WiFi CSI Based Temporal Activity Detection Via Dual Pyramid Network

Zhendong Liu[1], Le Zhang[1] *, Bing Li[1], Yingjie Zhou[2], Zhenghua Chen[3], Ce Zhu[1]

[1]School of information and Communication Engineering, University of Electronic Science and Technology of China
[2]College of Computer Science, Sichuan University
[3]Institude for Infocomm Research, Agency for Science, Technology and Research (ASTAR), Singapore
lzdjohn@std.uestc.edu.com, {lezhang, bing_li, eczhu}@uestc.edu.cn, yjzhou@scu.edu.cn, chen_zhenghua@i2r.a-star.edu.sg

Abstract

We address the challenge of WiFi-based temporal activity detection and propose an efficient Dual Pyramid Network that integrates Temporal Signal Semantic Encoders and Local Sensitive Response Encoders. The Temporal Signal Semantic Encoder splits feature learning into high and low-frequency components, using a novel Signed Mask-Attention mechanism to emphasize important areas and downplay unimportant ones, with the features fused using Contra-Norm. The Local Sensitive Response Encoder captures fluctuations without learning. These feature pyramids are then combined using a new cross-attention fusion mechanism. We also introduce a dataset with over 2,114 activity segments across 553 WiFi CSI samples, each lasting around 85 seconds. Extensive experiments show our method outperforms challenging baselines. Code and dataset are available at https://github.com/AVC2-UESTC/WiFiTAD.

1 Introduction

Using IoT devices to recognize human activities like walking, falling, and lying down has numerous applications (Kong and Fu 2022). Recently, there has been growing interest in not just short-term activity analysis but also long-term daily behavior monitoring, which is vital for real-world applications like health monitoring and medical statistics (Gu et al. 2019).

For long-term monitoring, researchers are increasingly focusing on processing extended data streams, specifically through Temporal Activity Detection (TAD), which aims to automatically identify activities and their timing within prolonged monitoring data (Shou, Wang, and Chang 2016). However, most of these efforts rely on camera sensors, which require a direct line of sight (LOS) to function effectively. This limitation restricts their use in low-light conditions and raises significant privacy concerns, making them less suitable for sensitive environments where confidentiality is important (Sun et al. 2022).

As a privacy-preserving alternative, researchers have turned to non-invasive technologies like WiFi Channel State Information (CSI) to sense and recognize human activities (Chen et al. 2018). Unlike cameras, WiFi CSI doesn't capture visual images, eliminating privacy issues. By detecting the multi-path effects caused by human movements, CSI allows for Non-Line-of-Sight (NLOS) monitoring along fixed transmitter-receiver paths. The widespread availability of WiFi makes this approach not only cost-effective but also highly accurate for activity recognition, even in challenging conditions (Zeng et al. 2018).

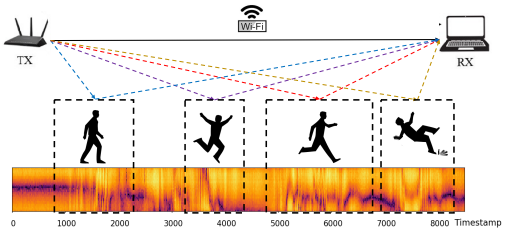

Figure 1: The variation of ubiquitous wireless signal caused by different human activities along the temporal axis.

Despite significant progress in using CSI for human activity recognition (HAR), most existing methods assume that input signals are pre-segmented into distinct activities, focusing on correctly identifying activities from a predefined set of classes (Chen et al. 2018; Meng et al. 2023; Li et al. 2021a; Chavarriaga et al. 2013; Xiao et al. 2020; Yousefi et al. 2017). While effective for well-defined segments, these methods struggle in more complex scenarios where activity boundaries are not predefined, and the signal is continuous and untrimmed, as shown in Fig. 1.

Temporal activity detection is well-developed in computer vision using visual inputs (Wang et al. 2023a), but applying these methods directly to WiFi CSI-based activity detection is challenging due to differences in data modality and characteristics. Unlike visual data, which provides rich spatial features and clear temporal sequences, WiFi CSI data is primarily temporal, noisy, and lacks intuitive spatial cues. Environmental factors introduce noise in WiFi signals, which is different from the noise typically encountered in visual data. Additionally, WiFi CSI suffers from a scarcity of annotated data, unlike the large labeled datasets available in computer vision. Moreover, WiFi CSI data, generated by inexpensive sensors, requires more efficient models than the computationally intensive approaches used in computer vision. These differences highlight the need for tailored methods for WiFi CSI, making it unsuitable to directly

*Corresponding author

apply computer vision-based approaches.

To address these challenges, we explore wireless Temporal Activity Detection (TAD) and introduce DPWiT, an end-to-end learning model. DPWiT uses a multi-scale dual pyramid structure that combines frequency-aware feature learning with fluctuation information to accurately identify activities and their precise locations within untrimmed, long-term signals. The core component, the Dual Pyramid Temporal Context Modeling (DPTCM), generates multi-scale features through Temporal Signal Semantic Encoders (TSSE) and Local Sensitive Response Encoders (LSRE), which are fused using Cross-attention Pyramid Fusion modules. We also collected and annotated a comprehensive untrimmed WiFi CSI dataset covering seven daily activities: walk, run, jump, wave, fall, sit, and stand. This dataset includes 553 untrimmed samples with 2,114 activity instances, each annotated with start time, end time, and category. To summarize, we contribute in:

- We systematically study the WiFi based Temporal Activity Detection (TAD) task and introduce a comprehensive solution. This includes developing a novel method, creating a real-world dataset of long-term, untrimmed multi-activity wireless signals, and establishing a new benchmark for future research.
- We explore the classification and localization sub-tasks of TAD, finding that high-frequency information is crucial for localization, while low-frequency information is better for identifying activity categories.
- We propose DPWiT, a model that combines frequency-aware learning with dual pyramid fusion, achieving state-of-the-art results on real-world datasets. For low-frequency learning, we introduce Signed Mask-Attention, which better highlights important areas and downplays unimportant ones, enhancing the model's focus on critical regions.

