Abstract—Radar-based human activity recognition is attracting a wide range of interest from both industry and academia because of its through-wall ability, privacy-preserving capability, and device-free detection. Currently, most radar-based systems consider signal analysis and feature extraction in the frequency domain or the temporal domain independently without fusing them together. In this paper, in order to model both frequency properties and temporal profiles of human activity, we proposed a spectro-temporal network (STnet) that integrates a temporal convolutional network (TCN) and a convolutional neural network (CNN). It can extract temporal patterns and micro-Doppler features from radar signals for human activity recognition. In the experiments, two radar sensors and one base station were used to build a low-power wireless radar sensor network. Fifteen activities were investigated in a real kitchen scenario by using this radar sensor network. Frequency spectrograms were obtained after signal processing using a short-time Fourier transform (STFT). They were further segmented using a short sliding window (2.5s), which enables a very small latency. The proposed STnet achieved 99.64% overall accuracy in testing, which is superior to the other three networks that we implemented in this work. Our work also can be used as a generic solution to other sensor-based (wearable sensors, WiFi CSI, etc.) activity recognition.

Index Terms—Human activity recognition, spectro-temporal network, radar sensor network, temporal convolutional network, micro-Doppler signatures

I. INTRODUCTION

HUMAN activity recognition has been gaining a lot of attention in recent years due to its wide range of applications in surveillance, human-computer interaction, healthcare, etc. For instance, in surveillance, thefts are possible to be detected by recognizing their abnormal activities [1]. In healthcare, sleep monitoring [2] helps people to improve their sleep conditions, fall detection [3] enables us to provide timely assistance for the elderly, and heartbeat monitoring [4] can provide stress and activity levels monitoring, and track heart rate during exercise. In human-computer interaction, people can contactlessly interact with electronic products such as AR/VR glasses by using gesture recognition [5], and electronic devices can perform automatic person authentication based on the biometrics extracted from human activity [6].

From the aspect of sensors, human activity recognition systems can be divided into two types: Devices-Bound and Devices-Free. In Devices-Bound systems, wearable sensors, including accelerometers, gyroscopes, and Bluetooth/WiFi receivers, are worn by people. Human activity recognition based on wearable sensors has advanced significantly because of widely used smartphones and smart bracelets. However, it is inconvenient for users to wear them all the time. The study in [7] found that the elderly are less likely to carry wearable sensors. Device-Free systems perform human activity recognition remotely without requiring users to wear any sensor by using cameras, radars, etc. Camera-based activity recognition is also well-developed because cameras are widely installed and videos are a visual form that can be easily perceived by our eyes. However, camera-based systems usually suffer from insufficient illumination which makes them unable to work in darkness and they have a narrow visual range. Both wearable sensor-based systems and camera-based systems are privacy-intrusive. For example, wearable sensor data usually contains the information of sensor IDs, which can infer the owners of the sensors; people’s faces may be exposed to cameras. However, radar-based approaches do not have these problems. Radar can provide contactless sensing without requiring users to carry any receivers. It does not intrude on users’ privacy [8]. It is light-insensitive and able to work in darkness. Radar has a certain ability of through-wall penetration that depends on its operating frequency. The same functionality is available with WiFi, but radar usually has wider bandwidths and higher spatial resolutions [9]. The development of radar manufacturing has made low-cost and low-power radar sensors more accessible.

Considering the above merits, radar-based human activity recognition is attracting a great deal of interests and becoming a research hotspot with a wide range of applications. Chen [10] proposed micro-Doppler for the first time in 2000. It refers to the additional modulations that are added to the main Doppler frequency shift. The modulations are created by additional movements of smaller parts of the target relative to the radar. Since then many researchers have begun to apply micro-Doppler signatures for object recognition and activity classification. For examples, micro-Doppler signatures have been used in breathing and heartbeat detection [11], human recognition [12], and fall detection [13], etc. Handcrafted features [14]–[17] were extracted from the frequency domain of radar signals, including frequency peaks, the mean of frequency, bandwidth, etc. However, handcrafted feature extraction is time and labor-consuming. It is unable to achieve online human activity recognition with handcrafted feature extraction. To solve this problem, several methods have been applied to perform automatic micro-Doppler feature extraction, including principal component analysis (PCA) [18], empirical mode decomposition (EMD) [19], [20], singular value decomposition (SVD) [21], etc. In recent years, there have been many advances in deep learning. CNN is one of the most popular

Fei Luo, Eliane Bodanese, Salabat Khan, and Kaishun Wu (corresponding author) are with the College of Computer Science and Software Engineering, Shenzhen University, Shenzhen, China. E-mail: luofei2018@outlook.com, eliane.bodanese@qmul.ac.uk, salabatwazir@gmail.com, wu@szu.edu.cn.
Deep learning structures, it has been frequently used in the classification of frequency spectrograms generated from radar signals for human activity recognition [22]–[24]. Automatic learning of frequency features of 2-dimensional spectrograms is feasible with CNNs due to their hierarchical structure. Although impressive performance has been achieved, these methods only focus on extracting or learning features on the frequency domain and they are not capable of fully capturing the temporal dynamics within the sequential radar signals.

