Advances in Computational Design, *Vol. 9, No. 3 (2024) 187-211*
https://doi.org/10.12989/acd.2024.9.3.187

CSI-based human activity recognition via lightweight compact convolutional transformers

Fahd Saad Abuhoureyah[1], Yan Chiew Wong*[1],
Malik Hasan Al-Taweel[1,2] and Nihad Ibrahim Abdullah[3]

*[1]Centre for Telecommunication Research and Innovation (CeTRI)
Fakulti Teknologi dan Kejuruteraan Elektronik dan. Komputer (FTKEK),
Universiti Teknikal Malaysia Melaka (UTeM), 76100 Durian Tunggal, Melaka, Malaysia
[2]Department of Communications Engineering, College of Engineering,
University of Diyala, Baqubah, Diyala, Iraq
[3]Computer Science, Sulaimani Polytechnic University, Sulaimani, KRG Iraq*

(Received March 13, 2024, Revised June 30, 2024, Accepted September 11, 2024)

Abstract. WiFi sensing integration enables non-intrusive and is utilized in applications like Human Activity Recognition (HAR) to leverage Multiple Input Multiple Output (MIMO) systems and Channel State Information (CSI) data for accurate signal monitoring in different fields, such as smart environments. The complexity of extracting relevant features from CSI data poses computational bottlenecks, hindering real-time recognition and limiting deployment on resource-constrained devices. The existing methods sacrifice accuracy for computational efficiency or vice versa, compromising the reliability of activity recognition within pervasive environments. The lightweight Compact Convolutional Transformer (CCT) algorithm proposed in this work offers a solution by streamlining the process of leveraging CSI data for activity recognition in such complex data. By leveraging the strengths of both CNNs and transformer models, the CCT algorithm achieves state-of-the-art accuracy on various benchmarks, emphasizing its excellence over traditional algorithms. The model matches convolutional networks' computational efficiency with transformers' modeling capabilities. The evaluation process of the proposed model utilizes self-collected dataset for CSI WiFi signals with few daily activities. The results demonstrate the improvement achieved by using CCT in real-time activity recognition, as well as the ability to operate on devices and networks with limited computational resources.

Keywords: activity recognition; channel state information; compact convolutional transformer; WiFi sensing

1. Introduction

Recent years have witnessed the widespread utilization of WiFi devices in indoor settings due to their cost-effectiveness and ease of deployment. This increased adoption of WiFi devices has prompted the exploration of Human Activity Recognition (HAR) applications across various domains, such as smart home environments (Guo *et al*. 2018), medical monitoring systems (Palanivelu and Srinivasan 2020), and public security initiatives, among others. Traditional

*Corresponding author, Ph.D., E-mail: ycwong@utem.edu.my

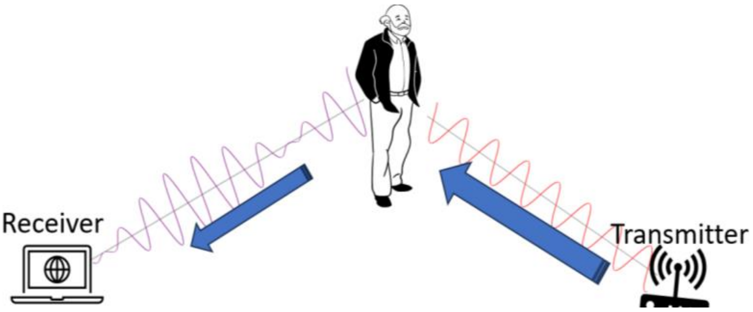


Fig. 1 The WiFi sensing principle

approaches for WiFi-based activity recognition rely on handcrafted features and traditional machine learning algorithms (Yang *et al*. 2022a). However, these methods struggle to capture complex patterns and long-range dependencies in the CSI data (Zhou *et al*. 2021). By introducing lightweight Compact Convolutional Transformers CCT into CSI-based HAR, there is an opportunity to overcome these challenges. The lightweight self-attention mechanism in CCT captures the long-range dependencies and complex relationships present in the CSI data.

By analyzing the patterns of WiFi signal variations, it becomes possible to detect movements and infer human activities within a given space (Zhang *et al*. 2021). Fig. 1 illustrates how changes in WiFi signal strength and reflections caused by movements are used to detect activities such as walking, gestures, or breathing (Abuhoureyah *et al*. 2024). The machine learning techniques enable the analysis of signals in the context of image processing for recognition by leveraging their ability to learn hierarchical representations (Yang *et al*. 2022b). These techniques enable the application of image-processing models for tasks such as object recognition, localization, detection, and pattern identification.

Deep learning models can learn and extract intricate features from raw CSI data. Likewise, CNNs, which are known for their hierarchical feature extraction through convolutional layers, excel at capturing local patterns and have dominated image recognition tasks (Ma *et al*. 2019). In contrast, Vision Transformer ViTs leverage the self-attention mechanism to process image data, enabling them to capture global context information (Lu *et al*. 2022). ViTs also exhibit strong performance in various vision tasks, often with fewer parameters, which is beneficial for resource-constrained applications (Dierickx *et al*. 2023). However, CNNs have a proven track record and a substantial body of pre-trained models, making them more accessible for many practical use cases. Nevertheless, ViTs are data-hungry and expensive, which limits their applicability in specific scenarios.

The CCT represents an integration in bridging the gap between ViTs and CNNs (Dierickx *et al*. 2023). CCT combines the efficiency of CNNs in capturing local visual patterns with the modeling capabilities of transformers for capturing global context. The hybrid approach retains the benefits of both architectures while mitigating their limitations. CCT incorporates compact self-attention mechanisms, enabling it to process images and making it more accessible than pure ViTs. The proposed lightweight CCT model has the potential to enhance the deployment of CSI-based sensing on resource-constrained devices. Therefore, this work aims to develop a lightweight CCT for CSI HAR applications through pruning techniques, efficient architecture design, and experimentation with model depth and width. The study tests how well the lightweight CCT model does in CSI-based HAR tasks, comparing it to other methods.

Table 1 CSI-based methods for activity classification for human activity recognition

Ref	Algorithm	Activity Attributes	Method	Preprocessing	Accuracy	Challenges
(Ahmed Ouameur *et al.* 2020)	Multilayer MNN	Localization	RSS	Fingerprints acquisition	66% to 80%	MNN suffers from training Complexity and Data Dependency
(Ahmed *et al.* 2020)	CNN and SVN/KNN /RF/NB	Gesture recognition	CSI	Convolutional filters	Up to 98%	Complexity variance, data efficiency of feature extraction
(Zhang *et al.* 2021)	CNN	HAR	CSI	Quasi-static offsets Convolutional filters	92.7%	Spatial resolution constraints, computational intensity, and susceptibility to signal variations
(Schäfer *et al.* 2021)	LSTM	HAR	CSI	sequential learning	96.9%	Computationally expensive
(Jiang *et al.* 2020)	GAN	Gesture recognition	CSI	Deep Convolution Feature Extraction	95.6	Computationally expensive
(X. Ding *et al.* 2021)	Meta-Learning	HAR	CSI	CNN-LSTM	Up to 99%	Encounters scalability issues and resource demands
(Showmik *et al.* 2023)		HAR	CSI	CNN	Avg 94%	Suffers Over-fitting

This work uses the CCT algorithm for CSI-based classification to perform classification tasks. The first section introduces the CCT algorithm, while the second section examines related works in image processing and CSI techniques. Section three describes the mathematical representation of CSI sensing, as well as the image classifiers used in WiFi-based sensing to establish the HAR model. Section four describes in detail the lightweight CCT algorithm, outlining its design and methodology. The fifth section evaluates the proposed method and discusses the results. Lastly, Section six concludes the work by summarizing the findings and outlining future research directions.

2. Related works

The utilization of CSI in HAR holds promise for enhancing activity recognition models' robustness. By taking advantage of the fact that WiFi networks are available, CSI-based HAR systems use the unique properties of wireless signals to make activity recognition models accurate and durable. Recent research has improved different structures of algorithms such as CNNs, Recurrent Neural Network RNNs (Abuhoureyah *et al.* 2023). These algorithms have the capabilities to focus on the intricate relationship between signal characteristics and model structures in the learning process (Luo *et al.* 2022). The integration incorporates poor-quality temporal and frequency information into activity recognition (Ding and Wang 2019). Moreover, WiFi activity detection systems scrutinize the received data using classical training models like the hidden Markov model (HMM) (Tiku *et al*. 2020) and the KNN.

Recent studies have investigated the application of CNNs in HAR, highlighting the advantages of CNN-based approaches in extracting temporal features from sensor data (Ding *et al*. 2021). Additionally, these networks excel in domains that seek local inductive bias. However, they need to catch up in capturing long-range interdependencies where transformers excel at using available data. The reliance on extensive datasets constrains utility in scientific and signal-sensing domains, given the formidable challenges associated with dataset acquisition. Furthermore, the often-required large computational resources limit the ability to reproduce and access machine learning techniques. While CNNs have made contributions to HAR, addressing these limitations remains a subject of ongoing research, with newer models and techniques seeking to enhance their robustness and adaptability in HAR scenarios (Showmik *et al*. 2023). Table 1 summarizes some of the procedural approaches utilized in HAR through WLAN, which are delineated in Table 1.

While utilizing similar datasets for CIS-based activity recognition, the varying algorithms in Table 1 exhibit differences in accuracy and computation. Although some methods achieve high accuracy rates, they suffer from computational intensity or scalability issues, limiting their practical deployment in real-time, pervasive systems. However, the lightweight CCT algorithm, which processes CSI data with balanced computational efficiency, holds promise for improving WiFi-based sensing classification accuracy. Its streamlined approach addresses the challenges of previous methodologies, enabling more precise and efficient activity recognition within a pervasive system.

On the other side, transformers have witnessed a meteoric rise in popularity in machine learning research, following the introduction of the “Attention is All You Need” paper (Vaswani, n.d.). While these models were first used for natural language processing, they have now been extended to computer vision through the Vision Transformer (ViT) framework. This framework shows how pure transformer backbones can change things (Lu *et al*. 2022). Transformers’ development underscored both the power of such models and the supremacy of large-scale training over inductive biases. ViTs have harnessed the transformative potential of the Transformer architecture to enhance image classification tasks. In contrast to traditional CNNs, ViTs adopt a sequence-based approach to processing images by dividing them into non-overlapping patches, treating each patch as a token (Lu *et al*. 2022).

