. Validation accuracy and loss of five ViTs on two CSI data sets. (a) Validation accuracy of ViTs on UT HAR dataset. (b) VALIDATION loss of ViTs on UT_HAR dataset. (c) Validation accuracy of ViTs on NTU-FI dataset. (d) Validation loss of ViTs. on NTU FI dataset

We trained the ViT models on an NVIDIA A100 GPU card. The optimizer we used is Adam, which is initialized with a learning rate of 0.001. All ViTs are trained on the UT-HAR data set and the NTU-Fi data set, respectively. Fig. 6 shows the accuracy and loss of five ViTs generated in each epoch on the UT-HAR and NTU-Fi data sets. As can be seen, the training of ViTs goes to overfitting after around 250 epochs on the UT-HAR data set and overfits on the NTU-FI data set after around 50 epochs. We adopted early-stopping [68] to stop training when generalization error begins to degrade. The final trained models we saved are trained for around 250 epochs on the UT-HAR data set and for around 50 epochs on the NTU-FI data set.

After the training, we finally evaluated the trained ViTs on the test data sets. Except for the accuracy, we also show the metrics of the number of parameters (Params) and the number of MACs (MACs) in Table I. As can be seen, the accuracy of five ViTs on the UT-HAR data set has a small difference but a bigger difference on the NTU-FI HAR data set. CaiT achieved the best accuracy in both data sets; which is 98.78% on the UT-HAR data set and 98.2% on the NTU-FI HAR data set. DeepViT achieved the second-best accuracy on the UT-HAR data set but the worst accuracy on the NTU-FI HAR data set. For the aspect of the model size and computation efficiency, SwinTransformer has the least parameters for both two data sets and the least MACs for the NTU-FI HAR data

TABLE I
EVALUATION OF VITS ON TWO DATA SETS

Model	UT-HAR			NTU-FI HAR		
	Accuracy (%)	Params	MACs	Accuracy (%)	Params	MACs
Vanilla ViT	96.74±0.1	10.58M	273.1M	90.31±0.94	834.53k	501.64M
SimpleViT	97.26±0.08	9.33M	232.82M	91±0.25	8.59M	8.15G
DeepViT	98.42±0.3	58.44M	192.22M	85±2.5	107.54M	2.69G
CaiT	98.78±0.19	1.16M	19.09M	98.2±0.5	2.3M	140M
SwinTransformer	96.58±0.38	84.09K	8.24M	87.5±2.5	202.81k	20.12M

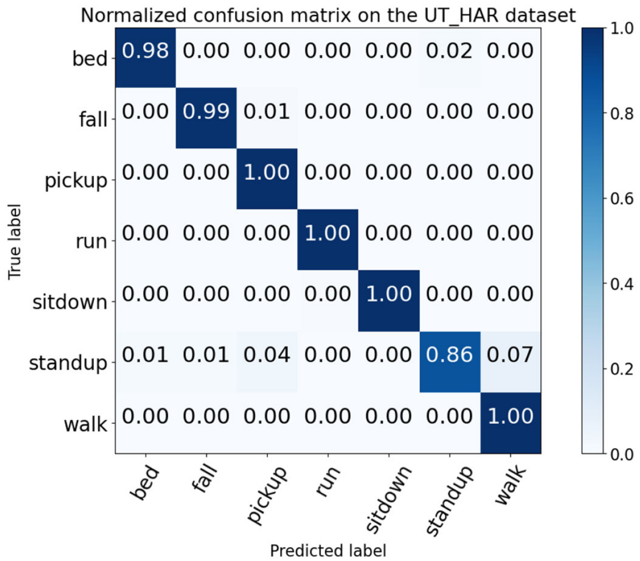


Fig. 7. Confusion matrix of CaiT on the data set UT-HAR data set.