In order to take advantage of the temporal dynamics of human activity, some efforts [25], [26] have been made by using dynamic time warping (DTW). DTW classifies signal sequences of different activities by measuring temporal similarity based on their alignments. However, DTW is usually computation-intensive, which makes it unscalable to a large dataset [27]. Hidden Markov models (HMMs) are also widely applied in temporal pattern recognition. An HMM is a generalization of a mixture model where the hidden variables (or latent variables) control the mixture component to be selected for each observation [28]. It provides a probabilistic framework for time series modelling of multivariate observations [29]. HMMs have been used to classify human motion based on micro-Doppler signatures [30]. However, HMMs are based on the Markov assumption [31], which states that the probability of being a given state at time \( t \) only depends on the state at time \( t - 1 \). This makes HMMs unable to model complex temporal patterns because many sequences have dependencies extending through several states. Recurrent neural networks (RNNs) become popular for time series modelling. By memorizing the previous information, RNNs can calculate the current output [32], which makes them useful for sequential learning. As a variant of RNNs, LSTM further introduces a gated memory cell to control the information flow in and out of each node, which enables them to maintain long term patterns. Some researchers have already applied LSTM to model the temporal profile of falling [33], human gait [34], and hand gestures [35], in order to classify them.

Although both frequency-based modelling and temporal modelling have been investigated for radar-based human activity recognition, it is still waiting for further research to perform spectro-temporal modelling that integrates frequency features on spectrograms and temporal features within sequential signals. Spectro-temporal modelling has been explored in audio analysis [36] and speech recognition [37]. Most current spectro-temporal models [38]–[40] are the combinations of CNNs and RNNs. In this paper, we proposed a spectro-temporal model (STnet) that consists of two streams, which are a spectro-stream and a temporal stream. The spectro-stream is a CNN that can extract micro-Doppler features in the frequency domain and the temporal stream is a temporal neural network (TCN) that can distill discriminative temporal patterns along with data sequences. A TCN can model the entire sequence length by stacking many dilated convolutions up to the entire length of the sequence [41]. TCNs outperform RNNs including LSTM across a diverse range of tasks and datasets [42], including music and language modelling, adding problem, trajectory classification [43], sequential image classification, etc. Based on the superiority of TCNs, we implemented a TCN to build the temporal stream of STnet.

In this paper, we use a wireless radar sensor network that contains two power-efficient and low-cost radar sensors. It can be applied in long-time continuous human activity recognition. Using this radar sensor network, fifteen activities were investigated in a kitchen setting. The segmentation of signals was performed in a short time window (2.5 seconds), enabling human activity recognition in near real-time. The STnet was proposed to perform human activity classification. By fusing frequency features and temporal patterns of human activity, our STnet outperforms the ResNet, TCN, and BiLSTM+ResNet that are implemented in this work. A preliminary version of this work has been presented as a conference paper [23]. In this work, we incorporate additional contents:

1) A two-stream model (STnet) that combines CNN and TCN was proposed for human activity recognition. The STnet outperforms the CNN in [23] with a wide margin. Apart from the STnet, we further implemented ResNet, TCN, and BiLSTM+ResNet. A comparison was made among them.

2) We described the signal processing including frequency analysis and signal segmentation in detail. The signal segmentation is performed continuously upon two data streams of two radars in the radar sensor network.

3) We analyzed the impact of sensor fusion. Since our radar sensor network contains two radar sensors. We compared the results of using two radars and one radar.

The remainder of this paper is organized as follows. Section II briefly reviews the related work. Section III introduces micro-Doppler theory in general. The architecture of the STnet is presented in Section IV. The experiments, data collection, and data processing are introduced in Section V. The performance evaluation is presented in Section VI. Finally, we conclude the paper in Section VII.

Note that although we focus on radar signal sequences in this paper, the proposed STnet can be applied as a generic solution to other sensor-based (wearable sensors, WiFi CSI, etc.) activity recognition on temporal sequences that contain rich frequency characteristics and temporal dynamics.

II. RELATED WORK

A wide range of research has been done for radar-based human activity recognition by using a wide variety of radar systems and techniques. In this section, we will review some work related to radar systems, feature extraction, and activity classification with machine learning.

A. Radar systems

Micro-Doppler-based human activity recognition is greatly influenced by radar systems. A radar system that operates at high frequency can detect micro-Doppler signatures with more detail since micro-Doppler signatures are represented in frequency scale [44]. Radar systems in a wide range of frequency bands including ultra-high frequency (UHF) band [45], S-band [46], C-band [47], X-band [48], K-band [49], and W-band [19] have been used in human activity recognition.
There has been some research using polarimetric radars [50]–[52]. Instead of transmitting and receiving radar signals with a single horizontal polarization, polarimetric radars transmit and receive radio waves with both horizontal and vertical polarizations. The work in [51] presents that the separability of different parts of human movements can be improved by using a fully polarimetric radar. The work in [52] shows cross-polarized observations of polarimetric radars emphasize the movements of legs and arms. Ultrasound radars also have been applied to detect human activity [53], [54]. Compared with electromagnetic radars, acoustic radar systems are lower-cost, and their signals are easier to process and are not affected by electromagnetic interference. But sound propagation is likely to be affected by factors such as humidity, atmospheric pressure, and temperature. Radar systems with different antenna configurations (monostatic radar, bistatic radar, multistatic radar) have been used in human activity recognition. Observing micro-Doppler signatures using monostatic radar systems is limited to the radial component. Radar cannot detect movement if the target moves perpendicular to the beam. Bistatic radar systems and multistatic radar systems can overcome the limitations of monostatic radar systems. The transmitter and receiver of a bistatic radar are located at different locations. Multistatic radar systems combine multiple radars to provide multiple observations of a target from multiple angles [55].