2 Related Work

2.1 CSI based Huam Activity Analysis

Recently, deep learning-based strategies have been increasingly applied to CSI-based human activity recognition, inspired by the success of deep neural networks in various fields (Krizhevsky, Sutskever, and Hinton 2012; Li et al. 2021a). Models like ABLSTM (Chen et al. 2018) and Transformers (Li et al. 2021b) have demonstrated the advantages of temporal context modeling in improving recognition accuracy. Some studies have extended WiFi signal use to long-term monitoring, achieving effective respiration monitoring in real home environments (Tian et al. 2018; Liu et al. 2021; Wang et al. 2023b). Although they analyze successive activity recognition from the temporal angle, these methods are generally limited to single, well-defined activity segments and regular patterns. The challenge of accurately detecting and recognizing various activities from untrimmed, unrestricted WiFi CSI data remains largely unresolved.

2.2 Temporal Activity Detection

Temporal activity detection is well-developed in computer vision using visual inputs (Wang et al. 2023a). Existing vision-based TAD methods are divided into one-stage and two-stage approaches. Two-stage methods (Chen et al. 2022; Xia et al. 2022) first generate potential instance proposals and then classify and refine them using independently trained detectors. In contrast, one-stage methods (Shi et al. 2023; Yang et al. 2020) use an end-to-end pipeline to simultaneously localize and recognize actions. Although temporal action detection (TAD) is well-studied in the vision community, WiFi-based TAD remains in its early stages. It requires precise modeling of signal temporal boundaries and an in-depth understanding of multiple actions. However, as demonstrated in our experiments, directly applying vision-based methods to WiFi CSI signals often yields suboptimal results due to the unique characteristics of wireless signals.

3 Method

3.1 Problem Definition

Given a dataset of long-term signals $\mathcal{D} = \{X_i\}_{i=1}^{n}$, where each signal instance X_i contains M_i action segments $Y_i = \{(s_m, e_m, c_m)\}_{m=1}^{M_i}$, with s_m representing the start time, e_m the end time, and c_m the corresponding action category, our task is to detect all action segments in Y_i based on the input signal X_i.

3.2 Some preliminary Results

Before delving into the detailed design of our proposed methods, we present some preliminary results that motivate our solution.

We empirically observed that CSI signals are highly complex, as illustrated in Fig. 3, primarily due to the inherently noisy characteristics caused by multi-path effects (Yang, Zhou, and Liu 2013). Despite this complexity, our task involves identifying the precise start and end times of each potential activity. We hypothesize that these temporal boundaries are largely driven by rapid changes in the signal, which can be effectively captured by high-frequency information. Additionally, identifying the specific activity category requires a comprehensive understanding of the semantic information across an entire input segment, which can be modeled by low-frequency information. Low-frequency components are more effective for distinguishing between different types of activities, as they capture the broader, more stable patterns in the signal that are crucial for accurate classification within the detected temporal boundaries. Similar approaches to signal analysis from a frequency perspective has also been validated in the vision community (Wang et al. 2022; Li et al. 2023).

Existing work (Wang et al. 2022; Li et al. 2023) has shed light on network design from a frequency perspective. Self-attention, by focusing on all parts of the input sequence, tends to aggregate information across the entire sequence, smoothing out variations and emphasizing the global structure. This behavior is akin to a low-pass filter in signal processing, making self-attention particularly effective at extracting low-frequency features that capture broader, more stable patterns—ideal for tasks requiring a global context or semantic understanding. Convolutional layers, on the other

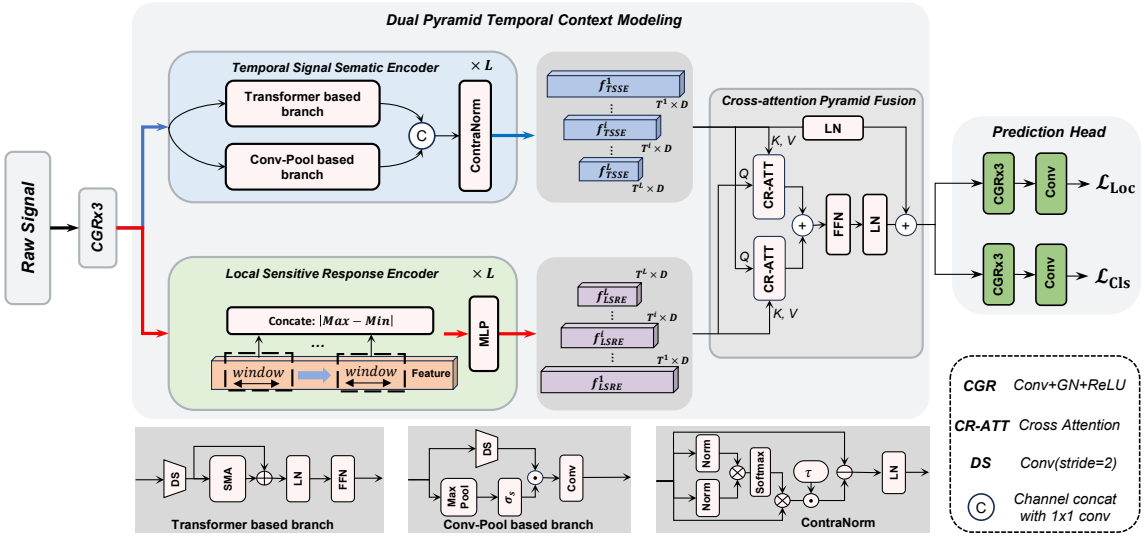

Figure 2: Overview of our network. Given a raw signal input, we employ 3 CGR layers to project the signal and generate output feature through dual pyramid temporal context modeling. Finally, the features are processed by the prediction head to transform into TAD results.