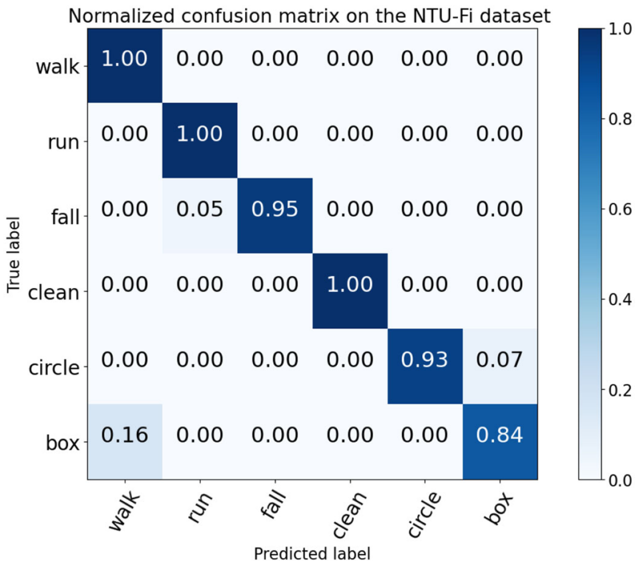


Fig. 8. Confusion matrix of CaiT on the data set NTU-FI HAR data set.

set; but the accuracy of SwinTransformer is almost the worst. Compared to CaiT, Vanilla ViT, SimpleViT, and DeepViT do not have advantages in the aspects of accuracy, model size, and computation efficiency. In short, CaiT has the best overall performance.

Since the CaiT achieves the best accuracy in two data sets, we illustrate the confusion matrices generated by CaiT on both data sets in Figs. 7 and 8. As can be seen, in the UT-HAR data set, the "Stand up" is the most difficult to recognize and has the lowest classification accuracy of 86%. In the NTU-FI HAR data set, the "box" is the hardest to classify and has only an accuracy of 84%. Most misclassified samples of the box are classified as the "walk".

C. Comparison to the Related Work

We implemented and compared five ViTs for WiFi CSI-based HAR. For presenting a complete research work, it is worth making a comparison with the latest related work. In Table II, we compared the CaiT implemented in our work with the related work on the UT-HAR data set. As can be seen, most work only provides the accuracy of HAR models while ignoring the investigation of model size and computation efficiency. In comparison, our CaiT achieved the best accuracy and its parameters and MACs are much less than the ResNet18 in [69]. In Table III, the accuracy of the CaiT on the NTU-Fi HAR data set is inferior to the work in [67] and [69], but its parameters and MACs are also much less than theirs. In general, the CaiT has superiority in accuracy compared to the related work, especially on the UT-HAR data set, it also can be considered as computation-efficient from the aspect of model size and MACs.

TABLE II
COMPARISON WITH THE RELATED WORK ON THE UT-HAR DATA SET

Model	UT-HAR		
	Accuracy (%)	Params	MACs
CaiT (ours)	98.78	1.16M	19.09M
Recurrent ConFormer [66]	96.16	0.5M	_
MLP [70]	98	_	_
Causal Net [71]	98.03	_	_
ResNet18 [69]	98.11	11.18M	99M
CIAM [72]	92.4	_	_

TABLE III
COMPARISON WITH THE RELATED WORK ON THE NTU-FI HAR DATA SET

Model	NTU-FI HAR		
	Accuracy (%)	Params	MACs
CaiT (ours)	98.2	2.3M	140M
MLP [69]	99.69	175.24M	350.48M
CIAM [72]	96.2	_	_
EfficientFi [67]	98.6	_	_

D. Analysis

From the evaluation, we can see that the different ViT performs differently on WiFi-based HAR. It is obvious that the architecture of ViTs has a great impact on their performance. We believe the introduction of CLS in

CaiT maximizes the information flow from patch embedding to class embedding, which results in CaiT achieving the best result in two WiFi-based HAR data sets. The performance of DeepViT follows the CaiT on the UT-HAR data set, which proves reattention mechanism can effectively improve the performance compared to the Vanilla ViT. The subtle modifications of SimpleViT improve a little performance. SwinTransformer performs poorly on both data sets, which indicates that the shifted window-based module in SwinTransformer is more suitable for high-resolution images but not for WiFi CSI data. Except for CaiT, the other four ViTs perform poorly on the NTU-FI HAR data set with large margins. It implies that the transformer architecture is not good at capturing frequency features since the NTU-FI data set is raw CSI data while the UT-HAR data set is CSI frequency spectrograms.