The above-mentioned radar systems have been used to recognize various human activities, including human gait, hand gestures, falling, respiration, etc. For example, Silvious et al. [56] used UHF band radars to analyze micro-Doppler characteristics of heartbeat, walking, etc. Lai et al. [57] used an L-band UWB radar to detect human respiration. Seifert et al. [58] used a K-band radar to observe and classify five different walking gait patterns. In [59], an FMCW radar system was used to collect six activities including walking, sitting, standing up, drinking water, picking an object from the floor, and falling across various people and locations. In [60], a MIMO UWB radar system was used to collect seven activities including stepping, picking up an object, waving, jumping etc., in a through-wall scenario. The radar sensor network built in this paper can be seen as a multistatic radar system that contains two C-band pulse-Doppler radars. It can be used to measure human activity from two different viewpoints, which are helpful to provide more information for human activity classification.

B. Feature extraction

In contrast to optical sensors, radar signals are more difficult for humans to interpret. In radar-based human activity recognition, micro-Doppler signatures and temporal characteristics are derived from radar signals collected from the subject or its movements. Time-frequency processing techniques, for example, Fourier transform [61] and wavelet transform [62], are widely used to represent micro-Doppler signatures. Handcrafted features [14]–[16] such as average Doppler frequency, Doppler bandwidth, and Doppler offset have been extracted from frequency spectrograms of radar signals. The handcrafted feature extraction method is too inefficient, time-consuming, and labour-intensive to process a large number of radar signals, let alone recognize human activity in real-time. Various methods have been applied in feature extraction of radar signals. Empirical mode decomposition (EMD) is an adaptive technique that decomposes a signal into time-frequency components called intrinsic mode functions (IMFs). In [63], [64], EMD was used for extracting micro-Doppler signatures with minimal interference and a clearer focus. Singular value decomposition (SVD) is used to reduce the dimension of the feature space [21], [65]. The output of SVD can provide information about target velocity, spectrum periodicity, and spectrum width [66]. Mel-frequency cepstrum (MFC) is usually used to present the short-term power spectrum of audio. MFCCs are coefficients that collectively make up an MFC. MFCCs are usually used in audio processing. Both radar signals and audios are time-series data, they have a similar structure. So some researchers [67], [68] have attempted to represent micro-Doppler signatures of human activity with MFCCs. As the most popular feature extraction tool, PCA has been widely used to extract human activity features from radar data [18], [69]. There are still many other feature extraction methods: Gabor wavelet filter, Independent Component Analysis (ICA), etc., that have been applied in feature extraction for radar-based human activity recognition.

C. Classification

Most existing work for radar-based human activity recognition adopts off-the-shelf classifiers. Traditional supervised algorithms, such as Support Vector Machine, Random Forests, Naïve Bayes, Decision Tree, K-Nearest Neighbor, Gaussian mixture model, and Linear Discriminant Analysis, have been widely applied to model human activity upon labeled radar signal sequences. However, these classification algorithms need features that are handcrafted or extracted using feature extraction methods. It is heavily dependent on the quality of the features for these algorithms to perform well. Deep learning has the ability of feature learning without relying on feature extraction methods. CNNs use multiple convolutional layers to capture structural features from a series of local small regions in an image. As frequency spectrograms generated from radar signals have a similar structure to images, CNNs have been used to learn micro-Doppler features upon these spectrograms. They outperform traditional supervised learning algorithms by a wide margin [70], [71]. However, radar echoes of human activity have high temporal sequentiality rather than the spatial grid structure in images and CNNs cannot take the advantage of such sequentiality [72]. RNNs are powerful in learning temporal patterns from time series. Recently, LSTM has been increasingly used to model temporal dynamics of human activity upon radar signals. For example, A stacked LSTM network has been used to capture temporal dependencies within radar signals for fall detection at home [33]. A network consisting of two LSTM layers has been built to differentiate walking from sitting and standing [73]. However, it is worthy to implement TCN for temporal modelling of radar-based human activity recognition because TCN is able
to capture longer temporal dependencies within time series without the problems of gradient vanishing and exploding in RNNs.

For taking the advantage of micro-Doppler signatures and temporal dynamics of radar signals, we built the STnet, which combines the local structural learning ability of CNNs and the temporal patterns distilling ability of TCN.

