WiFi-based sensing for human activity recognition (HAR) has recently become a hot topic as it brings great benefits when compared with video-based HAR, such as eliminating the demands of line-of-sight (LOS) and preserving privacy. Making the WiFi signals to ‘see’ the action, however, is quite coarse and thus still in its infancy. An end-to-end spatiotemporal WiFi signal neural network (STWNN) is proposed to enable WiFi-only sensing in both line-of-sight and non-line-of-sight scenarios. Especially, the 3D convolution module is able to explore the spatiotemporal continuity of WiFi signals, and the feature self-attention module can explicitly maintain dominant features. In addition, a novel 3D representation for WiFi signals is designed to preserve multi-scale spatiotemporal information. Furthermore, a small wireless-vision dataset (WV AR) is synchronously collected to extend the potential of STWNN to ‘see’ through occlusions. Quantitative and qualitative results on WV AR and the other three public benchmark datasets demonstrate the effectiveness of our approach on both accuracy and shift consistency.

Index Terms— WiFi sensing, human action recognition, 3D spatiotemporal, wireless-vision

1. INTRODUCTION

Recent years have witnessed increasing research interest in human action recognition (HAR) as it expands sensing areas and provides vast potential applications [1, 2] in various sensing scenarios, such as assisted living [3], health monitoring [4], surveillance [5], etc. Many new action sensing technologies [1,2] are continuously emerging, which enlarges signal acquisition range and enriches measurement data [2]. These sensing techniques motivate the breakthrough of long-time monitoring in the non-intrusive pattern [6].

Video-based systems demand the coverage area within line-of-sight (LOS) [1]. In general, lighting variation affects the quality of images and thus, analyzed information [7]. In the same way, perspective change, especially using a single view acquisition device, provides a limited visualization of activities being analyzed. This induces occlusion with its different types: self-occlusion where body parts occlude each other, occlusion of another object, and partial occlusion [7]. From a user perspective, the presence of cameras can affect privacy and cannot be employed in many environments. Therefore, a passive monitoring system based on RF sensing is a more sound way to sidestep such drawbacks [2]. Unlike video-based solutions, RF especially WiFi channel state information (CSI) based sensing is insensitive to lighting conditions and not intrusive without the privacy issue [8]. WiFi-based solutions have no requirements of LOS thereby enabling larger detection areas. Existing systems, however, are quite coarse [9]. Past systems focus primarily on manually designed features, dependent on prior knowledge and incapable of adequately mining spatiotemporal information in CSI streams [10]. Furthermore, separate stages for feature extraction and classifier learning may reduce the accuracy of recognition results. Therefore, it is worth exploring how to non-manually obtain spatiotemporal features and jointly optimize feature learning as well as the classification process.

To address the problems, we propose an end-to-end spatiotemporal WiFi-based neural network (STWNN) to exploit the spatiotemporal characteristics of CSI signals simultaneously. To summarize, our contributions are 1) We propose a novel method for representing WiFi signals in a multi-scale 3D spatiotemporal form; 2) We design a 3D convolution module and attention module to exploit the inherent spatial, temporal, and frequency features and embedded in a residual manner to reduce training burden; 3) We collect synchronous video and WiFi datasets (WV AR) to enable STWNN to ‘see’ through the occlusions; 4) We conduct experiments on three public benchmark datasets. The results show that our method outperforms competitive baselines with a good margin on the classification accuracy.
2. THE PROPOSED SPATIOTEMPORAL NEURAL NETWORK

2.1. Generation stage

2.1.1. Channel state information

The WiFi-based sensing principle is leveraging the influence of perceptual targets on the transmitted signal for recognition [11]. Generally, a WiFi system can be modeled as follows:

\[ B_s(i) = H_s(i)A_s(i) + \theta, \]

where \( s \in [1, \ldots, N_s] \) represents the index of the orthogonal frequency-division multiplexing (OFDM) subcarriers employed in the WiFi device, \( N_s \) is the total number of the OFDM subcarriers, \( i \) represents the index of the transmitted and received packets, \( A_s(i) \) and \( B_s(i) \) are the \( i \)-th transmitted and received packets associated with the OFDM subcarrier frequency \( s \), respectively, \( \theta \) represents the received noise, and \( H_s \) is a complex-valued matrix of dimensions \( N_T \times N_R \) that comprises the CSI measurements for the OFDM subcarrier frequency \( s \). \( N_T \) and \( N_R \) represent the number of transmitting and receiving antennas, respectively.