V. Conclusion

In this article, we implemented five ViTs for HAR via WiFi CSI. We analyzed the structural differences between the five ViTs and applied them to two benchmark WiFi CSI data sets. We evaluated and compared their performance from three aspects: accuracy, model size, and computation efficiency. Furthermore, we compared our work with the related work. We found that the CaiT achieved the SOTA on the UT-HAR data set and it also has advantages in model size and computation efficiency compared to the other four ViTs. As far as we are aware, our research is the first to comprehensively investigate the performance of different ViT architectures for WiFi CSI-based HAR. We believe our work can provide guidelines for the research of WiFi-based HAR modeling and model selection with the consideration of accuracy, model size, and computation efficiency. Our work also glimpses into the future painted by WiFi-based HAR, enriching the capabilities of ISAC and paving the way for NGMA networks that truly understand their users.

References

[1] M. S. Islam, M. K. A. Jannat, M. N. Hossain, W.-S. Kim, S.-W. Lee, and S.-H. Yang, "STC-NLSTMNet: An improved human activity recognition method using convolutional neural network with NLSTM from WiFi CSI," *Sensors*, vol. 23, no. 1, p. 356, 2022.

[2] A. Vijayvargiya, P. Singh, R. Kumar, and N. Dey, "Hardware implementation for lower limb surface EMG measurement and analysis using explainable AI for activity recognition," *IEEE Trans. Instrum. Meas.*, vol. 71, pp. 1–9, Aug. 2022.

[3] A. Klingenberg, V. Purrucker, W. Schüler, N. Ganapathy, N. Spicher, and T. M. Deserno, "Human activity recognition from textile electrocardiograms," in *Proc. 44th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, 2022, pp. 3434–3437.

[4] O. R. A. Almanifi, I. M. Khairuddin, M. A. M. Razman, R. M. Musa, and A. P. A. Majeed, "Human activity recognition based on wrist PPG via the ensemble method," *ICT Exp.*, vol. 8, no. 4, pp. 513–517, 2022.

[5] A. Li et al., "A contactless health monitoring system for vital signs monitoring, human activity recognition and tracking," *IEEE Internet Things J.*, early access, Nov. 23, 2023, doi: 10.1109/JIOT.2023.3336232.

[6] A. Li et al., "An integrated sensing and communication system for fall detection and recognition using ultra-wideband signals," *IEEE Internet Things J.*, vol. 11, no. 1, pp. 1509–1521, Jan. 2024.

[7] J. Liu, H. Liu, Y. Chen, Y. Wang, and C. Wang, "Wireless sensing for human activity: A survey," *IEEE Commun. Surveys Tuts.*, vol. 22, no. 3, pp. 1629–1645, 3rd Quart., 2020.

[8] H. Li, X. He, X. Chen, Y. Fang, and Q. Fang, "Wi-Motion: A robust human activity recognition using WiFi signals," *IEEE Access*, vol. 7, pp. 153287–153299, 2019.

[9] Y. Xie, Z. Li, and M. Li, "Precise power delay profiling with commodity WiFi," in *Proc. 21st Annu. Int. Conf. Mobile Comput. Netw.*, 2015, pp. 53–64.

[10] D. Halperin, W. Hu, A. Sheth, and D. Wetherall, "Tool release: Gathering 802.11n traces with channel state information," *ACM SIGCOMM Comput. Commun. Rev.*, vol. 41, no. 1, p. 53, 2011.