Recent work by Li *et al*. (2023) presents a novel solution to capturing rich and long-range semantic concepts in artworks using CNN-based style transfer methods. The authors introduce a compact transformer architecture called AdaFormer, which reduces the model scale by 20% compared to state-of-the-art transformers for style transfer. Additionally, they explore adaptive style transfer by allowing the content to select the detailed style element automatically, resulting in an appealing and reasonable output. Lee *et al*. (2023) introduces a novel architectural framework tailored for embedded systems, exhibiting superior performance in monocular depth estimation compared to existing methodologies. The authors carefully divide popular approaches using the Transformer paradigm into two groups: those that try to mimic the attention mechanism and those that support structures that combine CNN and Transformer elements.

In this work, we implement a lightweight, CCT that achieves over 96% accuracy on HAR for the dataset collected with minimum samples. We further enhance the system by integrating convolutional blocks into the tokenization process, thereby creating the CCT. These additions boost performance, culminating in a top-1% accuracy of 98%. The proposed model also outperforms most comparable CNN-based algorithms in this domain, showcasing its scalability and computational efficiency.

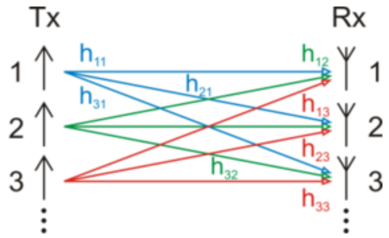


Fig. 2 Simplified structure of MIMO System

3. Preliminary

3.1 Channel state information

Within wireless communication, CSI serves as a source of information, offering essential insights into the amplitude and phase characteristics of distinct subcarriers. Multiple-input and Multiple-Output (MIMO) technology has gained widespread utilization to enhance data throughput and extend signal propagation range without enlarging bandwidth or augmenting transmission power, often involving deploying multiple antennas, as shown in Fig. 2 (Akhtar and Wang 2020).

In MIMO systems, the channel is represented by a complex coefficient for each antenna pair, forming a CSI matrix, as shown in Eq. (1). CSI matrix conveys amplitude and phase information for OFDM subcarriers in the WiFi protocol’s physical layer. WiFi standards like 802.11 a/g/b/n/ac are employed in virtual WiFi routers, offering higher data rates through MIMO and OFDM. These standards operate on 56, 114, and 232 subcarriers, facilitating various bandwidths (20MHz, 40MHz, and 80MHz) at either 2.4GHz or 5GHz. The mathematical representation of CSI is denoted as y and x, where y signifies the received signal, x denotes the transmitted signal, and their relationship relies on the CSI matrix data formatted in the frequency domain via OFDM (Li *et al*. 2021, Yang *et al*. 2013).

$$y = Hx + n \tag{1}$$

Whereby, Hx is a convolution matrix formed between received and transmitted signals r*t established by number of transmitting and receiving antennas represented in polar form at Eq. (2) (Muaaz *et al*. 2022).

$$H_{ij}^{s} = \|H_{ij}^{s}\| e^{jLH_{ij}^{x}} \qquad s \in [1, N_s], i \in [1, N_t], j \in [1, N_r] \tag{2}$$

The generated matrix H amplitude H_{ij}^{s} and $\angle H_{ij}^{s}$ denote of the matrix phase. Whereas Nt and Nr represent the numeral of antennas at the transmitter (TX) and receiver (RX). Likewise, i, j stands for the index of TX and RX antennas and Ns represents the subcarriers for each pair of transceiver antenna. Furthermore, for an activity structure matrix H of the CSI data at boundary t is expressed by Eq. (3).

$$H^{t} = \left(H_{ij}^{s}\right)^{t} = \begin{bmatrix} and(H_{11}^{s})^{t} & andand(H_{12}^{s})^{t} & andand(H_{13}^{s})^{t} \\ and(H_{21}^{s})^{t} & andand(H_{22}^{s})^{t} & andand(H_{23}^{s})^{t} \\ and(H_{31}^{s})^{t} & andand(H_{32}^{s})^{t} & andand(H_{33}^{s})^{t} \end{bmatrix} \tag{3}$$

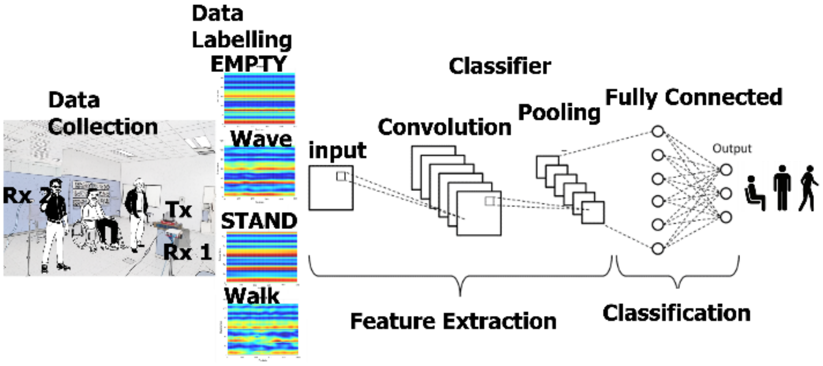


Fig. 3 A schematic of the CNN framework highlights the sequential flow of operations.

Since CSI is affected by environment, the channel impulse response has a multipath impact on the wireless channel (CIR) (Zhang *et al*. 2021). Under the premise of linear time invariance, the channel impulse representation of the response is represented in Eq. 4.

$$H(k) = \| H(k) \| e^{j\angle H(k)} \tag{4}$$

The matrix H(k) represents the CSI data for the k-th phase of the k-th subcarrier, and H(k) represents the amplitude of the k-th subcarrier (Sharma *et al*. 2021). The channel property of a communication link in wireless communication is known as CSI.

3.2 Image classifier

The preliminary phase of the CCT framework establishes the groundwork for subsequent stages, thereby enabling precise and efficient HAR through WiFi sensing. In the context of CSI-based HAR utilizing WiFi sensing, the initial step in the CCT involves the integration of a CNN. As depicted in Fig. 3, the process commences with the convolutional layers of the CNN, extracting hierarchical spatial features from CSI data. The Fig. visualizes CNN processing CSI data through convolutional filters, discerning essential patterns associated with human activities. CNNs, as a class of deep learning models, operate by learning spatial hierarchies of features from input images (Saw and Wong 2023). They identify basic features in initial layers and recognize more complex patterns in subsequent layers, employing pooling layers to reduce spatial dimensions. Fully connected layers at the network's end use learned features for classification or predictions. CNNs, with their hierarchical and localized learning approach, prove effective tasks such as image recognition, object detection, and image classification (Pan *et al*. 2019).

The framework underscores the ViT's efficacy in capturing global contextual information, complementing the local features acquired by CNNs, thus enhancing the overall capability of the CCT framework for HAR in WiFi sensing applications. Fig. 4 shows the process of the ViT transforming images by breaking them into patches. Initially, an image is divided into fixed patches, and each patch is embedded to form a linearly sequence. These patches, along with positional information, undergo self-attention processes in transformer encoder layers, enabling the network to capture global contextual information. CCT enhances CNNs by focusing on both local and global features. The ViT- framework is integrated into CCT, where CNNs and ViT collaborate to capture local and global features from CSI data.

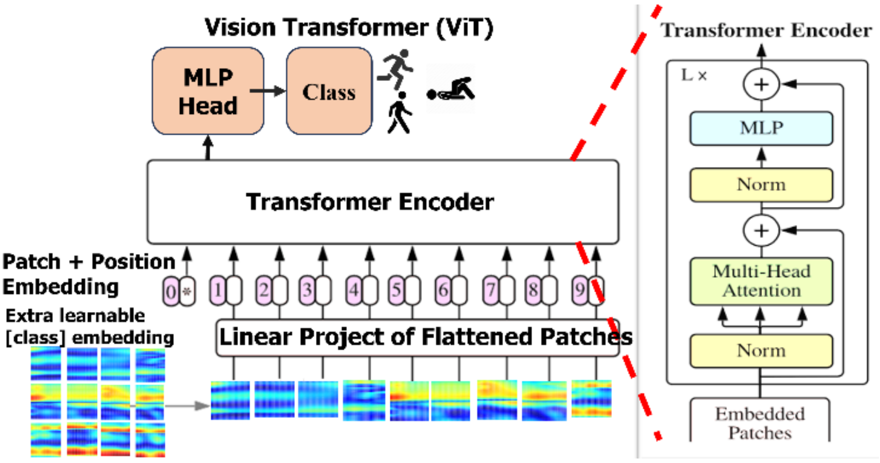


Fig. 4 Schematic representation of the Vision Transformer (ViT)

The Swin Transformer architecture demonstrates high developments performance across various vision tasks, surpassing preceding models in accuracy. Comparative assessments with transformer-based predecessors, such as DeiT, reveal that the Swin transformer achieves better accuracy while maintaining a comparable computational cost. Moreover, unlike ConvNet-based models like RegNet and EfficientNet, the Swin Transformer exhibits speed-accuracy trade-off, further solidifying its standing as a choice schematic in the landscape of computer vision tasks. An attribute contributing to Swin Transformer's efficacy is its hierarchical architecture, enabling nuanced design at different scales capturing local and global features. Additionally, the shifted windowing scheme within the Swin Transformer mitigates the computational complexity of self-attention computations, establishing linear computational scaling concerning image size.

4. Lightweight compact convolutional transformers

CCTs are a class of neural network architectures that blend the strengths of both CNNs and ViTs. They are designed to handle both spatial and token-based information within images (Dierickx *et al*. 2023). In a traditional CNN, convolutional layers are used to capture local spatial patterns in the input data. Let's denote the input feature map as $X \in \mathbb{R}^{H\times \mathrm{W} \times \mathrm{C}}$, H and W are the height and width of the feature map, and C is the number of channels. A convolution operation is represented as Y=X∗K where Y is the output feature map, K is the convolutional kernel, and $'*'$ denotes the convolution operation. The size and stride of the kernel, as well as the padding, determine the spatial dimensions of the output feature map.