2.1.2. Multi-scale 3D CSI data generation

2.2. Feature learning stage

2.2.1. 3D convolution module

Convolutional neural networks with 3D kernels can directly extract spatiotemporal features from videos, however, suffer from the heavy training burden due to a large number of their parameters. To mitigate the issue, we construct the network based on ResNet, which introduces shortcut connections that bypass a signal from one layer to the next. The connections pass through the gradient flows of networks from later layers to early layers, and ease the training of very deep networks. The connections bypass a signal from the top of the block to the tail. Our 3D convolution module consists of multiple residual blocks seen in Fig. 1.

2.2.2. Feature self-attention module

Inspired by the attention mechanism, our works formulate attention drift as a sequential process to capture different attended aspects. The learned sequential features by the 3D convolution module will be employed as the inputs of the attention model as self-attention with no prior information, as seen in Fig. 1. Given \( n \) feature vectors \( \alpha_i, i = 1, 2, \ldots, n \) derived from the 3D convolution module, a score function \( \Phi(\cdot) \) such as \( \text{tanh} \), \( \text{relu} \) and \( \text{linear} \) evaluates the importance of each feature vector by calculating a score \( \beta_i \) as follows:

\[ \beta_i = \Phi(\chi^T\alpha_i + b), \]

where \( \chi \) and \( b \) are weight vector and bias respectively. After obtaining the score for each feature vector, we can normalize it using the softmax function. The final Mask of the attention model is the multiplication of the feature vectors and their normalized scores, which is shown as follows:

\[ \text{Mask} = \sum_{i=1}^{n} (\text{softmax}(\beta_i) \cdot \alpha_i) = \sum_{i=1}^{n} \left( \frac{\exp(\beta_i)}{\sum_{j=1}^{n} \exp(\beta_j)} \cdot \alpha_i \right). \]

2.3. Task stage

The task stage is to leverage multi-scale spatiotemporal features learned above to compute the outputs for a specific task. Cross-entropy loss is a basic option to measure the difference between the network outputs \( O \) and the ground-truth values as follows.

\[ \mathcal{L} = -\lambda \sum_{j=1}^{J} G^j \log(M \cdot \text{Mask}^j \cdot O^j) - (1 - \lambda) \sum_{j=1}^{J} G^j \log(O^j) \]

where \( \cdot \) is the convolution operation, \( J \) is the snippet number of training samples, and \( \lambda \) is the weight coefficient. A typical value is \( \lambda = 0.5 \) in our experiments. In addition, we utilize the Stochastic Gradient Descent with Momentum to train the parameters.
3. EXPERIMENTS

Our WV AR dataset. WV AR collection was conducted in one spacious office apartment as shown in Fig. 3. Two volunteers were asked to implement nice activities and repeat for five trials in different motion details such as varied directions to ensure the diversity of the actions. The experimental hardware consists of two desktop computers as transmitter and receiver operating in IEEE 802.11n monitor mode at 5.4 GHz with a sampling rate of 100 Hz. The subcarriers \( N_S \) are equal to 30 and 3 antennas both in transmitter \( N_T \) and receiver \( N_R \) are activated. We employed the CSI extraction tool\(^1\) for CSI signals recording and CSI packets extraction. To synchronize the images and wireless data, a deep camera D435i\(^2\) was attached to our receiver desktop at the same location as the wireless card. We recorded the video at 20 FPS, i.e. every five CSI samples corresponding to one frame in the video.

HHI, CSLOS and WAR dataset. The dataset HHI [10] comprises 12 different interactions performed by 40 distinct pairs of subjects while performing different human-to-human interactions (HHI) inside an office with 10 different trials. Another cross-scene dataset (CSLOS) [12] is provided by the same group as the HHI. CSLOS is comprised of five experiments performed by 30 different subjects in two LOS environments. Each subject performed 20 trials for each of the experiments with different variations of human movements. The dataset WAR [13] consists of 6 persons for 6 activities with 20 trials for each in an indoor office. The sampling rate is 1 kHz.