hand, especially with small kernel sizes, are designed to capture local patterns by applying filters over small receptive fields. These operations are sensitive to sharp transitions and fine details, characteristic of high-frequency components. When combined with pooling, which down-samples the input, convolutional layers further amplify high-frequency features, making them effective at detecting edges, textures, and other fine details.

To validate our hypothesis, we conducted preliminary studies from a frequency perspective. In our first case study, we designed two networks: one featuring a transformer backbone and the other a convolutional network backbone. Both networks were configured with comparable numbers of parameters and employed the same classification and localization heads. We assessed these networks using two metrics: localization mean Intersection over Union (mIoU), classification precision. The mIoU measures the average IoU between the predicted and ground truth (GT) boundaries, regardless of the activity labels, evaluating the network's capability to detect potential activities within the input signal. Precision is defined as the ratio of correctly predicted activity labels—those whose predicted boundaries overlap with the GT boundaries within a predefined tIoU range of [0.3:0.7:0.1]—to the total number of predictions. Precision evaluate the network's ability to accurately understand the context of the input signal, assuming that rough segmentation has already been achieved. The results, summarized in Table 1, indicate that the transformer-based solution, which focuses on learning global dependencies and capturing low-frequency semantic information, excels in classifying activity categories. Conversely, the convolutional network-based solution, adept at capturing local patterns and high-frequency features, performs better in identifying activity boundaries. To further validate this phenomenon, we conducted additional experiments. Initially, we transformed the input signal into the frequency domain using the Fast Fourier Transform (FFT). We then identified the cut-off frequency at the point where the power spectrum decreased by 6 dB and divided the frequency spectrum into two parts: low-frequency and high-frequency, based on this cut-off point. Subsequently, we obtained the low-frequency and high-frequency signals by transforming the respective frequency bands back to the time domain using the Inverse FFT (IFFT). We then input these transformed signals into our proposed model, details of which will be elaborated in the following section, and evaluated their results. The outcomes are consistent with our previous study, further confirming the role of different frequency components in the task of Temporal Activity Detection. More results could be found in the supplementary material at https://github.com/AVC2-UESTC/WiFiTAD.

Network	mIoU	Precision
Transformer	47.1	21.9
Convolutional Network	49.4	18.3
Frequency Band	**mIoU**	**Precision**
Low-Frequency Inputs	43.5	19.1
High-Frequency Inputs	44.6	15.8

Table 1: Preliminary Results and Analysis from the frequency perspective.

3.3 Model Overview

Given the distinct roles of frequency components in Temporal Activity Detection (TAD), we propose a frequency-aware learning framework. As shown in Figure 2, the input signal is first processed through three CGR (Conv+GroupNorm+ReLU) layers. The resulting features are then passed to a dual pyramid temporal context modeling module, which includes $L\times$ Temporal Signal Semantic Encoders (TSSE) and Local Sensitive Response Encoders (LSRE), followed by a cross-attention pyramid fusion module. The TSSE consists of a transformer branch and a Conv-

Pool branch, designed to learn low and high frequencies, respectively. These features are integrated via a ContraNorm module. The LSRE captures signal fluctuations in a learning-free manner, and the features from both encoders are aligned through a cross-attention pyramid fusion mechanism. Finally, a prediction head outputs the detection results for training and inference.

3.4 Dual Pyramid Temporal Context Modeling

The feature encoders starts from the projected feature $f \in \mathbb{R}^{T\times D}$, where T represents the signal timestamp points and D represents the channels. Through the local sensitive response encoder and temporal signal semantic encoders, two multi-scale feature pyramids are created.

Temporal Signal Semantic Encoder We designed two distinct network branches, each featuring core modules of self-attention and convolutional-pooling operations, tailored to preferentially learn low and high frequencies, respectively.

More specifically, given a feature $f \in \mathbb{R}^{T\times D}$, we employ two branches to process the feature. The first branch is transformer-based, where we introduce a novel Signed Mask-Attention (SMA) mechanism to enhance the extraction of low-frequency features in the signal. These low-frequency features serve as crucial cues for achieving a comprehensive understanding of the semantic information across the entire segment of the input. For the input feature f, we divide it into multi-heads. For the *i*-th head $f_i \in \mathbb{R}^{T\times d_k}$, we compute the queries, keys, and values as follows:

$$Q_i = f_i W_i^Q, \quad K_i = f_i W_i^K, \quad V_i = f_i W_i^V \tag{1}$$

where $W_i^Q \in \mathbb{R}^{T\times d_k}$, $W_i^K \in \mathbb{R}^{T\times d_k}$, and $W_i^V \in \mathbb{R}^{T\times d_k}$ are projection matrices, with d_k representing the projection dimension, typically defined as $d_k = \frac{D}{M}$, where M denotes the number of attention heads. The Signed Mask-Attention matrix can be expressed as:

$$\mathbf{A} = \sigma_t(\|Q + K\|_1 W_{\varnothing}^T) \odot \sigma_s(QK^T) \tag{2}$$

where σ_t denotes the tanh activation function, which outputs values in the range $[-1, 1]$, and σ_s denotes the sigmoid activation function, which outputs values in the range $[0, 1]$. Additionally, $W_{\varnothing}$ is a learnable matrix with the same dimensions as Q. These activation functions introduce nonlinearity into the learning process while also constraining the matrix values to prevent them from becoming excessively large. Our newly designed attention matrix leverages the information in Q and K more effectively, adjusting the magnitude of the original attention mechanism. This approach emphasizes important areas while downplaying unimportant ones, enhancing the model's focus on critical regions.