To describe CCT, consider a hybrid architecture that combines these components (Li *et al*. 2023). The initial point starts by input image I∈ $\mathbb{R}^{C\times H\times \mathrm{W}}$ in the Convolutional Layers CCT which begins with a series of process to classify the input image Convi that represents the i-th convolutional layer. These layers extract local features from the images shown in Eq. (5).

$$\begin{gathered} X_1 = \mathrm{Conv}_1\,(I) \\ X_2 = \mathrm{Conv}_2(X_1) \end{gathered} \tag{5}$$

$$\vdots$$

$$X_n = \mathrm{Conv}_n(X_{n-1})$$

The CCT introduces transformer blocks to capture global and long-range dependencies after the initial convolutional layers using transformer blocks (Dierickx *et al.* 2023). Each transformer block consists of a self-attention mechanism and feedforward layers represented in Eq. (6).

$$\begin{gathered} Y_1 = \mathrm{Transformer}_1\ (X_n) \\ Y_2 = \mathrm{Transformer}_2\ (Y_1) \\ \vdots \\ Y_m = \mathrm{Transformer}_m\ (Y_{m-1}) \end{gathered} \tag{6}$$

CCTs present a paradigm shift in computer vision feilid and offers advantages with conventional CNNs and other established methodologies for image classification (Dierickx *et al.* 2023). Furthermore, transformer represents the i-th blocks allow the model to capture relationships between features across the spatial hierarchy. The choice of hyperparameters, such as the number of convolutional layers, transformer blocks, kernel sizes, and attention mechanisms, depends on the specific architecture and task requirements.

The attribute of CCTs resides in their ability to long-range dependencies within input data, a characteristic by integrating the transformer's self-attention mechanism. While CNNs are architected for the localized extraction of features, CCTs transcend these confines by forging connections between distant pixels or regions within an image, manifesting an enhanced capacity to discern global contextual cues. This trait assumes paramount in tasks characterized by object relationships that traverse spatial extents, such as the intricate domains of semantic segmentation.

The ViT enhances CNNs by reformulating image inputs into sequences of embedded patches. The mathematical expression demonstrated as $X \in \mathbb{R}^{N\times C}$, where N is the number of patches, and C is the dimension of each patch (Dosovitskiy *et al.* 2021). To preserve spatial information, positional embeddings $Xpos \in \mathbb{R}^{N\times C}$ are added. The self-attention mechanism in the transformer encoder layers is expressed as $\mathrm{Attention}\,(Q,K,V) = \mathrm{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$, where $Q,\ K$, and V query, key, and value matrices, and d_k is the dimension of the key vectors. Utilizing multiple self-attention heads in parallel, the multi-head attention is given by $\mathrm{MultiHead}\,(Q,K,V) = \mathrm{Concat}\,(\mathrm{head}_1,\ \dots,\ \mathrm{head}_h)W^O$, with $\mathrm{head}_i = \mathrm{Attention}\left(QW_i^Q,\ KW_i^K,\ VW_i^V\right)$, and $W_i^Q,\ W_i^K,\ W_i^V,\ W^O$ being learnable weight matrices. The feedforward layer operation is expressed as $\mathrm{FFN}\,(x) = \mathrm{ReLU}\,(xW_1 + b_1)W_2 + b_2$, where $W_1,\ b_1,\ W_2,\ b_2$ are learnable parameters.

4.1 System overview

The CCT architecture represents the system capability to various computer vision tasks. It achieves this by convolutional and transformer-based neural network components. At the inception of the CCT architecture lies a set of convolutional layers. These convolutional layers serve as the initial processing step to extract local features from the input data. Convolution operations, a fundamental component of these layers, involve the application of small kernels across the input data to capture intricate spatial hierarchies. Convolution is defined as defined in Eq. (7).

$$Y(i,j) = (X * K)(i,j) = \sum_{u,v} X(i-u,j-v)\cdot K(u,v) \tag{7}$$

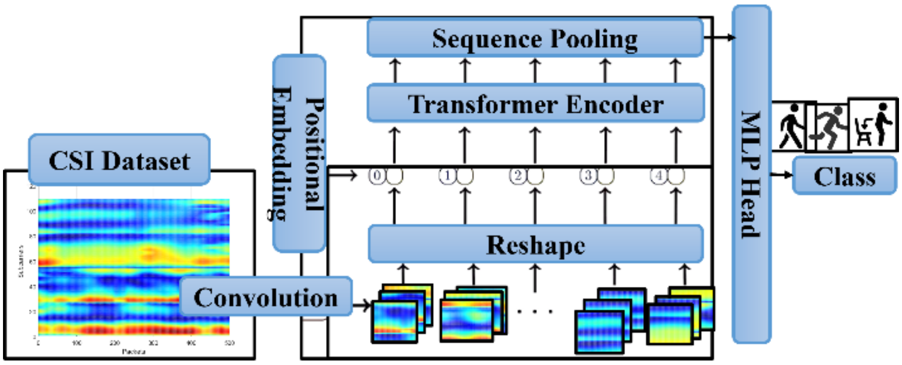


Fig. 5 An overview of the proposed CCT architecture

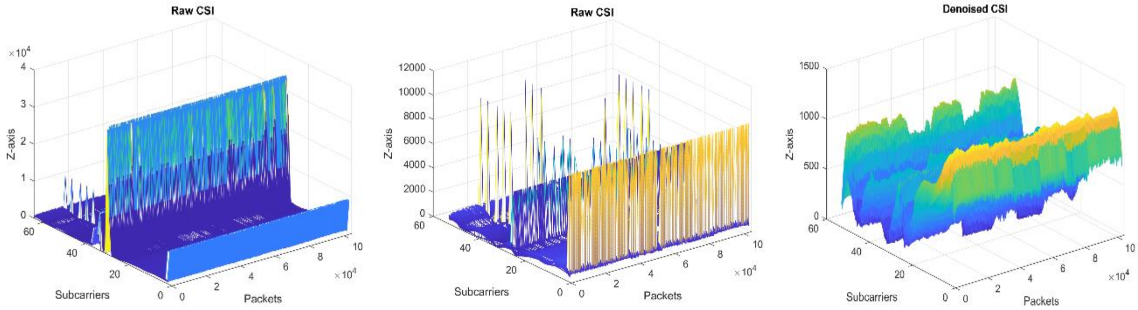
Y(i,j) represents the output feature at position (i, j) and X denotes the input data, and K is the convolutional kernel. Following the convolution layers, the architecture proceeds to a stage of reshaping. This step transforms the output from the convolutional layers into a suitable format for further processing by a encoder. Reshaping plays an important role between the convolutional and transformer components, facilitating the smooth flow of information.

Similarly, in the context of computer vision, this translates to capturing global context information within the image. The mathematical foundation of the self-attention mechanism in transformers involves computing weighted sums of query, key, and value matrices. The mechanism enables assigning varying degrees of importance to different elements in the input sequence, facilitating the capture of global relationships. After the Transformer encoder, the CCT architecture employs a sequence pooling layer. Sequence pooling serves the role of aggregating information from the transformed data to create a compact representation. Various pooling techniques are applied, such as max-pooling or mean-pooling, to capture essential features and reduce the spatial dimensionality of the data.

Eventually, the architecture concludes the process with an MLP (Multi-Layer Perceptron) head, which is often followed by a classification layer. MLP stage enables to achieve the fine-tuning and adapting of the learned features to specific downstream tasks, such as image classification. The MLP head comprises several fully connected layers to enable the model to learn complex patterns and representations from the pooled data.

4.2 Preprocessing

The denoising process of CSI datasets commenced with the initial step of eliminating null and pilot subcarriers. The data-denoising approach was employed to enhance the quality of the input dataset. The process consisted of multiple sequential steps, each with a distinct mathematical operation tailored to reduce noise and outliers. A two-dimensional median filter was applied to the data using the function with a specified neighborhood. In its local neighborhood, the operation replaced each value with the median value. Subsequently, the filling out function was employed to replace outliers with central tendency values by the median function. Furthermore, a moving median filter was implemented with a move median function for the data with a non-overlapping window size of one. Eq. (8) shows the representation, which involves calculating the median of data within a sliding window.



(a) CSI raw data

(b) Removing Null and Pillot subcarriers

CSI after filtering

Fig. 6 Filtering preprocess workflow for wireless signal data, commencing with removing null and pilot subcarriers.

$$\text{output }(x, y) = \text{median }(\text{neighborhood }(x, y))$$
$$\text{output }(t) = \text{median }(\text{window }(t)) \tag{8}$$
$$\text{output}(t) = \text{median}(\text{window}(t))$$

Fig. 6 illustrates the denoising process applied to the signal, involving the removal of null and pilot subcarriers, mitigating noise interference, and enhancing signal clarity.

4.3 Network architecture

The initial stage of the CCT architecture involves convolution layers. Using convolutional operations to find local patterns and characteristics in the image data, convolution layers are important for extracting features from the input data. In the CCT, the tokenizer stage forms an all-convolutional mini-network with a variable number of layers. The purpose is to generate adaptable depth image patches. The next step is customization based on the model's configuration, ensuring flexibility across different tasks and datasets. The CCT Tokenizer is part of the process of preparing the input data for subsequent stages and laying the groundwork for feature extraction. The variability in depth caters to the complexities present in various tasks. Table 2 provides a summary of the convolution module's network architecture and the output shape that aligns with its specific task, such as image classification.

4.4 Convolutional model

The convolutional module breaks down the two-dimensional operator into separate one-dimensional temporal and spatial convolutional (Dierickx *et al*. 2023, Dosovitskiy *et al*. 2021). The proposed design employs 'k' kernels in the initial layer of the module, characterized by dimensions of (1, 25), while implementing a stride of (1, 1). The configuration enables convolution operations to be executed along the temporal dimension, capturing temporal nuances inherent to CSI data (Li *et al*. 2023). The second layer 'k' kernels employed with dimensions of (ch, 1), where 'ch' represents the number of channels relevant to CSI signals. Furthermore, the design incorporates batch normalization techniques to scale the training process and counteract overfitting.