III. MICRO-DOPPLER THEORY

A. Micro-Doppler theory

A frequency shift to the reflected signals is generated when a subject is moving towards a radar sensor, and this phenomenon is the well-known Doppler Effect. The movement of smaller parts of the subject induces additional modulations to the main Doppler frequency shift, called the micro-Doppler effect [74], [75]. The observed micro-Doppler effect of an object or a process can generate some distinctive characteristics, which are called micro-Doppler signatures. Due to the fact that various parts of this complex object can move at different speeds and in different directions in relation to the radar. It provides the target’s micro-Doppler signatures, and can be used to recognize subjects and human activities. Usually, micro-Doppler signatures are presented in two-dimensional time-frequency space.

Fig. 1 shows the typical micro-Doppler spectrograms of five human activities generated from raw radar signals collected by the BumbleBee radars. These spectrograms are generated by using the STFT and they can be described and analyzed as follows:

**Walking**: a person walks forward and back. In Fig. 1(a), the fluctuations resulted by arms and legs are attached on the main Doppler frequency produced by the torso.

**Drinking**: picking up a cup from the table, drinking from it, then putting it down. In Fig. 1(b), the frequency 0 is surrounded by four lumps. Picking up the cup caused the first two. The front one reflects elbow movement, the latter one reflects arm movement. Putting the cup back results in the last two. Conversely, the front one represents arm movement, and the latter one represents elbow movement.

**Eating**: eating with a spoon, then putting it down. In Fig. 1(c), there are two lumps. Picking up the spoon generates the front one, and putting it down generates the later one.

**Standing**: standing up from a chair. In Fig. 1(d), this movement produces a big lump. The lower parts of the lump are more prominent than the upper parts.

**Sitting**: sitting down. Contrary to the standing, the upper part is more prominent than its lower part as shown in Fig. 1(e).

Spectrograms of these activities may not be the same as described above because of variables such as people’s position, habits, and the multipath effect, but spectrograms of one activity tend to be similar. Due to their simplicity, we only present spectrograms for five activities. There are many other activities that are more complex and scenario-dependent, but every activity has its own micro-Doppler signature and temporal patterns that can be used for recognition.

IV. SPECTRO-TEMPORAL MODELLING

A. The proposed STnet

We proposed a STNet for spectro-temporal modelling of human activity recognition using the radar sensor network. As shown in Fig. 2, the STnet has two streams, one stream is the spectro-stream, and another stream is the temporal stream. The features generated from two streams are finally fused together.
Let’s denote \( I \) as the input (fused spectrograms) of the STnet. The spectro-stream extracts micro-Doppler features \( F_s \) on the input \( I \).

\[
F_s = S(I),
\]

where \( S(\cdot) \) denotes the operations of the spectro-stream.

The temporal stream extracts temporal features \( F_t \) on the input \( I \).

\[
F_t = T(I),
\]

where \( T(\cdot) \) denotes the operations of the temporal stream.

After extracting micro-Doppler features and temporal features, we further conduct feature fusion.

\[
F_{st} = \text{Fuse}(F_s, F_t),
\]

where \( F_{st} \) denotes the fused features that consist of \( F_s \) and \( F_t \). \( \text{Fuse}(\cdot) \) denotes the operation of feature fusion.

### B. The Spectro-stream

The Spectro-stream in this study is a ResNet composed primarily of residual blocks. The residual block consists of stacked layers. The input of each block is added back to its output in order to create identity mappings [76]. Residual blocks can extend a network to be very deep by avoiding the vanishing gradient problem. The structure of a residual block can be illustrated in Fig. 3. It can be formulated as [77]:

\[
y = \mathcal{F}(x, \{W_i\}) + x,
\]

where \( x \) and \( y \) represent the input and output, \( \mathcal{F}(x, \{W_i\}) \) denotes the residual mapping to be learned, \( W_i \) denotes the learned weights in \( i \)-th layer within a block. A residual block usually contains two weight layers, so

\[
\mathcal{F}(x, \{W_i\}) = W_2 \sigma(W_1 x),
\]

where \( \sigma \) denotes the nonlinear activation (ReLU). The operation \( \mathcal{F} + x \) is performed by a shortcut connection and element-wise addition. ReLU is performed after the addition. In the implementation, the weight layers in each block are convolutional layers with the kernel size of \( 3 \times 3 \) and the step size of 1. Batch normalization is applied right after each convolution and before activation.

As shown as the spectro-stream in Fig. 2, one convolutional layer is used to extract shallow features. The first convolutional layer extracts features \( F_0 \) from the input \( x \).

\[
F_0 = \text{Conv}(x),
\]

where \( \text{Conv}(\cdot) \) denotes convolution operation. The shallow feature \( F_0 \) is used as the input to a set of residual blocks. Suppose the spectro-stream has \( n \) residual blocks, the extracted micro-Doppler features \( F_s \) of the \( n \)-th blocks can be obtained by

\[
F_s = \mathcal{H}_n(\mathcal{H}_{n-1}(\cdots(\mathcal{H}_1(F_0)\cdots))),
\]

where \( \mathcal{H}_n \) denotes the operations of \( n \)-th residual block.

### C. The temporal stream

The temporal stream is actually a TCN that consists of a set of stacked temporal residual blocks. The structure of the temporal residual block (TCN block) also can be represented as Fig. 3. Differently, a temporal residual block uses dilated causal convolutions to exponentially increase the receptive field for capturing long-term dependencies within a sequence. In dilated causal convolutions, the current output is calculated by using samples from previous time-steps with dilated causal convolutions.