[11] X. Zhang, C. Tang, K. Yin, and Q. Ni, "WiFi-based cross-domain gesture recognition via modified prototypical networks," *IEEE Internet Things J.*, vol. 9, no. 11, pp. 8584–8596, Jun. 2022.

[12] G. Wang, Y. Zou, Z. Zhou, K. Wu, and L. M. Ni, "We can hear you with Wi-Fi!" in *Proc. 20th Annu. Int. Conf. Mobile Comput. Netw.*, 2014, pp. 593–604.

[13] H. Abdelnasser, K. A. Harras, and M. Youssef, "UbiBreathe: A ubiquitous non-invasive WiFi-based breathing estimator," in *Proc. 16th ACM Int. Symp. Mobile Ad Hoc Netw. Comput.*, 2015, pp. 277–286.

[14] S. M. Hernandez and E. Bulut, "Performing WiFi sensing with off-the-shelf smartphones," in *Proc. IEEE Int. Conf. Pervasive Comput. Commun. Workshops (PerCom Workshops)*, 2020, pp. 1–3.

[15] M. Schulz, D. Wegemer, and M. Hollick. "Nexmon: The C-based firmware patching framework." 2017. [Online]. Available: https://nexmon.org

[16] M. Yang, H. Zhu, R. Zhu, F. Wu, L. Yin, and Y. Yang, "WiTransformer: A novel robust gesture recognition sensing model with WiFi," *Sensors*, vol. 23, no. 5, p. 2612, 2023.

[17] Y. Gu and J. Li, "A novel WiFi gesture recognition method based on CNN-LSTM and channel attention," in *Proc. 3rd Int. Conf. Adv. Inf. Sci. Syst.*, 2021, pp. 1–4.

[18] N. Dua, S. N. Singh, and V. B. Semwal, "Multi-input CNN-GRU based human activity recognition using wearable sensors," *Computing*, vol. 103, pp. 1461–1478, Mar. 2021.

[19] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 10012–10022.

[20] D. Zhou et al., "DeepVit: Towards deeper vision transformer," 2021, *arXiv:2103.11886*.

[21] H. Touvron, M. Cord, A. Sablayrolles, G. Synnaeve, and H. Jégou, "Going deeper with image transformers," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 32–42.

[22] Y. Zhou, H. Huang, S. Yuan, H. Zou, L. Xie, and J. Yang, "MetaFi++: WiFi-enabled transformer-based human pose estimation for metaverse avatar simulation," *IEEE Internet Things J.*, vol. 10, no. 16, pp. 14128–14136, Aug. 2023.

[23] L. Beyer, X. Zhai, and A. Kolesnikov, "Better plain ViT baselines for ImageNet-1k," 2022, *arXiv:2205.01580*.

[24] S. Yousefi, H. Narui, S. Dayal, S. Ermon, and S. Valaee, "A survey on behavior recognition using WiFi channel state information," *IEEE Commun. Mag.*, vol. 55, no. 10, pp. 98–104, Oct. 2017.

[25] J. Yang et al., "SenseFi: A library and benchmark on deep-learning-empowered WiFi human sensing," *Patterns*, vol. 4, no. 3, 2023, Art. no. 100703.

[26] E. Oughton, G. Geraci, M. Polese, and V. Shah. "Prospective evaluation of next generation wireless broadband technologies: 6G versus Wi-Fi 7/8," SSRN. 2023. [Online]. Available: https://ssrn.com/abstract=4528119

[27] F. Li, M. A. A. Al-Qaness, Y. Zhang, B. Zhao, and X. Luan, "A robust and device-free system for the recognition and classification of elderly activities," *Sensors*, vol. 16, no. 12, p. 2043, 2016.

[28] Y. He, Y. Chen, Y. Hu, and B. Zeng, "WiFi vision: Sensing, recognition, and detection with commodity MIMO-OFDM WiFi," *IEEE Internet Things J.*, vol. 7, no. 9, pp. 8296–8317, Sep. 2020.