Table 2 Network architecture of the convolution module

Component	Number of Layers	Description
Convolution Layers	2	Initial layers for local feature extraction from the input data.
CCT Tokenizer	Variable	All-convolution mini network to produce image patches.
Positional Embedding	-	Optional positional embeddings for sequences.
Stochastic Depth	Variable	Regularization technique applied to transformer blocks.
MLP for Transformers Encoder	Variable	Multi-layer perceptron for the Transformers encoder.
Data Augmentation	-	Geometric data augmentations, including random cropping and flipping.
Attention Pooling	-	Weighted pooling of Transformer encoder outputs for classification.
CCT Model	-	Combines the above components to create the final CCT model.

Additionally, the proposed architecture adopted a rectified linear unit (ReLU) to introduce nonlinearity into the convolutional module. This part contributes to improving the model's ability to capture intricate patterns within CSI signals. The third layer of this convolutional module then applies an average pooling operation along the temporal dimension. The pooling operation is executed with a kernel size of (1, 75) and a stride of 1. Its function is to smooth out temporal features present in the data, thereby rendering a dual benefit. Initially, it mitigates the risk of overfitting, ensuring the robustness of the system during training. The reorganization entails compression of the channel dimension and the transposition of the convolution channel dimension for the temporal dimension. The transformative approach serves the purpose of furnishing all feature channels at each distinct temporal point as independent tokens, thus enabling their subsequent processing in the ensuing module with utmost precision.

4.5 Self-attention

It is postulated that incorporating context-dependent representations within low-level temporal-spatial features would confer important advantages to the task of signals, given the inherent coherence of neural activities (Abuhoureyah *et al*. 2024). Within the proposed module, we leverage the self-attention mechanism to discern and capture global temporal dependencies inherent to CSI features. It is valuable because it complements the previous convolutional module's limited receptive field. The organized tokens from the preceding module undergo a linear transformation, resulting in triads known as query (Q), key (K), and value (V), each of which shares the same shape. The token correlations are evaluated through a dot product operation applied to Q and K. Additionally, the scaling factor is introduced to mitigate the risk of vanishing gradients, thereby ensuring stable model training. Subsequently, the computed results traverse through a Softmax function, yielding a weighting matrix called the attention score.

The proposed schematic introduces two fully connected feed-forward layers downstream to enhance the model's capacity for feature representation. Notably, this operation preserves the input and output dimensions. The entire attention computation process is performed 'N' times within the self-attention module. Additionally, a multi-head strategy is adopted to augment representation diversity. The tokens are evenly segmented into 'h' segments and processed through the self-attention module. The outcomes from these separate heads are then concatenated to form the final module output (Dosovitskiy *et al*. 2021). Eq. (9) illustrates the multi-head attention process.

$$\begin{gathered}\operatorname{MHA}(Q,K,V)=[\text{ head }_0;\cdots;\text{ head }_{h-1}],\\ \text{head }_l=\text{ Attention }(Q_l,K_l,V_l)\end{gathered} \tag{9}$$

where ‘MHA’ signifies multi-head attention, and ‘Ql,’ ‘Kl,’ and ‘Vl’ in ‘l-th’ head denote the query, key, and value vectors obtained via linear transformation of the divided tokens, respectively. In essence, MHA mechanism enables to capture and weigh the importance of various temporal dependencies within CSI features, contributing to its ability to learn and represent complex patterns.

4.6 Classification

In the final stage of the model, the weighted representation, denoted as R, is propagated through a dense layer to produce the logits, symbolized as L, which form the foundation for subsequent classification decisions. Mathematically, it is expressed in Eq. (10).

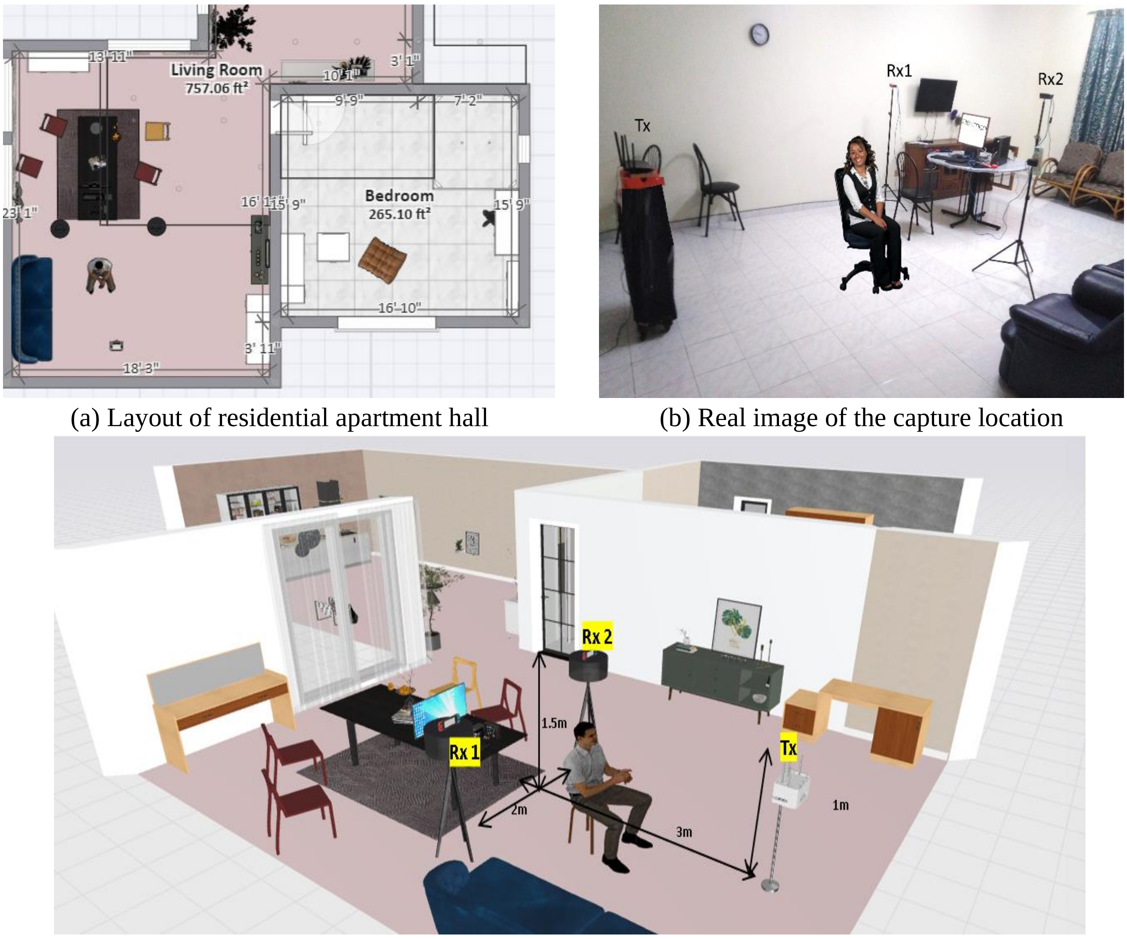
$$L=W\cdot R+b \tag{10}$$

Here, W represents the weights of the dense layer, and b is the bias term. The logits L are the raw scores associated with each class, capturing the model’s confidence in assigning an input image to a particular category. These logits are utilized in the classification process. The dense layer is employed at the final stage of the model to perform classification based on the extracted features. This layer is pivotal in transforming the learned representations from the transformer blocks into class predictions. Specifically, after applying sequence pooling to aggregate information across the encoded patches, a dense layer with a Softmax activation function is used to compute the final logits. This layer takes the weighted representation of the encoded patches and output logits corresponding to the number of classes.

CCT components are crafted to enhance the the framework amalgamates by using the principles of CNNs with self-attention mechanisms, drawing inspiration from transformers. The initial stage of the CCT unfolds with data augmentation procedures to input images aimed at enhancing the model’s generalization capabilities. Subsequently, image tokenization is carried out employing the CCT Tokenizer, which comprises a series of convolutional layers followed by max-pooling operations. The tokenization process partitions the image into discrete patches and meticulously extracts spatial characteristics. Optionally, positional embeddings, denoted as PE, are added to the tokenized features. These embeddings furnish valuable spatial positional information within the sequence. Mathematically, Eq (11) represents the addition of positional embeddings.

$$\text{Tokenized with Positional Embeddings}=\text{Tokenized Features}+\text{PE} \tag{11}$$

To infuse regularization into the model and deter overfitting, the technique of stochastic depth is incorporated. The mechanism deactivates a proportion of layers during the training process. Stochastic depth is employed to improve the robustness and generalization capability of the CCT model. It involves omitting entire transformer blocks during training, which creates a form of ensemble learning. By dropping layers, the model learns to adapt better across different datasets and variations in input data. Several stages are undertaken in the sequential operations within a transformer block to enhance the model’s understanding of input features. The process begins with Layer Normalization (LN1), where input features undergo normalization for improved stability.



(a) Layout of residential apartment hall

(b) Real image of the capture location

(c) Simulated 3D representation

Fig. 7 The environmental location for data collection

Multi-Head Self-Attention is employed to compute self-attention scores between tokens, capturing contextual information. The output of self-attention operation is then combined with the original tokenized features through the first Skip Connection. Subsequently, Layer Normalization (LN2) is applied to normalize the output further. The introduction of a Multi-Layer Perceptron (MLP) follows, utilizing feed-forward neural networks to capture intricate patterns. The MLP's output is integrated with the previous skip connection through Skip Connection 2. The structured sequence of operations empowers the model to capture localized and global image features. Similarly, the output of the culminating transformer block is utilized for sequence pooling. The attention weights denoted as α are computed across tokens. These attention weights are then employed to construct a weighted representation of the tokens. They are expressed in Eq. (12).

$$\text{Weighted Representation} = \alpha . \text{Tokenzied Features} \tag{12}$$

where the weighted representation is fed into a classification layer to derive predictions, marking the conclusion of the CCT's role in the image classification pipeline.