For a 1-D sequence input \( x[n] \), the output \( y[n] \) is:

\[
y[n] = \sum_{k=0}^{N-1} c_k \cdot x[n - M \cdot k],
\]

where \( M \) is the dilation factor, \( N \) is the filter size of the dilated causal convolution, \( c_k \) is the \( k \)-th element in the filter. The receptive field of \( y \) has the size of \( M(N-1)+1 \). By stacking dilated causal convolutions, the TCN can achieve large receptive fields with a limited number of parameters. Fig. 4(a) shows a dilated causal convolution with dilation factors \( d = 1, 2, 4 \) and filter size \( k = 2 \).
TCNs consist of several stacked temporal residual blocks. Temporal residual blocks are characterized by a set of transformations \( \tau \). As shown in 4(b), a temporal residual block contains two groups including a dilated casual convolution layer, a ReLU activation layer, a weight normalization layer, and a dropout layer. In the implementation, the dimension of the input is firstly reduced by using an average pooling operation. Suppose the temporal stream has \( m \) temporal residual blocks, the output \( F_t \) of the \( m \)-th blocks can be obtained by

\[
F_t = \tau_m(\tau_{m-1} = \tau_m(\tau_{m-1}(\cdots(\tau_1(avgpool(x))\cdots))), (9)
\]

where \( \tau_m \) denotes the operations of \( m \)-th temporal residual block. \( avgpool(\cdot) \) denotes the average pooling operation. \( F_t \) denotes the temporal features extracted by the temporal stream.

### D. Feature fusion

A good feature fusion method is supposed to be capable of taking advantages of the static micro-Doppler features and dynamic temporal features to produce better hybrid features. Based on the architecture of the proposed STNet, we can obtain micro-Doppler features \( F_s \) from the spectro-stream and temporal features \( F_t \) from the temporal stream. However, \( F_s \) and \( F_t \) have different dimensions. \( F_s \) is three-dimensional (width, height, channel) and \( F_t \) is two-dimensional (length, channel). For feature fusion, we further conduct global average pooling to reduce the spatial dimension of \( F_s \) and use fully connected network to further distill \( F_t \), and the outputs of them are concatenated together. Suppose the dimension of \( F_s \) is \( h \times w \times d \), global average pooling reduces its size to \( 1 \times 1 \times d \) that each feature map is reduced by simply taking the average of all elements and its output is further reshaped to \( d \). The Equation 3 can be re-formulated as

\[
F_{st} = \text{Concat}(Gavgpool(F_s) + Fc(F_t)),
\]

\[
Gavgpool(F_s) = \left\{ f^0, f^1, \cdots, f^{d-1} \right\},
\]

\[
f^k = \frac{1}{h \times w} \sum_{i=1}^{h} \sum_{j=1}^{w} e_{ij}^k, k \in [0, 1, \cdots, d-1],
\]

where \( \text{Concat}(\cdot) \) denotes the concatenation operation, \( Fc(\cdot) \) denotes the operation of a fully connected layer, \( Gavgpool(\cdot) \) denotes the operation of global average pooling. \( f^k \) denotes the average of \( k \)-th feature map in \( F_s \), \( e_{ij}^k \) denotes the value at the \( i \)-th row and \( j \)-th column of \( k \)-th feature map.

### E. Implementation details

The implementation of our STNet is illustrated in Fig. 5. There are two radar nodes in our radar sensor network that allow us to observe human activity from two different perspectives. To fuse the signals generated by two radar sensors, we downsampled their spectrograms to the size of \( 80 \times 80 \), then overlapped them, resulting in an input of \( 80 \times 80 \times 2 \). The input can be seen as an image with only two channels. In the spectro-stream, input data is first processed by a 2D convolutional layer with a kernel size of \( 7 \times 7 \), a stride of 2, and an output channel size of 64. A max-pooling with a \( 2 \times 2 \) filter is used to reduce the dimension after it. Then three residual blocks are applied to extract deep micro-Doppler features. All these three residual blocks use \( 3 \times 3 \) kernels. The output channel size of ‘Residual Block 1’ and ‘Residual Block 2’ are both 64 and that of ‘Residual Block 3’ is 128. We applied batch normalization after each 2D convolutional layer. In the temporal stream, an average pooling with a \( 2 \times 2 \) filter is firstly used to reduce the dimension size of the input. After it, seven temporal residual blocks are applied to extract temporal features of human activity. All temporal residual blocks use \( 1 \times 7 \) convolution kernels and their output channel size is 25. However, these temporal residual blocks have different dilation factors, which are \{1, 2, 4, 8, 16, 32, 64\} respectively. For fusing micro-Doppler features and temporal features, a global average pooling is applied after the spectro-stream to generate 128 units as the output of the spectro-stream has 128 channels, and a fully connected layer with 128 hidden units is applied after the temporal stream. Finally, all these units are concatenated together for feature fusion.

In the training process, dropout with the initialized rate of 0.2 is applied for the temporal stream to reduce overfitting. Fig. 4(b) shows that each dilated casual convolutional layer is followed by a dropout layer. The optimization function was Adam [78] with an initial learning rate of 0.0001. A batch size of 128 is used.