[29] X. Liu, J. Cao, S. Tang, and J. Wen, "Wi-Sleep: Contactless sleep monitoring via WiFi signals," in *Proc. IEEE Real-Time Syst. Symp.*, 2014, pp. 346–355.

[30] J. Liu, Y. Wang, Y. Chen, J. Yang, X. Chen, and J. Cheng, "Tracking vital signs during sleep leveraging off-the-shelf WiFi," in *Proc. 16th ACM Int. Symp. Mobile Ad Hoc Netw. Comput.*, 2015, pp. 267–276.

[31] K. Qian, C. Wu, Z. Yang, Y. Liu, and Z. Zhou, "PADS: Passive detection of moving targets with dynamic speed using PHY layer information," in *Proc. 20th IEEE Int. Conf. Parallel Distrib. Syst. (ICPADS)*, 2014, pp. 1–8.

[32] H. F. T. Ahmed, H. Ahmad, and C. Aravind, "Device free human gesture recognition using Wi-Fi CSI: A survey," *Eng. Appl. Artif. Intell.*, vol. 87, Jan. 2020, Art. no. 103281.

[33] W. Wang, A. X. Liu, M. Shahzad, K. Ling, and S. Lu, "Device-free human activity recognition using commercial WiFi devices," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 5, pp. 1118–1131, May 2017.

[34] X. Li et al., "IndoTrack: Device-free indoor human tracking with commodity Wi-Fi," *Proc. ACM Interact., Mobile, Wearable Ubiquitous Technol.*, vol. 1, no. 3, pp. 1–22, 2017.

[35] D. Zhang, F. Zhang, D. Wu, J. Xiong, and K. Niu, "Fresnel zone based theories for contactless sensing," *Contactless Human Activity Analysis*. Berlin, Germany: Springer, 2021, pp. 145–164.

[36] Y. Wang, K. Wu, and L. M. Ni, "WiFall: Device-free fall detection by wireless networks," *IEEE Trans. Mobile Comput.*, vol. 16, no. 2, pp. 581–594, Feb. 2017.

[37] K. Ali, A. X. Liu, W. Wang, and M. Shahzad, "Keystroke recognition using WiFi signals," in *Proc. 21st Annu. Int. Conf. Mobile Comput. Netw.*, 2015, pp. 90–102.

[38] W. He, K. Wu, Y. Zou, and Z. Ming, "WiG: WiFi-based gesture recognition system," in *Proc. 24th Int. Conf. Comput. Commun. Netw. (ICCCN)*, 2015, pp. 1–7.

[39] H. Li, W. Yang, J. Wang, Y. Xu, and L. Huang, "WiFinger: Talk to your smart devices with finger-grained gesture," in *Proc. ACM Int. Joint Conf. Pervasive Ubiquitous Comput.*, 2016, pp. 250–261.

[40] M. M. Najafabadi, F. Villanustre, T. M. Khoshgoftaar, N. Seliya, R. Wald, and E. Muharemagic, "Deep learning applications and challenges in big data analytics," *J. Big Data*, vol. 2, no. 1, pp. 1–21, 2015.

[41] Z. Zhang, Y. Xu, J. Yu, and S. Gao, "Saliency detection in 360 videos," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 488–503.

[42] T. Nakamura, M. Bouazizi, K. Yamamoto, and T. Ohtsuki, "Wi-Fi-CSI-based fall detection by spectrogram analysis with CNN," in *Proc. IEEE Global Commun. Conf.*, 2020, pp. 1–6.

[43] X. Shen, Z. Ni, L. Liu, J. Yang, and K. Ahmed, "WiPass: 1D-CNN-based smartphone keystroke recognition using WiFi signals," *Pervasive Mobile Comput.*, vol. 73, Jun. 2021, Art. no. 101393.

[44] L. Jia, Y. Gu, K. Cheng, H. Yan, and F. Ren, "BeAware: Convolutional neural network (CNN) based user behavior understanding through WiFi channel state information," *Neurocomputing*, vol. 397, pp. 457–463, Jul. 2020.