5. Experimental evaluation

The evaluation involved a comprehensive assessment using data procured through the Nexmon CSI extraction tool. The tool facilitated gathering CSI data from a Raspberry Pi 4B conFig.d for very high-throughput mode, operating with a total bandwidth of 80 MHz. The CSI samples acquired provided intricate channel information, encompassing 242 samples using 80 MHz bandwidth and 56 samples with 20 MHz data subchannels for each transmit-receive antenna pair. The setup included a Broadcom BCM43455c0 NIC with a Raspberry Pi 4B unit as the receiver and a TP-Link AC1350 router as the transmitter. The Raspberry Pi devices operated on Linux version 5.10.92 firmware and employed the Nexmon tool for CSI extraction. The receiver and transmitter adhered to IEEE 802.11n/ac standards and supported multi-user MIMO functionality. The setup ensured compatibility and adherence to industry standards, enhancing the reliability and validity of the experimental evaluation. CSI datasets were acquired within three distinct indoor settings, specifically within the confines of a residential apartment hall, as delineated by the distance spacing parameters illustrated in Fig. 7.

Two clustered Raspberry Pi units, each measuring 1.5 m in height and separated by 2 m, collected data, with the Rx positioned 3 m away from the router Tx device in Fig. 7's simulated location. During data collection, one individual engaged in one of five specific activities within each location. These activities included empty rooms (E), walking (W), swiping (S), continuous hand cycles (c), and standing in place (ST). The CSI samples were collected in three distinct indoor environments: a home at Hall, a classroom at FTKEK University at University Technical Melaka Malaysia, and a university laboratory. These environments were chosen to investigate the effects of varying room geometries and the presence of static obstacles on CSI data.

5.1 Data augmentation

Enhancing the diversity of the training dataset through data augmentation fosters robustness and generalization capabilities in the model. This specific illustration employs geometric augmentations like random cropping and horizontal flipping. The initial step involves a "rescaling" layer that normalizes pixel values to a scale between 0 and 1. Subsequently, a "RandomCrop" layer randomly extracts portions of the input images, introducing variability in object positioning. The final augmentation layer, "RandomFlip," imparts horizontal flipping to the images with a specified probability, addressing variations in object orientation. A Keras sequential model named "data_augmentation" encapsulates the sequence of augmentations, facilitating a structured and sequential application of the augmentation procedures.

5.2 Training process

The presented model segment encapsulates the training and evaluation processes of a CCT. The experimental workflow is orchestrated through a function where the system is compiled with a specialized AdamW optimizer conFig.d with a learning rate of 0.001 and weight decay of 0.0001. The loss function is defined as categorical cross-entropy, with label smoothing set to 0.1 to foster generalization. Evaluation metrics encompass categorical and top-5 categorical assessments of the model's performance. During training, a checkpoint callback is employed to monitor the validation accuracy and save the best-performing weights. The training data is subjected to a specified number of epochs with a designated batch size. A 20% validation split assesses the model's

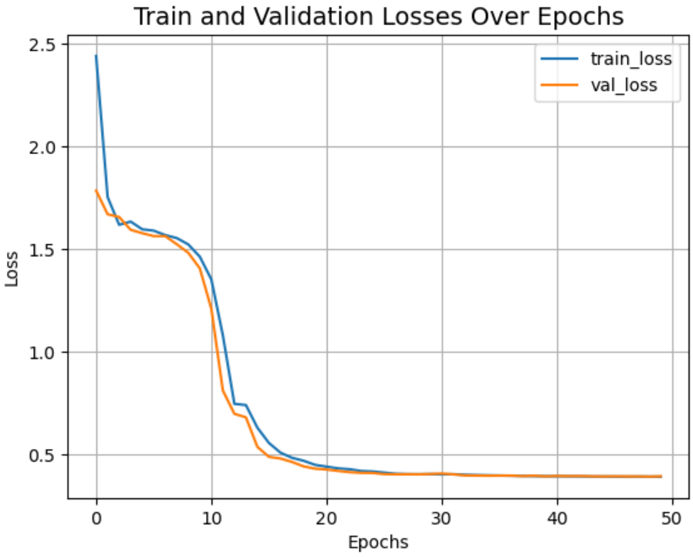


Fig. 8 Training and validation loss curves over successive epochs, illustrating the model's learning dynamics

generalization on a subset of the training data. Notably, the weights yielding the highest validation accuracy are loaded for the final evaluation. The training and evaluation process is executed systematically, ensuring a rigorous examination of the model's efficacy in learning from the training data and generalizing to unseen test data, providing valuable insights into its overall performance. Fig. 8 describes validation and training losses and represents the model's learning progression over epochs. The distinct curves, training loss, and validation loss showcase how the model refines its parameters and generalizes to new data.

5.3 Visualization of distribution changes

The results of the CCT, which leverages WiFi-based CSI sensing, were evaluated to recognize five distinct activities. The innovative algorithm sought to harness the unique characteristics of WiFi signals to discern activities in a transportation context. The confusion matrix shown in Fig. 9 provides a breakdown of the model's classification performance.

5.4 Comparative analysis

Direct comparisons between CNN, ViT, Swin Transformer, and CCT are intricate due to their disparate architectural paradigms and design principles; however, CCT Transformer has demonstrated its merit as a competitive alternative. CNNs, having long stood as the de facto standard in computer vision, have delivered impressive results across various tasks. However, It and Swin transformers' empirical success in outperforming established state-of-the-art models in accuracy across varied vision tasks underscores its potential as a formidable contender. Leveraging their hierarchical architecture and innovative shifted windowing scheme, ViT and Swin Transformer prove adept at feature extraction, offering a promising alternative to CNNs as a

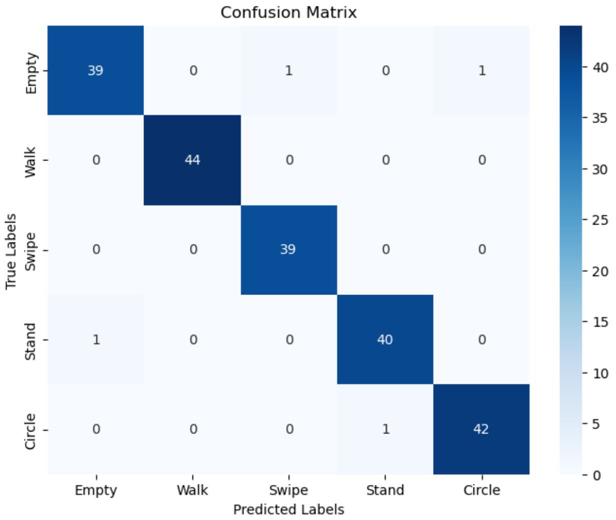
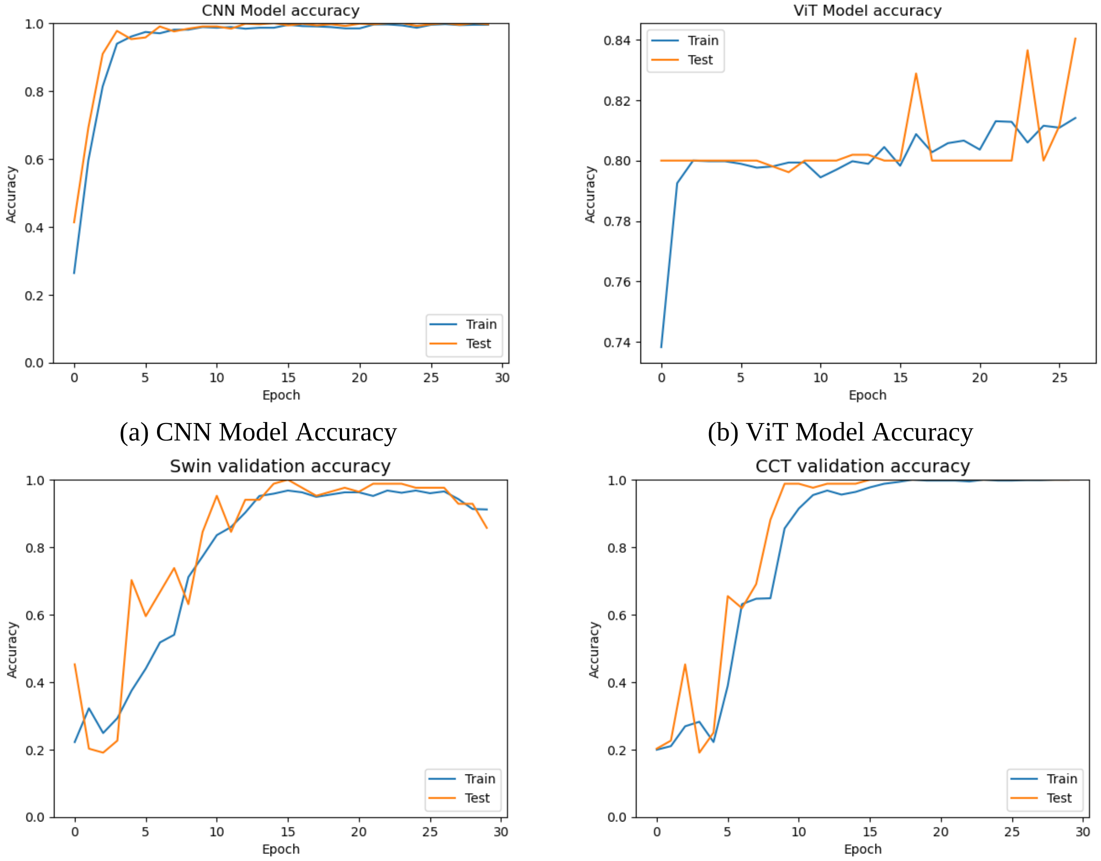


Fig. 9 Confusion matrix of CCT model employs WiFi-based CSI sensing



(a) CNN Model Accuracy

(b) ViT Model Accuracy

(c) SWIN Model Accuracy

(d) CCT Model Accuracy

Fig. 10 Analysis of Validation Accuracy for CSI vs. CNN, ViT, SWIN, and CCT

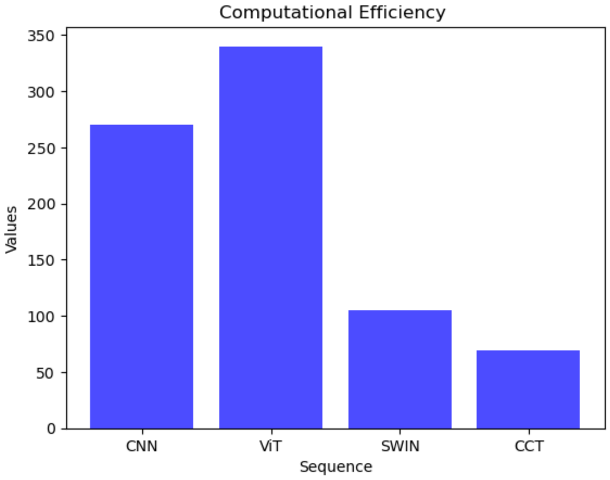


Fig. 11 Analysis of Execution Time for CNN, ViT, SWIN, and CCT

versatile and efficient backbone for addressing a spectrum of challenges in computer vision. Fig. 10 examines the classification across four distinct deep learning architectures: CNN, ViT, Swin Transformer, and CCT.