$$\text{SMA}(Q_i, K_i, V_i) = \text{Softmax}\left(\mathbf{A_i}/d_k\right) \times V_i \tag{3}$$

Subsequently, this result is added to the original features, followed by the application of a feedforward network. The output is then added again to obtain the final vector. Therefore, the SMA branch can be computed as:

$$f_{sma} = \text{FFN}(\text{LN}(\text{SMA}(DS(f)) + DS(f))) \tag{4}$$

where LN is the LayerNorm, DS is the down-sampling layer, FFN is the FeedForward Network.

Additionally, we designed a convolutional network-based branch to more effectively extract the high-frequency information from the input signal. This is accomplished through the use of convolution and max-pooling operations. Specifically, we have:

$$f_{pool} = Conv(\sigma_s(\text{Maxpool}(f)) \odot \text{DS}(f)), \tag{5}$$

Finally, we employ a mixed module that combines channel-wise concatenation with ContraNorm (Guo et al. 2023) to aggregate the features from the two distinct branches. The ContraNorm operation has been demonstrated to effectively disentangle representations in the embedding space, thereby enhancing generalization performance.

$$\begin{aligned} f_c &= Conv_{1x1}(Cat[f_{sma}, f_{pool}]) \\ f_{TSSE} &= f_c - \tau \cdot \text{softmax}(f_c \times f_c^\top) f_c \end{aligned} \tag{6}$$

where $Conv_{1x1}$ denotes a convolutional layer with a kernel size of 1 and a stride of 1, the channel is set as D. The channel-wise concatenation module is used to fuse f_{ssa} and f_{pool}, τ is the parameter.

Our TSSE design allows our network to leverage the unique strengths of each frequency band, thus enhancing the precision and robustness of our temporal activity detection system. By fusing the separately learned high and low-frequency features, we create a more comprehensive representation of the signal. This fused representation effectively combines the detailed temporal boundaries derived from the high-frequency features with the contextual insights from the low-frequency features. As a result, our model excels in accurately identifying both the timing and nature of activities, even amidst complex and noisy environments, as evidence by the numerical results in the experimental section.

Local Sensitive Response Encoder We further design another encoder to capture the fluctuation information to enhance the localization. As depicted in Fig. 2, the LSRE employs a channel-wise window to slide on temporal axis to extract regional information. The scale of window is determined by the LSRE order l. To capture the regional fluctuations, we compute the maximum and minimum values within the window to obtain the aggregated feature. This process can be formalized as:

$$f_w = Concat_t\{\max(f_{in}[t : t+2^l]) - \min(f_{in}[t : t+2^l])\}_{t=0}^{T-w} \tag{7}$$

where f_{w} represents the aggregated feature, and the t denotes the window's position, $Concat_t$ means to concatenate the features in the temporal dimension. The sliding window effectively captures the local fluctuation by computing the difference between the maximum and minimum values within each window, thereby enhancing the regional saliency of the input signal. Most importantly, this operation is achieved in a learning-free manner. In this way, it reduces complexity and computational overhead, enabling faster processing and lower memory usage. It's also less

prone to overfitting, making it an efficient and generalizable method for capturing essential signal characteristics. For the aggregated features, a MLP layer is utilized to transform the raw regional difference into a more robust and discriminative feature space

$$f_{\text{LSRE}} = \text{MLP}(f_{\text{w}}) \tag{8}$$

where the output feature f_{LSRE} encapsulates crucial regional saliency information, serving as a robust and informative input for subsequent processing.

Cross-attention Pyramid Fusion By obtaining the outputs from each encoder, we construct the feature pyramids. One pyramid contains the multi-scale output of TSSE, denoted as $Set_T = \{f^1_{TSSE}, f^2_{TSSE}, \ldots, f^L_{TSSE}\}$, and the other is the LSRE pyramid, denoted as $Set_L = \{f^1_{LSRE}, f^2_{LSRE}, \ldots, f^L_{LSRE}\}$. The scale of these pyramids is determined by a stride of 2. To efficiently and comprehensively fuse these two pyramid sets, we designed the Cross-Attention Pyramid Fusion, which aims to align and integrate the feature information from both pyramids. Specifically, for the features at the l^{th} level, this process can be formulated as:

$$\begin{cases} f^l_{c1} \leftarrow \text{CrossAttention}(f^l_{LSRE}, f^l_{TSSE}), \\ f^l_{c2} \leftarrow \text{CrossAttention}(f^l_{TSSE}, f^l_{LSRE}) \end{cases} \tag{9}$$

where fl_{LSRE} represents the feature at the l^{th} level of the LSRE pyramid, and f^l_{TSSE} represents the corresponding feature at the l^{th} level of the TSSE pyramid. The CrossAttention operation between these features enables the model to capture and exchange complementary information from both pyramids. This bi-directional attention mechanism ensures that the fused features f^l_{c1} and f^l_{c2} incorporate both local and global context, enhancing the overall representation capability of the model.

After that, we have post-processed the two features and add additional f^l_{TSSE} through layernorm for facilitating training

$$f^l_{det} = LN(FFN(f^l_{c1} + f^l_{c2}) + LN(f^l_{TSSE}) \tag{10}$$

where f^l_{det} represents the final fused feature used for detection at the l^{th} level of the pyramid.