[45] A. H. Ribeiro, K. Tiels, L. A. Aguirre, and T. Schön, "Beyond exploding and vanishing gradients: Analysing RNN training using attractors and smoothness," in *Proc. Int. Conf. Artif. Intell. Statist.*, 2020, pp. 2370–2380.

[46] J. Zhang et al., "Data augmentation and dense-LSTM for human activity recognition using WiFi signal," *IEEE Internet Things J.*, vol. 8, no. 6, pp. 4628–4641, Mar. 2021.

[47] X. Ming, H. Feng, Q. Bu, J. Zhang, G. Yang, and T. Zhang, "HumanFi: WiFi-based human identification using recurrent neural network," in *Proc. IEEE SmartWorld, Ubiquitous Intell. Comput., Adv. Trusted Comput., Scalable Comput. Commun., Cloud Big Data Comput., Internet People Smart City Innov. (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI)*, 2019, pp. 640–647.

[48] J. Ding and Y. Wang, "WiFi CSI-based human activity recognition using deep recurrent neural network," *IEEE Access*, vol. 7, pp. 174257–174269, 2019.

[49] L. Guo et al., "Towards CSI-based diversity activity recognition via LSTM-CNN encoder-decoder neural network," *Neurocomputing*, vol. 444, pp. 260–273, Jul. 2021.

[50] Y. Zhang, Q. Liu, Y. Wang, and G. Yu, "CSI-based location-independent human activity recognition using feature fusion," *IEEE Trans. Instrum. Meas.*, vol. 71, pp. 1–12, Oct. 2022.

[51] J. Li, L. Yao, B. Li, X. Wang, and C. Sammut, "Multi-agent transformer networks for multimodal human activity recognition," in *Proc. 31st ACM Int. Conf. Inf. Knowl. Manage.*, 2022, pp. 1135–1145.

[52] S. Sowmiya and D. Menaka, "Transformer model for human activity recognition using IoT wearables," in *Proc. Int. Conf. Comput. Vis., High-Perform. Comput., Smart Devices, Netw.*, 2022, pp. 287–300.

[53] J. Wensel, H. Ullah, and A. Munir, "ViT-ReT: Vision and recurrent transformer neural networks for human activity recognition in videos," *IEEE Access*, vol. 11, pp. 72227–72249, 2023.

[54] B. Li, W. Cui, W. Wang, L. Zhang, Z. Chen, and M. Wu, "Two-stream convolution augmented transformer for human activity recognition," in *Proc. AAAI Conf. Artif. Intell.*, vol. 35, 2021, pp. 286–293.

[55] M. Abdel-Basset, H. Hawash, N. Moustafa, and N. Mohammad, "H2HI-Net: A dual-branch network for recognizing human-to-human interactions from channel-state information," *IEEE Internet Things J.*, vol. 9, no. 12, pp. 10010–10021, Jun. 2022.

[56] Y. Yao, C. Zhao, Y. Pan, C. Sha, Y. Rao, and T. Wang, "Human gesture recognition based on CT-A hybrid deep learning model in WiFi environment," *IEEE Sensors J.*, vol. 23, no. 22, pp. 28021–28034, Nov. 2023.

[57] M. Cominelli, F. Gringoli, and F. Restuccia, "Exposing the CSI: A systematic investigation of CSI-based Wi-Fi sensing capabilities and limitations," in *Proc. IEEE Int. Conf. Pervasive Comput. Commun. (PerCom)*, 2023, pp. 81–90.

[58] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.

[59] E. D. Cubuk, B. Zoph, J. Shlens, and Q. V. Le, "RandAugment: Practical data augmentation with no separate search," 2019, *arXiv:1909.13719*.

[60] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "Mixup: Beyond empirical risk minimization," 2017, *arXiv:1710.09412*.

[61] G. Huang, Y. Sun, Z. Liu, D. Sedra, and K. Q. Weinberger, "Deep networks with stochastic depth," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 646–661.