CNN, a longstanding and conventional architecture, has demonstrated commendable accuracy in image classification applications. Meanwhile, ViT, representing a novel paradigm with its attention-based mechanism, is anticipated to excel in capturing long-range dependencies within images. SWIN, another transformer-based architecture, brings a hierarchical structure to image understanding. Conclusively, CCT is a fusion of convolutional and transformer components to leverage the strengths of both paradigms by assessing the classification of these models to discern the strengths and limitations of each architecture in handling intricate image categorization tasks. The perceptions are instrumental in finding the optimal choice of deep learning architecture contingent on the specific characteristics and requirements of the classification task at hand.

The algorithm incorporates positional embeddings and employs two convolutional layers with a projection dimension of 128. To enhance model robustness, a stochastic depth rate of 0.1 is applied. Training parameters include a learning rate of 0.001, weight decay of 0.0001, and a batch size of 128 over 30 epochs. The input images are resized to 32x32 pixels, and the dataset consists of labeled images categorized into classes represented as vectors. The model is trained on a dataset split into training and testing sets using the split function from scikit-learn, ensuring high accuracy classification on unseen data.

5.5 Computational efficiency

The computational efficiency shown in Fig. 11 serves as a quantitative metric for each model’s temporal performance in undertaking a given computational task. The comparison shows that CCT exhibits the lowest execution time, indicating superior performance compared to the other algorithms. ViT, on the other hand, is associated with the highest execution time among the examined algorithms.

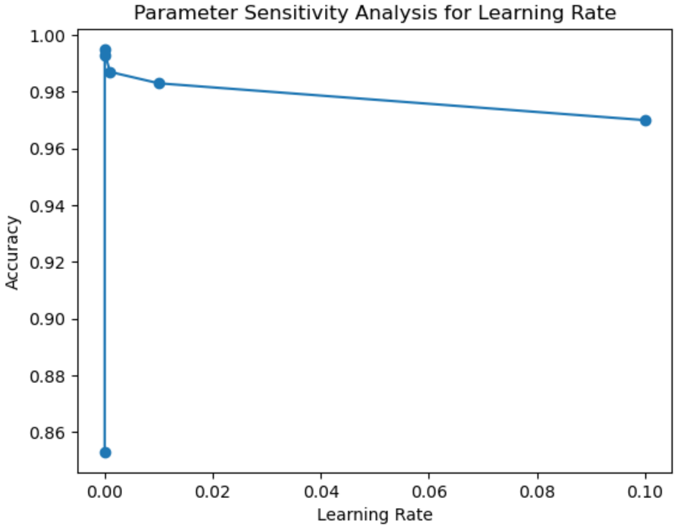


Fig. 12 Parameter sensitivity analysis for the CCT model

5.6 Parameter sensitivity analysis

The model's parameter sensitivity analysis extends beyond the learning rate to encompass additional hyperparameters, delving into their intricate interplay and collective influence on performance. The findings unfold across defined hyperparameter space, varying parameters such as batch size, the number of transformer layers, and the attention mechanism's configuration. The analysis aims to clarify the model's alterations in each hyperparameter, offering understanding of their individual impacts on the model's accuracy. The x-axis of the plotted Fig. delineates discrete values of the explored hyperparameters, presenting a spectrum of configurations ranging from conservative to aggressive parameter choices. Fig. 12 unveils nuanced patterns and trends by traversing this parameter space, providing invaluable insights into the model's sensitivity to alterations in various hyperparameters. The findings analysis extends beyond the learning rate, offering an aggregation on the intricate relationship between hyperparameters and the consequential impact on the model's performance.

5.7 Ablation study

An ablation analysis to explore and delineate the performance characteristics offered by the CCT architecture compared to traditional models. The approach facilitated a nuanced analysis of the impact of each component on the overall model performance. The ANOVA results shown in Fig. 14 visualized through a boxplot provide insights into potential statistical differences between the groups. The initial cohort of analyses pertains to the impact of sample rates, revealing a positive correlation wherein heightened sample rates during training result in superior performance. An inverse relationship emerges within the second set focusing on activity classes, indicating that elevated class numbers correspond to diminished accuracy. Furthermore, the investigations demonstrate that the model's performance exhibits enhancement with a reduction in the number of patches employed.

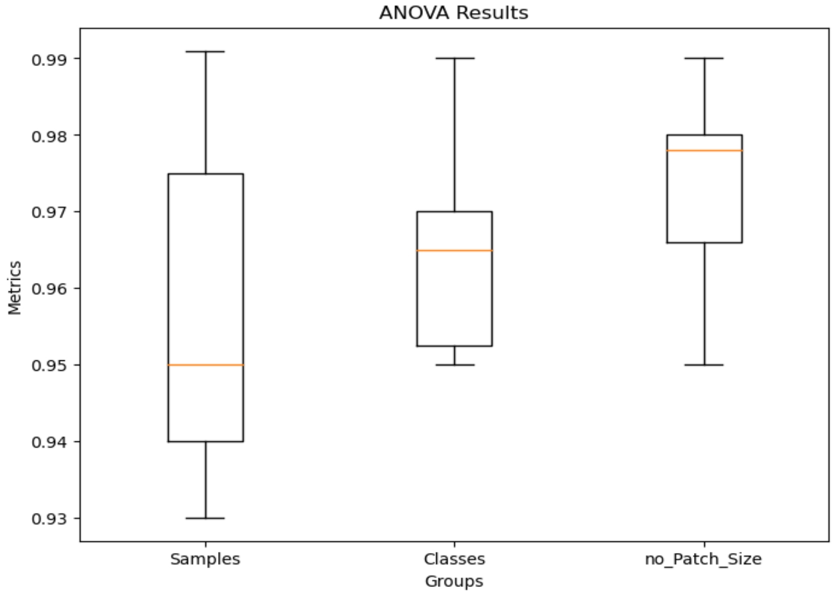


Fig. 13 Comparison of performance metrics across different experimental conditions

The ablation findings revealed that the hybrid architecture of CCT, amalgamating convolutional and transformer features, enhanced design performance. The convolutional components facilitated local feature extraction, while the transformer components enabled practical global context understanding, thereby combining the strengths of both architectural paradigms. Furthermore, the analysis explored a lightweight variant of CCT tailored for resource-constrained environments. The findings demonstrate promising results performance while reducing computational requirements. The attribute positions the lightweight CCT variant as a viable option for deployment in scenarios where computational resources are limited.

5.8 Comparison between CCT and current studies in CSI HAR

To evaluate the accuracy of the various listed algorithms, the self-collected dataset, which includes recordings of five activities, was evaluated. Compared to CNN-based models, CCT exhibits a distinct advantage in capturing local and global contextual information, surpassing the limitations of traditional CNN architectures. Additionally, it outperforms other state-of-the-art algorithms such as LSTM, GRU, and attention-based schematic in modelling both short-term and long-term dependencies in the input data. Furthermore, CCT demonstrates superior performance in comparison to ensemble methods, transfer learning approaches, and reinforcement learning algorithms used in similar domains. While CNNs (Showmik *et al*. 2023) excel at extracting local features through convolutional layers, they often struggle to capture long-range dependencies in sequential data. In contrast, CCT integrates transformer components which enables it to handle the global context of CSI sequences and capture dependencies feature between distant samples and improve the recognition performance. Additionally, CCT achieves a balanced fusion of local and global information by combining the best parts of both architectures.

In addition to LSTM CNN (Ding *et al*. 2021), meta, and LSTM models, CCT offers unique

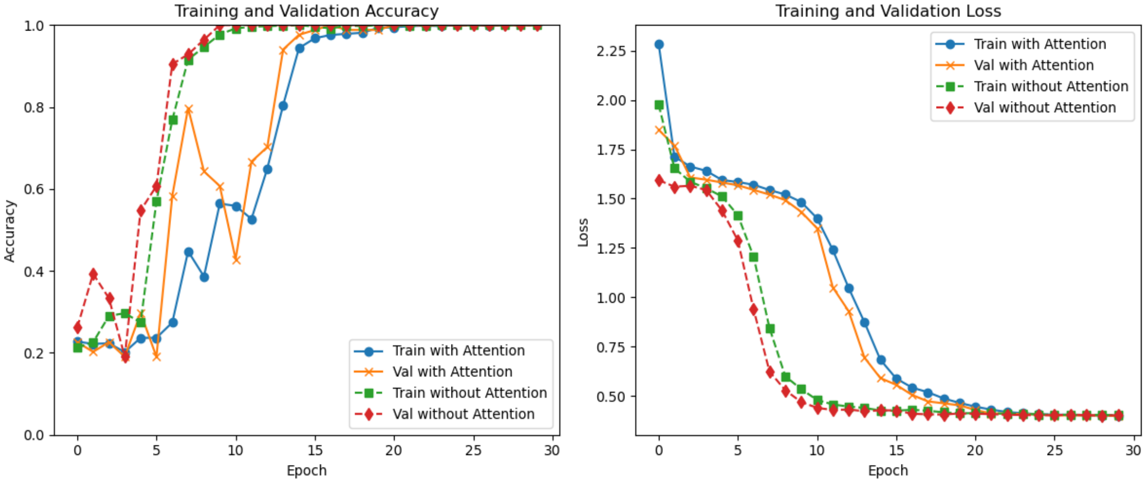


Fig. 14 Validation comparison when applying attention mechanisms

features and advantages for temporal modelling. Likewise, LSTM CNN models incorporate both convolutional and recurrent layers. Although they capture temporal dependencies, they suffer from the vanishing gradient problem and struggle with longer sequences. On the other hand, meta-learning approaches aim to learn the architecture for a given task. While these methods adapt to new tasks quickly, they require extensive labeling data and computational resources during the training phase. In contrast, CCT combines convolutional and transformer components addressing the limitations of LSTM CNN models and meta-learning approaches. CCT captures long-range dependencies while benefiting from the efficient local feature extraction of CNNs by leveraging the self-attention mechanism of transformers resulting in improving scalability for CSI-based HAR tasks.