### V. Experiments

#### A. Experiment setup and data collection

Two BumbleBee radars are used in our low-power, low-cost radar sensor network [79] (see Fig. 6). The Bumblebee radar operates at 5.8 GHz and is a low-power Pulse Doppler radar. Its coverage is up to 10 meters. The input can be seen as an image with only two channels. As it only consumes 12 mAh, it can operate at full duty cycle for about 8 days when using 1.5v alkaline batteries with a capacity of 2400 mAh [23]. The main specifications of the Bumblebee radar are shown in Table I. A BumbleBee radar usually works with a TelosB mote. The TelosB mote offers low-power wireless communication (IEEE 802.15.4). It is operating on TinyOS, which is an open-source operating system that supports large-scale, self-assembling sensor networks.

As shown in Fig. 6, the sensor network consists of one base station and two nodes (‘Node 1’ and ‘Node 2’). Each node contains a TelosB mote and a Bumblebee radar. A Bumblebee radar sensor is connected to a TelosB mote that transmits radar signals via ZigBee to the base station.
Fig. 5. Implementation of our STnet

**TABLE I**

**BUMBLEBEE RADAR SPECIFICATIONS**

<table>
<thead>
<tr>
<th>Feature</th>
<th>Specification</th>
</tr>
</thead>
<tbody>
<tr>
<td>Center frequency</td>
<td>5.8 GHz</td>
</tr>
<tr>
<td>Coverage pattern</td>
<td>60 degree conical coverage pattern</td>
</tr>
<tr>
<td>Detection range</td>
<td>Up to 10m</td>
</tr>
<tr>
<td>Range gate sharpness</td>
<td>0.2m</td>
</tr>
<tr>
<td>Total power draw</td>
<td>About 12 mA</td>
</tr>
<tr>
<td>Antenna</td>
<td>Onboard antenna</td>
</tr>
<tr>
<td>Coherent output</td>
<td>I &amp; Q channels</td>
</tr>
<tr>
<td>Responds to radial velocity</td>
<td>2.6 cm/s to 2.6 m/s</td>
</tr>
</tbody>
</table>

Fig. 6. The proposed radar sensor network

We deployed the radar sensor network in a kitchen scenario. Fig. 7(a) shows the plan of the kitchen. A pair of sensor nodes (green rectangles in Fig. 7(a)) were placed at diagonal corners with an approximate angle of 60° to their right wall. Both nodes have a height of 1.2 meters. Fig. 7(b) shows a 3D model of the kitchen. The kitchen is equipped with a long rectangular table, a cabinet, a chair, a microwave oven, and a sink.

The number of participants is three, two men and one female, their age is from 25 to 46, and their height is from 162cm to 183cm. We investigated fifteen activities in the kitchen, such as (a) Drinking, (b) Eating, (c) Walking, (d) Sitting, (e) Washing, (f) Standing, (g) Open door and get out, (h) Open door and get in, (i) Close cabinet, (j) Open cabinet, (k) Close oven, (l) Close freezer, (m) Open freezer, (n) Open oven, and (o) No activity. Fig. 8 illustrates the spectrograms of eight investigated human activities that were observed by ‘Node 1’ and ‘Node 2’.

**B. Data processing**

The value of radar signals has two components $I + jQ$ as shown in Fig. 9(a). The radar signals were transformed from the time-amplitude domain to the time-frequency domain using STFT. The duration of different activities is different. People usually take a longer time to perform eating and drinking than other activities. So, the sliding window of the STFT should be chosen appropriately to contain the main motion cycle of each activity and also to ensure low recognition latency. With an interval of 0.5s, a sliding window of 2.5s is used for segmentation. We choose the sliding window of 2.5s because it is an average time to complete one motion cycle for these fifteen activities. The sampling frequency of each radar sensor is 250 Hz, so a sliding window contains 625 signals. By moving the sliding window continuously, a spectrogram can be extracted every 0.5 seconds. The signal processing can be shown in Fig. 9.

Fig. 7. The plan of the kitchen and its 3D model

Fig. 8. Spectrograms of eight investigated human activities

Fig. 9. Signal processing diagram

Fig. 10. Number of samples of each activity
In the experiment, 15350 spectrograms were obtained in total after radar signal processing. The sample size of each activity is shown in Fig 10.

VI. EVALUATION

This section evaluates the performance of our STnet in three ways: validation on the test dataset, a comparison with the other three networks, and a comparison to our previous work. We also analyzed the impact of radar sensor fusion.

A. Validation on the test dataset

We used 80% of the spectrogram samples for training and validation, and 20% for testing in the classification. The training and test datasets are mixed. The test dataset was not exposed during training and validation. Because each activity class has a different number of samples. For overcoming the problem of class-imbalanced data, we assigned different weights to each class in training. Weight ratios were inversely proportional to spectrogram class proportions. For the training, we used cross-entropy as the loss function; Adam as the optimizer with a learning rate of 0.0001; and 180 epochs. The achieved overall accuracy by STnet in testing is 99.64%.