3.5 Prediction Head

We build the prediction head to process the pyramid features across multi levels. It can be divided into two symmetric classification branch and localization branch which are both realized by 3 CGR layers with a single convolution layer. The difference between the two branches is the classification branch conv projects the feature $f^l_{det} \in \mathbb{R}^{T^l \times D}$ into the class score predictions $f^l_{cls} \in \mathbb{R}^{T^l \times cls}$ and localization branch conv projects the feature $f^l_{reg} \in \mathbb{R}^{T^l \times 2}$ into the boundary locations, respectively at different time stamps.

For an instant t^l in the l_{th} level, the prediction head estimates the boundary distance $\hat{d}^l_{st}$ and $\hat{d}^l_{et}$, the class scores $\hat{c}^l_t$ of all categories with the background is also obtained. Then the candidate activity segments $\hat{o}^l_t = (\hat{a}^l_t, \hat{s}^l_t, \hat{e}^l_t)$ can be decoded by

$$\hat{a}^l_t = \arg\max(\hat{c}^l_t), \quad \hat{s}^l_t = t^l - \hat{d}^l_{st}, \quad \hat{e}^l_t = t^l + \hat{d}^l_{et} \tag{11}$$

3.6 Training and Inference

Training In the training stage, the network outputs predicted candidates $\hat{o}$, we optimize the model by aligning the predicted results with the ground truth annotations. The objective function of the proposed DPWiT optimization follows the design in (Zhang, Wu, and Li 2022), which has two sub-functions, the first sub-function $\mathcal{L}_{\text{Cls}}$ is a focal loss(Lin et al. 2017) for classification, the second sub-function $\mathcal{L}_{\text{Loc}}$ is a DIoU loss(Zheng et al. 2020) for distance regression. The objective function is defined as

$$\mathcal{L} = \sum_t (\mathcal{L}_{\text{Cls}} + \lambda \mathcal{L}_{\text{Loc}})/T_+ \tag{12}$$

where T_+ is the number of positive predictions and λ is an hyper-parameter to modulate the ration of $\mathcal{L}_{\text{Cls}}$ and $\mathcal{L}_{\text{Loc}}$.

Inference At inference stage, the predicted candidates $\hat{o}$ with classification scores higher than threshold β and their corresponding instances are kept. We then assemble all predictions and process them with Soft-NMS(Bodla et al. 2017) to duplicate overlapped instances.

4 Experiments

4.1 Dataset

We collected CSI samples to create the dataset used in our experiments. The CSI data was gathered in an empty office room measuring $7m \times 12m \times 2.5m$. Our test bed consists of two laptops equipped with commercial Intel 5300 NICs, functioning as the transmitter (TX) and receiver (RX), respectively. Both the TX and RX each have one antenna, with thirty sub-channels available. Three student volunteers participated in the experiment, where they were asked to randomly perform a set of predefined daily activities between the TX and RX. The key statistics of our dataset are summarized in Table 2. The device's sampling frequency is 100Hz. The signal recordings cover seven daily activities—walking, running, jumping, waving, falling, sitting, and standing—and include 553 untrimmed samples with a total of 2,114 activity instances. Each instance is meticulously annotated with its start time, end time, and activity category. The whole dataset is split with a 7:3 ratio as the training and testing subsets.

Category	Walk	Run	Jump	Wave	Fall	Sit	Stand
Num.	394	361	347	335	332	225	120
Avg.(s)	16	17	13	18	13	13	11
Max.(s)	30	25	20	30	25	40	20
Min.(s)	10	5	5	10	5	5	5

Table 2: Details of the dataset

4.2 Experimental Setups

Metrics and Baselines We evaluate our model's performance using Mean Average Precision (mAP) at several temporal Intersection over Union (tIoU) thresholds. tIoU is defined as the ratio of the intersection to the union of two temporal windows, determining localization accuracy. If tIoU

Model	$mAP_{0.3}$	$mAP_{0.4}$	$mAP_{0.5}$	$mAP_{0.6}$	$mAP_{0.7}$	mAP_{avg}	GFlops	Time/ms
Baseline-ResNet1d(He et al. 2016)	19.7	19.2	16.7	10.9	7.5	14.8	0.29	**7.8**
Baseline-THAT(Li et al. 2021a)	21.2	20.6	16.2	10.6	6.7	15.1	1.64	8.2
AFSD(Lin et al. 2021)	46.6	45.4	42.4	37.9	25.4	39.5	91.0	71.8
BREM(Hu et al. 2022)	48.8	46.4	43.5	38.7	29.7	41.4	205.1	74.7
ActionFormer(Zhang, Wu, and Li 2022)	60.8	58.6	56.6	50.5	35.8	52.5	148.3	98.5
TADTR(Liu et al. 2022)	63.3	59.9	57.4	51.8	38.2	54.1	185.4	75.7
Tridet(Shi et al. 2023)	62.4	60.8	58.5	53.8	39.5	55.0	239.5	117.4
TemporalMaxer(Tang, Kim, and Sohn 2023)	64.5	62.0	60.0	54.9	40.5	56.4	189.6	76.1
DyFADet(Yang et al. 2024)	66.8	64.2	62.4	56.4	40.1	58.0	304.0	106.0
Ours	**85.5**	**83.0**	**77.3**	**72.1**	**54.5**	**74.5**	**44.1**	61.8

Table 3: Comparison of the model mAPs, GFLOPs and inference time(ms) of different methods.

exceeds a threshold, the window is validated for correct action classification. mAP is then obtained by averaging the Average Precision (AP) across all categories. Following Thumos14 (Idrees et al. 2017), tIoU thresholds range from 0.3 to 0.7 in steps of 0.1.