[62] Y. Liu, S. Mai, X. Chen, C.-J. Hsieh, and Y. You, "Towards efficient and scalable sharpness-aware minimization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 12360–12370.

[63] S. Yun, D. Han, S. J. Oh, S. Chun, J. Choe, and Y. Yoo, "CutMix: Regularization strategy to train strong classifiers with localizable features," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 6023–6032.

[64] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 10347–10357.

[65] V. Sovrasov. "Ptflops: A flops counting tool for neural networks in pyTorch framework." 2022. [Online]. Available: https://github.com/sovrasov/flops-counter.pytorch

[66] M. Shang and X. Hong, "Recurrent ConFormer for WiFi activity recognition," *IEEE/CAA J. Automatica Sinica*, vol. 10, no. 6, pp. 1491–1493, Jun. 2023.

[67] 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 Things J.*, vol. 9, no. 15, pp. 13086–13095, Aug. 2022.

[68] L. Prechelt, "Early stopping—But when?" in *Neural Networks: Tricks of the Trade*. Berlin, Germany: Springer, 2002, pp. 55–69.

[69] J. Yang et al., "Deep learning and its applications to WiFi human sensing: A benchmark and a tutorial," 2022, *arXiv:2207.07859*.

[70] D. Butyrskaya, "Human movement recognition using deep learning on visualized CSI Wi-Fi data," M.S. thesis, Eng. Data-Intensive Intell. Softw. Syst., Univ. L'Aquila, L'Aquila, Italy, 2023.

[71] K. Xu, J. Wang, H. Zhu, and D. Zheng, "Self-supervised learning for WiFi CSI-based human activity recognition: A systematic study," 2023, *arXiv:2308.02412*.

[72] Q. Zhou, S. Wu, C. Jiang, R. Zhang, and X. Jing, "Cloud-edge–terminal collaboration enabled device-free sensing under class-imbalance conditions," *IEEE Internet Things J.*, vol. 11, no. 4, pp. 5980–5992, Feb. 2024.

Fei Luo received the B.Sc. degree from Jiangxi University of Science and Technology, Ganzhou, China, in 2012, the M.Sc. degree in surveying and mapping from Wuhan University, Wuhan, China, in 2016, and the Ph.D. degree from the Queen Mary University of London, London, U.K., in 2020.

He is currently working as a Postdoctoral Fellow with Shenzhen University, Shenzhen, China, and also with Great Bay University, Dongguan, China. His research interests include geographic information systems, human activity detection, and machine learning.

Salabat Khan (Member, IEEE) received the Ph.D. from the School of Computer Science and Technology, Beijing Institute of Technology, Beijing, China.

He is a Postdoctoral Fellow with the College of Computer Science and Software Engineering, Shenzhen University, Shenzhen, China, and also with with Qilu Institute of Technology, Jinan, China. His current research interest includes secure and transparent public-key infrastructure, transparent and secure key management in VANETs, distributed ledger technology, and blockchain.

Bin Jiang (Member, IEEE), photograph and biography not available at the time of publication.

Kaishun Wu (Fellow, IEEE) received the Ph.D. degree in computer science and engineering from Hong Kong University of Science and Technology (HKUST), Hong Kong, in 2011.

After that, he worked as a Research Assistant Professor with HKUST. In 2013, he joined Shenzhen University, Shenzhen, China, as a Distinguished Professor. He has co-authored two books and published over 100 high quality research papers in international leading journals and primer conferences, like IEEE TRANSACTIONS ON MOBILE COMPUTING, IEEE TRANSACTIONS ON PARALLEL AND DISTRIBUTED SYSTEMS, ACM MobiCom, and IEEE INFOCOM. He is the inventor of six U.S. and over 100 Chinese pending patents.

Dr. Wu received the 2012 Hong Kong Young Scientist Award, the 2014 Hong Kong ICT Awards: Best Innovation, and the 2014 IEEE ComSoc Asia–Pacific Outstanding Young Researcher Award.