Fig. 11 illustrates the normalized confusion matrix generated by the STnet in test. As it can be seen, the classification rates of all activity classes are above 98%, and most of them achieve the classification rate of 100%. Some activities have similarities that led to misclassification. For example, some samples of 'Drinking' are classified into 'Eating' and 'No activity' because 'Drinking' is similar to 'Eating', and there is a short period that people have no movement while he/she is drinking.

B. Ablation study

As STnet comprises the spectro-stream and temporal stream, there are two hyperparameters in STnet, which are the number of residual blocks $r$ in the spectro-stream and the number of temporal residual blocks $t$ in the temporal stream. In this section, we performed an ablation study by investigating these two hyperparameters $r$ and $t$. In this study, we tuned $r$ in the range of $\{1, 3, 5, 7, 9\}$ and $t$ in the range of $\{3, 5, 7, 9, 11\}$. STnet $(rm, tn)$ denotes the spectro-stream has $m$ blocks and the temporal stream has $n$ blocks. As shown in Table II, the accuracy of STnet increases when the number of residual blocks in the spectro-stream increases from 1 to 3, while it declines when the number of the blocks increases from 3 to 9. The number of blocks in the temporal stream has little effect on the performance of STnet, but still, STnet has slightly higher accuracy when the temporal stream has 7 blocks. Through the ablation experiments, it can be found that STnet $(r3, t7)$ achieved the best performance. In this paper, the STnet defaults as STnet $(r3, t7)$. 
Fig. 9. Signal processing. (a) original signal sequence $I + jQ$, (b) a spectrogram with a sliding window for segmentation, (c) sensor fusion by overlapping segments from two nodes.

Fig. 11. Normalized confusion matrix
TABLE II
THE PERFORMANCE OF STNet WITH DIFFERENT NUMBERS OF BLOCKS IN TWO STREAMS

<table>
<thead>
<tr>
<th>Model</th>
<th>OA</th>
<th>Recall</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>STnet (r3, t7)</td>
<td>99.64%</td>
<td>99.61%</td>
<td>99.64%</td>
</tr>
<tr>
<td>STnet (r1, t7)</td>
<td>98.93%</td>
<td>98.97%</td>
<td>98.96%</td>
</tr>
<tr>
<td>STnet (r5, t7)</td>
<td>98.05%</td>
<td>98.01%</td>
<td>98.15%</td>
</tr>
<tr>
<td>STnet (r7, t7)</td>
<td>64.37%</td>
<td>64.38%</td>
<td>66.72%</td>
</tr>
<tr>
<td>STnet (r9, t7)</td>
<td>36.51%</td>
<td>37.77%</td>
<td>37.77%</td>
</tr>
<tr>
<td>STnet (r3, t3)</td>
<td>99.62%</td>
<td>99.61%</td>
<td>99.62%</td>
</tr>
<tr>
<td>STnet (r3, t5)</td>
<td>99.62%</td>
<td>99.61%</td>
<td>99.60%</td>
</tr>
<tr>
<td>STnet (r3, t9)</td>
<td>99.57%</td>
<td>99.57%</td>
<td>99.57%</td>
</tr>
<tr>
<td>STnet (r3, t11)</td>
<td>99.12%</td>
<td>99.13%</td>
<td>99.13%</td>
</tr>
</tbody>
</table>

C. Comparison with other three networks

For validating the superiority of our STnet, we compared it with the other three networks: ResNet, TCN, and BiLSTM+ResNet from three aspects (accuracy, memory consumption, and computation efficiency). The ResNet and TCN are actually the same as the spectro-stream and temporal stream respectively in the STnet. As shown in Fig. 12, the BiLSTM+ResNet has a similar structure as the STnet. Differently, it uses two bidirectional LSTM layers to construct the temporal stream and each bidirectional LSTM layer has 128 units. Bidirectional LSTM is used to model sequences and to analyze the temporal patterns in the sequence forwards and backwards [80].

As shown in Table III, the STnet achieved the best performance in activity classification, which is 99.64% in overall accuracy (OA), 99.61% in Recall, and 99.64% in f1. It is about 1.3% higher than BiLSTM+ResNet, which presents that TCN has better capability in temporal pattern learning than LSTM. Both STnet and BiLSTM+ResNet are better than TCN and ResNet, which denotes spectro-temporal modeling for radar-based human activity can efficiently integrate micro-Doppler features and temporal features to improve the recognition rate of human activity. From the aspect of memory consumption, STnet consumes 10.7 MB, which is more than TCN, ResNet, and BiLSTM+ResNet. All of them use a small memory, which is not a problem for most edge devices (including Raspberry pi and Arduino). From the aspect of computation efficiency, we use the FLOPs as a metric. FLOPs denotes the total floating-point operations of a model. It has been frequently used to measure the inference time of a deep learning model. In Table III, we use MFLOPs that denotes the number of millions of floating-point operations in a model. As can be seen, STnet has 1.4 MFLOPs, which is more than TCN and ResNet but less than BiLSTM+ResNet. From the above, it can conclude that, compared to BiLSTM+ResNet, STnet has superiority in accuracy and inference time (computation complexity) but occupies more memory; TCN and ResNet have much less memory consumption and FLOPs, but their accuracies are also much lower than the STnet. Fig. 13 illustrates the lines of validation accuracy and validation loss obtained in validation using four networks. Obviously, the STnet proposed in this paper achieves the highest accuracy and lowest loss.