The baseline comparison methods are primarily adapted from the vision community and include the following: **AFSD** (Lin et al. 2021), which learns salient boundary features without anchors; **BREM** (Hu et al. 2022), which predicts multi-scale boundary quality to improve proposal scores; **ActionFormer** (Zhang, Wu, and Li 2022), which combines local self-attention with multiscale features for long-range temporal context; **TADTR** (Liu et al. 2022), which uses temporal deformable attention to focus on key snippets; **Tridet** (Shi et al. 2023), which models action boundaries with a relative probability distribution; **TemporalMaxer** (Tang, Kim, and Sohn 2023), which emphasizes feature extraction while minimizing long-term context modeling; and **DyFADet** (Yang et al. 2024), which introduces Dynamic Feature Aggregation to adapt kernel weights and receptive fields over time. To ensure fair comparison, we modified their pipelines to fit signal data and replaced their backbones with the same encoder used in our network. Additionally, we implemented two sliding-window baseline methods: one based on a Convolutional Network and the other on a Transformer. The input is pre-segmented according to ground truth annotations, with a window size matching the maximum ground truth boundary. These networks are trained to classify activity categories and background, and during inference, a sliding window approach is used to identify activities within each segment.

Implementation Details The model is implemented in PyTorch, using Adam as the optimizer with an initial learning rate of 4e-5 and a weight decay coefficient of 1e-3. Training was conducted on a workstation equipped with an Intel(R) Xeon(R) Gold 5218R CPU @ 2.10GHz and two Nvidia 3090 GPUs, with a batch size of 2. Training our dataset required 40 epochs, taking approximately 4 hours to complete. During the inference stage, the model outputs were processed by Soft-NMS, with a sigma value of 0.95 and a confidence threshold of 0.01. For both training and inference, single signal samples were divided into clips, each with a length of 4096 time stamps (approximately 41 seconds, covering over 2 activities), with a stride of 0.5. We utilized 8 TSSE and LSRE backbones as feature encoders, and the output features from the last 4 layers were used for detection. Regarding the hyperparameters, the coefficient λ in the objective function was set to 10, the scale τ in ContraNorm was set to 0.1, and the confidence threshold in focal loss β was set to 0.9.

Method	$mAP_{0.3}$	$mAP_{0.5}$	$mAP_{0.7}$	mAP_{avg}
Decomposed Frequency-aware Learning				
w/o Transformer-based branch	77.8	71.9	49.9	68.6
w/o Conv-Pool based branch	78.9	70.7	46.2	66.6
Dual Encoder				
LSRE	33.1	26.6	12.3	24.7
TSSE	79.0	70.5	44.6	66.2
Design in LSRE				
min	80.0	73.9	46.4	68.8
mean	83.6	74.1	47.9	70.3
max	82.1	75.1	50.5	71.3
max-min	**85.5**	**77.3**	**54.5**	**74.5**
Pyramid Fusion				
Pyramid-wise Add	80.8	72.3	48.6	69.0
Signed Mask-Attention				
Self-Attention	70.7	64.0	35.1	59.4
Proposed Method				
Ours	**85.5**	**77.3**	**54.5**	**74.5**

Table 4: Ablation studies from various aspects.

4.3 Main Results

We report our main results in Table 3, which demonstrate that our model outperforms all baselines and achieves state-of-the-art performance on the dataset. Notably, DyFADet achieves close to 60% accuracy, while TemporalMaxer, using MaxPooling for dimensionality reduction, loses semantic information, resulting in a lower mAP of 56.4%. Single-stage models like AFSD and BREM achieve 39.5% and 41.4% mAP, respectively, while TadTR, adapted from DETR, scores 54.1%. These results highlight the critical role of model design in temporal activity detection using WiFi-based data. Cross-person evaluation is provided in the supplementary material at https://github.com/AVC2-UESTC/WiFiTAD.

4.4 Ablation Study

To further verify the efficacy of our contributions, we conduct extensive ablation studies on Dataset for our method in

Table 4.

First, we evaluated the benefits of the **decomposed frequency-aware learning** mechanism. Removing either the Transformer or Convolution-pooling branch significantly reduced performance, confirming the value of high and low-frequency decomposition. Second, we assessed the impact of **Dual Encoders**. While the system with only LSRE struggled, combining it with TSSE showed substantial improvements. Third, we examined the **Design in LSRE** and found that capturing both maximum and minimum values empirically delivered the best results. Fourth, we analyzed the **pyramid fusion** mechanism. Conventional methods, like pairwise addition, performed poorly, underscoring the need to model feature interactions. Finally, replacing the proposed **Signed Mask-Attention** with conventional self-attention led to significantly worse results, demonstrating its effectiveness in focusing on informative input areas.

Error Type	Definition
Background	Correctly labeled predictions with an IoU of less than 0.1
Localization	Correctly labeled predictions with an IoU between 0.1 and 0.5
Wrong Label	Predictions with more than 0.5 IoU but with incorrect class labels
Confusion	Incorrectly labeled predictions with an IoU between 0.1 and 0.5
Double Detection	Correctly labeled predictions with an IoU over 0.5, but the GT is already matched with a higher-scoring wrong prediction

Table 5: Definition of error metrics in the False Positive Analysis.