5.9 Evaluating CCT model

To evaluate the impact of the multi-head attention mechanism within the proposed CCT model, we constructed two versions of the model: one incorporating the multi-head attention mechanism and another excluding it. Both models were trained and evaluated on the same dataset, with consistent hyperparameters and training procedures to ensure a fair comparison. The results were recorded and analyzed. The analysis revealed that the model with multi-head attention exhibited performance in both training and validation phases, as evidenced by higher accuracy and lower loss values. This indicates that the multi-head attention mechanism enhances the model's ability to capture and utilize complex patterns within the data, thereby improving its predictive capability. These findings validate the efficacy of incorporating multi-head attention in transformer models for image classification tasks, demonstrating its pivotal role in enhancing model performance as shown in Fig. 15.

The evaluation of stochastic depth in the CCT model was conducted by training two versions: one with stochastic depth enabled and one without. The validation accuracies and losses were plotted over 30 epochs to compare their performance. The results showed improved inference with reduced overfitting, as evidenced by the lower validation loss and higher validation accuracy than

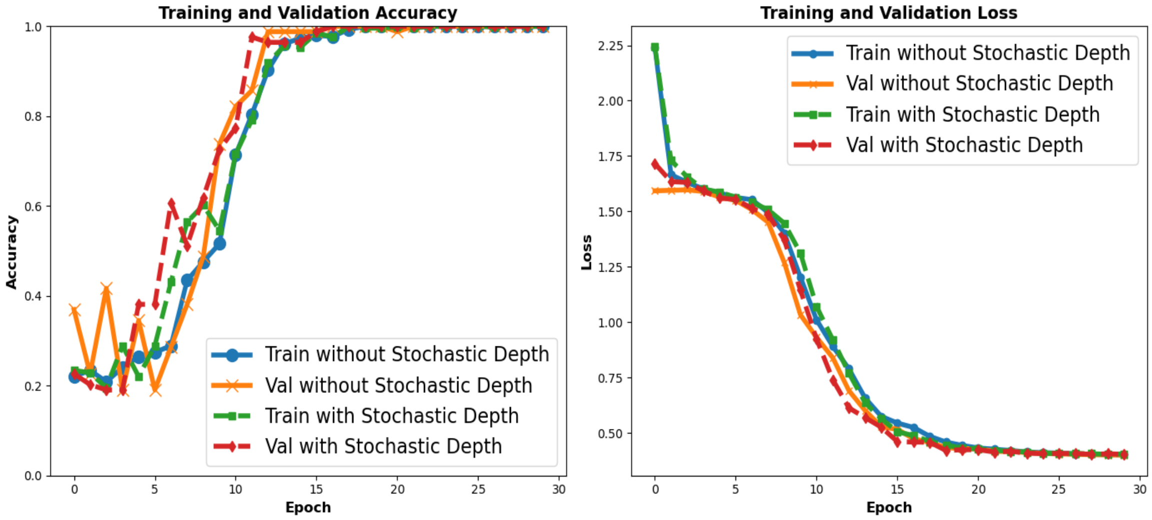


Fig. 15 Validation Performance of CCT Models with and without Stochastic Depth

the model without stochastic depth, as shown in Fig. 16. The results support the hypothesis that stochastic depth enhances the robustness and performance of deep learning models by introducing random regularization during training.

5.10 Constraints and future works

Deploying a lightweight CCT for CSI WiFi-based sensing introduces certain constraints that require careful consideration. For example, the technique still inherent trade-off between model complexity and classification capability. As the CCT is struggles to operate in resource-constrained environments, its lightweight design may limit the depth and intricacy of its architecture impacting the model's ability to discern delicate variations in the complex and dynamic channel state information. This condition requires further exacerbated in scenarios with extensive signal interference or changing wireless environments. Moreover, WiFi signals traverse multiple paths and exhibit disparities in signal strengths, rendering the accurate estimation of an individual's location based solely on WiFi measurements a formidable task. The environment impacts the wireless signals used to capture the CSI measurements and the accuracy of HAR is also influenced by the positioning and orientation of the person being monitored as indicated in Fig. 16.

Eq. (13) represents the power propagated in space with varied gains between transmitter and receiver. By determining the environmental aspect, it is possible to analyze the environmental effects of the signal to eliminate the dependency on the environment.

$$P_r = \frac{Pt * G_{t*} G_{r*} \lambda^2 * F}{(4\pi * R)^2} \tag{13}$$

Transmitted power Pt and transmitted and receiver gains $G_{t,} G_r$ directly affect the received power P_r. In addition to the signal wavelength λ and F is the propagation factor coming from the environment. Moreover, the distances between Tx and Rx, such as propagation range R and other

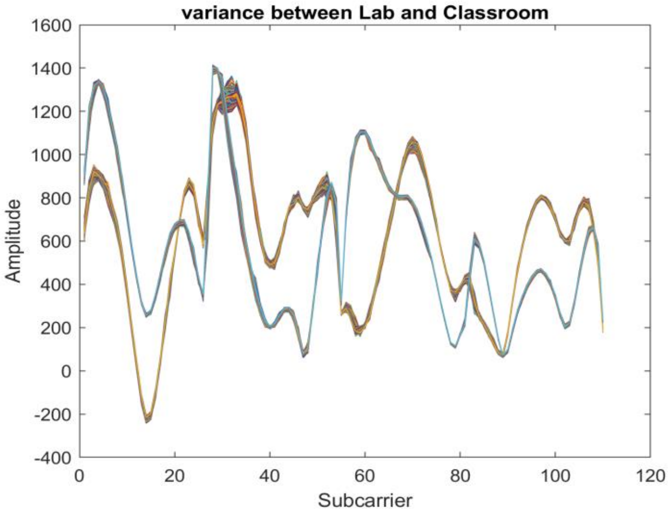


Fig. 16 Comparison Variability in CSI amplitude across distinct locations during identical activities

limitations, might be analyzed further in future developments for lightweight CCT, especially in using CSI WiFi-based sensing. The first pertains to the need for continuous optimization to strike an optimal balance between performance and computational efficiency. Future work is required to enhance the method with neural architecture quantization, refine the lightweight CCT architecture, and enhance its flexibility to different sensing conditions. Another avenue for exploration involves robustness assessments in scenarios with non-ideal signal conditions or limited training data.

CCT architecture demonstrates versatility across a range of applications. The algorithms can be applied to several tasks to highlight their adaptability. For image classification, CCT achieved in standard benchmarks such as CIFAR-10 and CIFAR-100. In object detection, it performed well on the dataset, identifying and localizing objects accurately. The model can also be used in semantic segmentation, defect detection, and labeling while excelling in segmenting urban scenes.

In future endeavors, attention must be directed toward expanding the exploration of multi-modal sensing approaches, integrating with other sensor types, and incorporating transfer learning techniques could fortify the model's generalization capacity across different sensing scenarios. Additionally, efforts to enhance the interpretability and ability of the model's decisions will be pivotal, fostering trust in its outcomes for real-world applications. The evolution of lightweight CCT in CSI WiFi-based sensing holds promising solutions, with ongoing research poised to address existing constraints and pave the way for broader deployment in practical sensing applications.

6. Conclusions

In conclusion, using CCT in CSI WiFi sensing presents a promising avenue for the state-of-the-art in image classification and wireless sensing application. The amalgamation of convolutional and transformer architectures within CCT exhibits advantages in capturing both spatial and temporal dependencies inherent in CSI datasets. The hierarchical feature extraction facilitated by the initial convolutional layers of CCT proves instrumental in discerning patterns within the wireless channel. Furthermore, the proposed model is underscored by its ability to mitigate the associated challenges, such as the requirement of using large-scale datasets and offering solutions in scenarios where dataset acquisition is complex. CCT's contributions are in the precision metrics achieved by classifying activities within the CSI wireless sensing refinement holds promising applications ranging from human complex signals in the wireless domain, reinforcing its potential technology for indoor environment sensing applications. Furthermore, as activity recognition to environmental monitoring, future research needs to involve in the optimization of self-learning and labeling techniques with the use of techniques such as reinforcement learning.

Acknowledgment

The authors acknowledge the technical and financial support by the Ministry of Higher Education, Malaysia, under the research grant no. FRGS/1/2024/ICT02/UTEM/02/3 and Universiti Teknikal Malaysia Melaka (UTeM)

References

Abuhoureyah, F.S., Wong, Y.C., Sadhiqin, A. and Mohd, B. (2024), "WiFi-based human activity recognition through wall using deep learning", *Eng. Appl. Artif. Intell.*, **127**(PA), 107171. https://doi.org/10.1016/j.engappai.2023.107171.

Abuhoureyah, F., Yan Chiew, W., Bin Mohd Isira, A.S. and Al-Andoli, M. (2023), "Free device location independent WiFi-based localisation using received signal strength indicator and channel state information", *IET Wireless Sensor Syst.*, **13**(5), 163-177. https://doi.org/10.1049/wss2.12065

Ahmed, H.F.T., Ahmad, H., Narasingamurthi, K., Harkat, H. and Phang, S.K. (2020), "DF-WiSLR: Device-Free Wi-Fi-based sign language recognition", *Pervasive Mobile Comput.*, **69,** 101289. https://doi.org/10.1016/j.pmcj.2020.101289.