TABLE III
COMPARISON OF THE CLASSIFICATION MODELS

<table>
<thead>
<tr>
<th>Model</th>
<th>OA</th>
<th>Recall</th>
<th>F1</th>
<th>Memory</th>
<th>MFLOPs</th>
</tr>
</thead>
<tbody>
<tr>
<td>STnet</td>
<td>99.64%</td>
<td>99.61%</td>
<td>99.64%</td>
<td>10.7 MB</td>
<td>1.40</td>
</tr>
<tr>
<td>TCN</td>
<td>93.11%</td>
<td>92.97%</td>
<td>93.22%</td>
<td>5.86 MB</td>
<td>0.76</td>
</tr>
<tr>
<td>ResNet</td>
<td>97.95%</td>
<td>97.94%</td>
<td>97.95%</td>
<td>6.4 MB</td>
<td>0.83</td>
</tr>
<tr>
<td>BL-ResNet</td>
<td>98.36%</td>
<td>98.35%</td>
<td>98.37%</td>
<td>8.5 MB</td>
<td>2.40</td>
</tr>
</tbody>
</table>

D. The effect of sensor fusion

In previous experiments, we evaluated the models on the dataset that integrates two radar sensors. To show the impact...
of sensor fusion, we compared the classification results of STnet by using two radars and one radar in this section. The location of radar sensor 1 and radar sensor 2 is shown in Fig. 7(a). With one radar sensor, the input shape becomes (80, 80, 1), and the structure of STnet keeps the same. As shown in Table V, with the data of radar 1, the STnet achieves 95.68% OA; with the data of radar 2, the STnet achieves 94.65% OA. They are lower than 99.64% with around 4% and 5% respectively, which proves radar sensor fusion can improve human activity recognition. The observation perspective also affects the recognition rate as radar 1 and radar 2 provides different recognition rate and STnet+radar 1 offers 1% higher accuracy than STnet+radar 2. Fig. 14 illustrates the lines of validation accuracy and validation loss of STnet+radar 1 and STnet+radar 2 obtained in the validation.

![Validation accuracy and loss of STnet with radar 1 and radar 2 respectively](image)

**Fig. 14.** Validation accuracy and loss of STnet with radar 1 and radar 2 respectively

### E. Comparison with our previous work

In our previous work [23], we built a CNN with 3 convolutional layers and 2 max-pooling layers to perform activity classification on the same dataset collected by the radar sensor network. As a result of the CNN, the OA was 92.81%, the Recall was 93.14 %, and the F1 was 93.83%. It occupies 6.6 MB RAM and has 0.6 MFLOPs. Obviously, the STnet proposed in this work outperforms the CNN by a wide margin. However, it is worth noting that the STnet is bigger than the CNN in [23] and much more computation-intensive to be deployed on edge computers.

### VII. Discussion

As our data is strongly related to the things (door, oven, etc.) in the kitchen, and these things are distributed in different locations, which results in that our activity classification is probably dependent on the location. Because the radar signals differ from different locations due to the multipath effect. This problem has been long discussed. In [81], [82], the authors used transfer learning and meta-learning to adapt the model across different locations with a small number of data samples. However, these methods cannot effectively address the problem of location dependency and the model still needs some samples from the new locations. The most practical way to alleviate location dependency is to collect a large amount of data from different locations and orientations, which is laborious and time-consuming. In our work, we used two radar sensors to observe human activity from two different viewpoints, which improve the capability of human activity recognition. With multiple radar sensors, we think it is also possible to collect the samples by using the radars at different locations and viewpoints rather than letting people perform activities in different locations. The total samples of each radar sensor have different distances between the radar and the activity performer, which might be helpful to alleviate the location dependency of the model, but it needs further investigation which certainly is our future research plan.

### VIII. Conclusion

In this paper, we propose a two-stream network, STnet, which incorporates a spectro-stream and a temporal stream in order to extract micro-Doppler signatures and temporal patterns for radar-based human activity recognition. A radar sensor network with two radar sensor nodes was constructed in order to detect human activity without requiring users to carry any sensors. An investigation of 15 kitchen activities was conducted. An STFT was used to generate spectrograms, and a sliding window with a size of 2.5s and a step of 0.5s was applied to segment and generate samples in signal processing. We achieved 99.64% accuracy in testing with STnet, which is superior to BiLSTM+ResNet, ResNet, and TCN. By combining micro-Doppler features and temporal patterns, we demonstrate the great potential of spectro-temporal modeling in radar-based activity recognition. It also presents that the low-power low-cost radar sensor network is suitable to be deployed indoors for human activity recognition, which can be helpful in intrusion detection, healthcare, and smart home design. Our approach is power-saving and privacy persevering, and it does not require wearable sensors. The STnet can be used as a generic solution to other sensor-based activity recognition and sequence modelling. We will continue to implement our approach to other scenarios and sensors and investigate more activities in the future. We will also investigate model compression techniques that miniaturize activity recognition models while keeping a high recognition rate, which enables the models to be deployed on edge computers.

### References