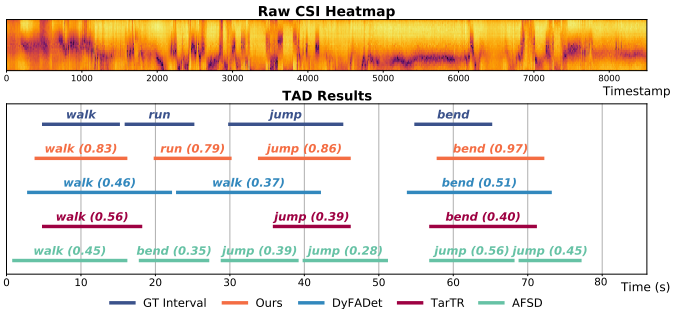

Figure 3: Prediction Visualizations form different methods.

4.5 Analyze

Visual Analyze We present a qualitative visualization of our dataset in Fig. 3. The figure shows the ground truth activities in the CSI data alongside the best-predicted proposals from four different models. As observed, our method accurately identifies candidate actions and provides reliable temporal locations. In contrast, DyFADet correctly identifies the location but misclassifies the action, TADTR misses the location, and AFSD produces scattered predictions that fail to form a coherent result.

False Positive Analyze To further diagnose the errors predicted by our proposed method, we follow the process outlined in (Alwassel et al. 2018) to conduct a False Positive (FP) Analysis. Specifically, this framework involves five error metrics, which are defined in Table 5. The results are

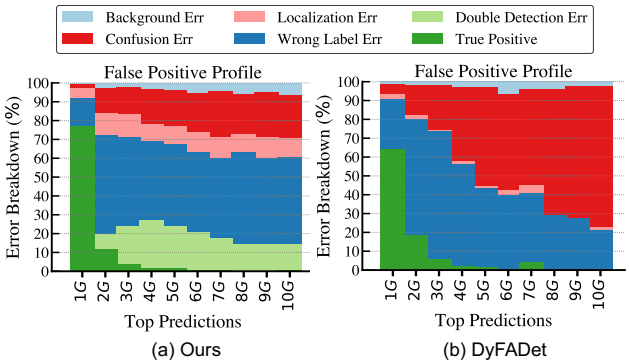

Figure 4: False Positive Profiling on wireless dataset, we use DPWiT as the model and compare our net with DyFADet with their best mAP results.

illustrated in Fig. 4. This analysis allows us to evaluate the error profile of the top-10G predictions, where G represents the number of ground truth instances. We select the top predictions in a per-class manner, meaning we choose the top-$10G_j$ predictions from class j, where G_j is the number of instances in class j. Additionally, to observe the trend of each error type, we divide the top-10G predictions into ten equal splits and examine the breakdown of the five FP error types in each split. Compared to the leading competing method, DyFADet, our method generates more informative predictions that improve positive localization and reduces confusion errors, thereby minimizing false positive judgments.

5 Conclusion

This work tackles the challenging problem of wireless temporal activity detection. We propose a Dual Pyramid Network that integrates high- and low-frequency features via a Temporal Signal Semantic Encoder and refines them using a Local Sensitive Response Encoder and cross-attention pyramid fusion. To support this task, we introduce a dataset with 2,114 activity segments from 553 WiFi CSI samples. Extensive experiments show our method significantly outperforms existing baselines, advancing temporal activity detection.

6 Acknowledgments

This work was supported by the Key Program for International Cooperation of Ministry of Science and Technology of China (No.2024YFE0100700) and the National Natural Science Foundation of China (NSFC) under Grant 62020106011. The work was also supported in part by National Natural Science Foundation of China (NSFC) under Grant No. 62171302, the 111 Project under Grant No. B21044 and Sichuan Science and Technology Program under Grant No. 2023NSFSC1965.

References

Alwassel, H.; Caba Heilbron, F.; Escorcia, V.; and Ghanem, B. 2018. Diagnosing Error in Temporal Action Detectors. In *The European Conference on Computer Vision (ECCV)*.

Bodla, N.; Singh, B.; Chellappa, R.; and Davis, L. S. 2017. Soft-NMS–improving object detection with one line

of code. In *Proceedings of the IEEE international conference on computer vision*, 5561–5569.

Chavarriaga, R.; Sagha, H.; Calatroni, A.; Digumarti, S. T.; Tröster, G.; Millán, J. d. R.; and Roggen, D. 2013. The Opportunity challenge: A benchmark database for on-body sensor-based activity recognition. *Pattern Recognition Letters*, 34(15): 2033–2042.

Chen, G.; Zheng, Y.-D.; Wang, L.; and Lu, T. 2022. DCAN: improving temporal action detection via dual context aggregation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, 248–257.

Chen, Z.; Zhang, L.; Jiang, C.; Cao, Z.; and Cui, W. 2018. WiFi CSI based passive human activity recognition using attention based BLSTM. *IEEE Transactions on Mobile Computing*, 18(11): 2714–2724.

Gu, Y.; Zhang, C.; Wang, Y.; Liu, Z.; Ji, Y.; and Li, J. 2019. A Contactless and Fine-Grained Sleep Monitoring System Leveraging WiFi Channel Response. In *ICC 2019 - 2019 IEEE International Conference on Communications (ICC)*, 1–5.

Guo, X.; Wang, Y.; Du, T.; and Wang, Y. 2023. ContraNorm: A Contrastive Learning Perspective on Oversmoothing and Beyond. In *The Eleventh International Conference on Learning Representations*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Hu, J.; Guo, C.; Zhuang, L.; Wang, B.; Ge, T.; Jiang, Y.; and Li, H. 2022. Estimation of Reliable Proposal Quality for Temporal Action Detection. In *Proceedings of the 30th ACM International Conference on Multimedia*, MM '22, 6685–6695. New York, NY, USA: Association for Computing Machinery. ISBN 9781450392037.