Abuhoureyah, F., Sim, K.S. and Wong, Y.C. (2024), "Multi-user human activity recognition through adaptive location-independent WiFi signal characteristics", *IEEE Access*, **12**, 112008-112024. https://doi.org/10.1109/ACCESS.2024.3438871.

Ahmed Ouameur, M., Caza-Szoka, M. and Massicotte, D. (2020), "Machine learning enabled tools and methods for indoor localization using low power wireless network", *Internet Thing.*, **12**, 100300. https://doi.org/10.1016/j.iot.2020.100300.

Akhtar, Z.U.A. and Wang, H. (2020), "Wifi-based driver's activity monitoring with efficient computation of radio-image features", *Sensors*, **20**(5). https://doi.org/10.3390/s20051381

Darabkh, K.A., Al-Akhras, M., Zomot, J.N. and Atiquzzaman, M. (2022), "RPL routing protocol over IoT: A comprehensive survey, recent advances, insights, bibliometric analysis, recommendations, and future directions", *J. Netw. Comput. Appl.*, **207**(2021), 103476. https://doi.org/10.1016/j.jnca.2022.103476.

Dierickx, P., Van Damme, A., Dupuis, N. and Delaby, O. (2023), "Comparison between CNN, ViT and CCT for channel frequency response interpretation and application to G.Fast", *IEEE Access*, **11**, 24039-24052. https://doi.org/10.1109/ACCESS.2023.3247877.

Ding, J. and Wang, Y. (2019), "WiFi CSI-based human activity recognition using deep recurrent neural network", *IEEE Access*, **7**, 174257-174269, https://doi.org/10.1109/ACCESS.2019.2956952

Ding, X., Jiang, T., Zhong, Y., Huang, Y. and Li, Z. (2021), "Wi-Fi-based location-independent human activity recognition via meta learning", *Sensors*, **21**(8). https://doi.org/10.3390/s21082654.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J. and Houlsby, N. (2021), "An image is worth 16X16 words: Transformers for image recognition at scale", *Proceedings of the ICLR 2021 9th International*

Conference on Learning Representations.

Fard Moshiri, P., Shahbazian, R., Nabati, M. and Ghorashi, S.A. (2021), "A CSI-based human activity recognition using deep learning", *Sensors*, **21**(21), 1-19. https://doi.org/10.3390/s21217225.

Abuhoureyah, F., Yan Chiew, W. and Zitouni, M.S. (2024), "WIFI based human activity recognition using multi-head adaptive attention mechanism", *J. Intell. Fuzzy Syst., Preprint*, 1-16.

Guo, L., Wang, L., Liu, J., Zhou, W. and Lu, B. (2018), "HuAc: Human activity recognition using crowdsourced WiFi signals and skeleton data", *Wireless Commun. Mobile Comput.,* **2018**, 1-15. https://doi.org/10.1155/2018/6163475.

Jiang, D., Li, M. and Xu, C. (2020), "Wigan: A wifi based gesture recognition system with gans", *Sensors*, **20**(17), 1-19. https://doi.org/10.3390/s20174757.

Lee, D.J., Lee, J.Y., Shon, H., Yi, E., Park, Y.H., Cho, S.S. and Kim, J. (2023), "Lightweight monocular depth estimation via token-sharing transformer", *Proceedings of the IEEE International Conference on Robotics and Automation*, 2023-May(Icra), 4895-4901. https://doi.org/10.1109/ICRA48891.2023.10160566.

Li, T., Shi, C., Li, P. and Chen, P. (2021), "A novel gesture recognition system based on CSI extracted from a smartphone with nexmon firmware", *Sensor*, **21**(1), 1-19. https://doi.org/10.3390/s21010222.

Li, Y., Xie, X., Fu, H., Luo, X. and Guo, Y. (2023), "A compact transformer for adaptive style transfer", *Proceedings of the IEEE International Conference on Multimedia and Expo*, 2023-July, 2687-2692. https://doi.org/10.1109/ICME55011.2023.00457.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S. and Guo, B. (2021), "Swin transformer: Hierarchical vision transformer using shifted windows", *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 10012. https://doi.org/10.48550/arXiv.2103.14030.

Lu, Z., Xie, H., Liu, C. and Zhang, Y. (2022), "Bridging the gap between vision transformers and convolutional neural networks on small datasets", *Adv. Neural Inform. Proc. Syst.*, **35**(NeurIPS), 1-15. https://doi.org/10.48550/arXiv.2210.05958.

Luo, Z., Cheng, X. and Yang, Y. (2022), "Computational electromechanical approach for stability/instability of smart system actuated with piezoelectric NEMS", *Adv. Comput. Des.*, **7**(3), 21. https://doi.org/10.12989/acd.2022.7.3.211

Ma, Y., Zhou, G, and Wang, S. (2019), "WiFi sensing with channel state information: A survey", *ACM Comput. Surveys*, **46**(1), 1-32. https://doi.org/10.1186/1687-6180-2011-10

Muaaz, M., Chelli, A., Gerdes, M.W. and Pätzold, M. (2022), "Wi-Sense: A passive human activity recognition system using Wi-Fi and convolutional neural network and its integration in health information systems", *Annal. Telecommun.*, **77**(3-4), 163-175. https://doi.org/10.1007/s12243-021-00865-9

Palanivelu, R. and Srinivasan, P.S.S. (2020), "Safety and security measurement in industrial environment based on smart IOT technology based augmented data recognizing scheme", *Comput. Commun.*, **150**, 777-787. https://doi.org/10.1016/j.comcom.2019.12.013.

Pan, X., Jiang, T., Li, X., Ding, X., Wang, Y. and Li, Y. (2019), "Dynamic hand gesture detection and recognition with WiFi Signal Based on 1D-CNN", *Proceedings of the 2019 IEEE International Conference on Communications Workshops, ICC Workshops 2019*, 0-5. https://doi.org/10.1109/ICCW.2019.8756690.

Saw, C.Y. and Wong, Y.C. (2023), "Neuromorphic computing with hybrid CNN-Stochastic Reservoir for time series WiFi based human activity recognition", *Comput. Electr. Eng.*, **111**(PA), 108917. https://doi.org/10.1016/j.compeleceng.2023.108917

Schäfer, J., Barrsiwal, B.R., Kokhkharova, M., Adil, H. and Liebehenschel, J. (2021), "Human activity recognition using csi information with nexmon", *Appl. Sci.*, **11**(19). https://doi.org/10.3390/app11198860.

Sharma, L., Chao, C., Wu, S.L. and Li, M.C. (2021), "High accuracy wifi-based human activity classification system with time-frequency diagram cnn method for different places", *Sensors*, **21**(11). https://doi.org/10.3390/s21113797.

Showmik, I.A., Sanam, T.F. and Imtiaz, H. (2023), "Human activity recognition from Wi-Fi CSI data using principal component-based wavelet CNN", *Digital Signal Proc. Rev. J.*, **138**, 104056. https://doi.org/10.1016/j.dsp.2023.104056.

Tiku, S., Pasricha, S., Notaros, B. and Han, Q. (2020), "A Hidden Markov Model based smartphone heterogeneity resilient portable indoor localization framework", *J. Syst. Arch.*, **108**(8), 101806. https://doi.org/10.1016/j.sysarc.2020.101806

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N. and Polosukhin, I. (2017), "Attention is all you need", *Adv. Neural Inform. Proc. Syst.*, **30**.

Wang, D., Yang, J., Cui, W., Xie, L. and Sun, S. (2022), "AirFi: Empowering WiFi-based passive human gesture recognition to unseen environment via domain generalization", *IEEE T. Mobile Comput.*, **23**(2), 1156-1168. https://doi.org/10.1109/TMC.2022.3230665

Yang, J., Chen, X., Wang, D., Zou, H., Lu, C.X., Sun, S. and Xie, L. (2022a), "Deep learning and its applications to WiFi human sensing: A benchmark and a tutorial", *arXiv preprint*, arXiv:2207.07859.

Yang, J., Chen, X., Wang, D., Zou, H., Lu, C.X., Sun, S. and Xie, L. (2022b), "SenseFi: A library and benchmark on deep-learning- empowered WiFi human sensing", *Patterns*, **4**(3), 100703. https://doi.org/10.1016/j.patter.2023.100703

Yang, J., Zou, H. and Xie, L. (2022), "SecureSense: Defending adversarial attack for secure device-free human activity recognition", *IEEE T. Mobile Comput.*, **1**(11). https://doi.org/10.1109/TMC.2022.3226742

Yang, Z., Zhou, Z. and Liu, Y. (2013), "From RSSI to CSI: Indoor localization via channel response", *ACM Comput. Surveys*, **46**(2), 1-32. https://doi.org/10.1145/2543581.2543592

Zhang, Y., Yin, Y., Wang, Y., Ai, J. and Wu, D. (2023), "CSI-based location-independent Human Activity Recognition with parallel convolutional networks", *Comput. Commun.*, **197**(2022), 87-95. https://doi.org/10.1016/j.comcom.2022.10.027

Zhang, Y., Zheng, Y., Qian, K., Zhang, G., Liu, Y., Wu, C. and Yang, Z. (2021), "Widar3.0: Zero-effort cross-domain gesture recognition with Wi-Fi", *IEEE T. Pattern Anal. Machine Intell.*, **8828**(c). https://doi.org/10.1109/TPAMI.2021.3105387

Zhou, C., Yang, L., Liao, H., Liang, B. and Ye, X. (2021), "Ankle foot motion recognition based on wireless wearable sEMG and acceleration sensors for smart AFO", *Sensors Actuat. A Phys.*, **331**, 113025. https://doi.org/10.1016/j.sna.2021.113025

Zhou, R., Hou, H., Gong, Z., Chen, Z., Tang, K. and Zhou, B. (2021), "Adaptive device-free localization in dynamic neural networks", *IEEE Sensors J.*, **21**(1), 548-559. https://doi.org/10.1109/JSEN.2020.3014641

CC