Idrees, H.; Zamir, A. R.; Jiang, Y.-G.; Gorban, A.; Laptev, I.; Sukthankar, R.; and Shah, M. 2017. The thumos challenge on action recognition for videos "in the wild". *Computer Vision and Image Understanding*, 155: 1–23.

Kong, Y.; and Fu, Y. 2022. Human action recognition and prediction: A survey. *International Journal of Computer Vision*, 130(5): 1366–1401.

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25.

Li, A.; Zhang, L.; Liu, Y.; and Zhu, C. 2023. Feature modulation transformer: Cross-refinement of global representation via high-frequency prior for image super-resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 12514–12524.

Li, B.; Cui, W.; Wang, W.; Zhang, L.; Chen, Z.; and Wu, M. 2021a. Two-stream convolution augmented transformer for human activity recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 286–293.

Li, W.; Bocus, M. J.; Tang, C.; Piechocki, R. J.; Woodbridge, K.; and Chetty, K. 2021b. On CSI and passive Wi-Fi radar for opportunistic physical activity recognition. *IEEE Transactions on Wireless Communications*, 21(1): 607–620.

Lin, C.; Xu, C.; Luo, D.; Wang, Y.; Tai, Y.; Wang, C.; Li, J.; Huang, F.; and Fu, Y. 2021. Learning salient boundary feature for anchor-free temporal action localization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 3320–3329.

Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; and Dollár, P. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, 2980–2988.

Liu, J.; Zeng, Y.; Gu, T.; Wang, L.; and Zhang, D. 2021. WiPhone: Smartphone-based Respiration Monitoring Using Ambient Reflected WiFi Signals. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 5(1).

Liu, X.; Wang, Q.; Hu, Y.; Tang, X.; Zhang, S.; Bai, S.; and Bai, X. 2022. End-to-end temporal action detection with transformer. *IEEE Transactions on Image Processing*, 31: 5427–5441.

Meng, W.; Liu, Z.; Li, B.; Cui, W.; Zhou, J. T.; and Zhang, L. 2023. GrapHAR: A Lightweight Human Activity Recognition Model by Exploring the Sub-carrier Correlations. *IEEE Transactions on Wireless Communications*.

Shi, D.; Zhong, Y.; Cao, Q.; Ma, L.; Li, J.; and Tao, D. 2023. Tridet: Temporal action detection with relative boundary modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18857–18866.

Shou, Z.; Wang, D.; and Chang, S.-F. 2016. Temporal action localization in untrimmed videos via multi-stage cnns. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1049–1058.

Sun, Z.; Ke, Q.; Rahmani, H.; Bennamoun, M.; Wang, G.; and Liu, J. 2022. Human action recognition from various data modalities: A review. *IEEE transactions on pattern analysis and machine intelligence*.

Tang, T. N.; Kim, K.; and Sohn, K. 2023. TemporalMaxer: Maximize Temporal Context with only Max Pooling for Temporal Action Localization. *arXiv preprint arXiv:2303.09055*.

Tian, Y.; Lee, G.-H.; He, H.; Hsu, C.-Y.; and Katabi, D. 2018. RF-based fall monitoring using convolutional neural networks. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 2(3): 1–24.

Wang, B.; Zhao, Y.; Yang, L.; Long, T.; and Li, X. 2023a. Temporal action localization in the deep learning era: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Wang, F.; Gao, Y.; Lan, B.; Ding, H.; Shi, J.; and Han, J. 2023b. U-Shape Networks Are Unified Backbones for Human Action Understanding From Wi-Fi Signals. *IEEE Internet of Things Journal*.

Wang, P.; Zheng, W.; Chen, T.; and Wang, Z. 2022. Anti-Oversmoothing in Deep Vision Transformers via the Fourier Domain Analysis: From Theory to Practice.

Xia, K.; Wang, L.; Zhou, S.; Zheng, N.; and Tang, W. 2022. Learning To Refactor Action and Co-Occurrence Features for Temporal Action Localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13884–13893.

Xiao, C.; Lei, Y.; Ma, Y.; Zhou, F.; and Qin, Z. 2020. DeepSeg: Deep-learning-based activity segmentation framework for activity recognition using WiFi. *IEEE Internet of Things Journal*, 8(7): 5669–5681.

Yang, L.; Peng, H.; Zhang, D.; Fu, J.; and Han, J. 2020. Revisiting anchor mechanisms for temporal action localization. *IEEE Transactions on Image Processing*, 29: 8535–8548.

Yang, L.; Zheng, Z.; Han, Y.; Cheng, H.; Song, S.; Huang, G.; and Li, F. 2024. DyFADet: Dynamic Feature Aggregation for Temporal Action Detection. In *European Conference on Computer Vision (ECCV)*.

Yang, Z.; Zhou, Z.; and Liu, Y. 2013. From RSSI to CSI: Indoor localization via channel response. *ACM Computing Surveys (CSUR)*, 46(2): 1–32.

Yousefi, S.; Narui, H.; Dayal, S.; Ermon, S.; and Valaee, S. 2017. A survey on behavior recognition using WiFi channel state information. *IEEE Communications Magazine*, 55(10): 98–104.

Zeng, Y.; Wu, D.; Gao, R.; Gu, T.; and Zhang, D. 2018. FullBreathe: Full Human Respiration Detection Exploiting Complementarity of CSI Phase and Amplitude of WiFi Signals. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 2(3).

Zhang, C.-L.; Wu, J.; and Li, Y. 2022. Actionformer: Localizing moments of actions with transformers. In *European Conference on Computer Vision*, 492–510. Springer.

Zheng, Z.; Wang, P.; Liu, W.; Li, J.; Ye, R.; and Ren, D. 2020. Distance-IoU loss: Faster and better learning for bounding box regression. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, 12993–13000.