Baselines. We design a 2D baseline with the same neural network structure as STWNN (2DWNN). Besides, the classic SVM [14] is deployed for comparison.

Quantitative Results. Table 1 shows the classification accuracy of the WV AR dataset. We tested the data from all scenarios (All), the scenes with partial (S-p) and full occlusion (S-f), respectively. For all scenarios, SVM (non_gen) without the generation stage performs worse than SVM, indicating the generation stage’s effectiveness. The overall accuracy (OA) of STWNN is higher than these of others and over 85% on all the actions. It is possible that all activities have obvious trajectories in the spatial domain over time. STWNN can pay attention to the characteristics of both spatiotemporal domains. For the scene with partial and full occlusions, STWNN surpasses the other two methods with a margin of around 6%. These results indicate that robust to the influence of the environmental disturbance.

### Table 1: Classification accuracy on the dataset WV AR.

| Scene   | Method   | fall (skeleton) | talk (skeleton) | seat (skeleton) | drink (skeleton) | OA   |
|---------|----------|----------------|----------------|----------------|-----------------|------|
| All     | SVM      | 0.85           | 0.80           | 0.80           | 0.80            | 0.73 | 0.50 | 0.04 | 0.79 |
|         | STWNN    | 0.88           | 0.80           | 0.80           | 0.80            | 0.73 | 0.50 | 0.04 | 0.79 |
|         | 2DWNN    | 0.90           | 0.92           | 0.91           | 0.93            | 0.73 | 0.50 | 0.04 | 0.79 |
|         | SVM      | 0.93           | 0.96           | 0.95           | 0.97            | 0.73 | 0.50 | 0.04 | 0.79 |
|         | STWNN    | 0.94           | 0.96           | 0.96           | 0.97            | 0.73 | 0.50 | 0.04 | 0.79 |
|         | 2DWNN    | 0.95           | 0.96           | 0.96           | 0.97            | 0.73 | 0.50 | 0.04 | 0.79 |
|         | SVM      | 0.97           | 0.98           | 0.97           | 0.97            | 0.73 | 0.50 | 0.04 | 0.79 |
|         | STWNN    | 0.98           | 0.98           | 0.98           | 0.97            | 0.73 | 0.50 | 0.04 | 0.79 |
|         | 2DWNN    | 0.99           | 0.99           | 0.98           | 0.97            | 0.73 | 0.50 | 0.04 | 0.79 |

1It is available at https://github.com/dhalperi/linux-80211n-csitool-supplementary

### Table 2: Classification accuracy on the dataset WAR.

| Methods  | lie_down | fall | run | sit_down | stand_up | walk | OA   |
|----------|----------|------|-----|----------|----------|------|------|
| RF [15]  | 0.53     | 0.60 | 0.81| 0.88     | 0.49     | 0.57 | 0.65 |
| LSTM [13]| 0.52     | 0.72 | 0.92| 0.96     | 0.76     | 0.52 | 0.73 |
| SVM      | 0.91     | 0.96 | 0.93| 0.96     | 0.71     | 0.87 | 0.93 |
| 2DWNN    | 0.93     | 0.93 | 0.93| 0.98     | 0.90     | 0.86 | 0.95 |
| STWNN    | 0.96     | 0.99 | 0.97| 0.97     | 0.96     | 0.93 | 0.97 |

Table 1 also reports the skeleton-based classification based on the WiFi and video data on the WV AR dataset. Inspired by the work [6, 21], the skeletons derived from Alphapose [22] are used to train the STWNN in LOS scenes.
Table 3. Classification accuracy on the dataset HHI.

| Methods      | approaching | departing | hand, shaking | high five | kicking, left_leg | kicking, right_leg | pointing, left_hand | pointing, right_hand | punching, left_hand | punching, right_hand | pushing | OA   |
|--------------|-------------|-----------|---------------|-----------|------------------|-------------------|--------------------|--------------------|--------------------|--------------------|---------|------|
| GoogleNet [17] | 0.93        | 0.93      | 0.79          | 0.76      | 0.64             | 0.64              | 0.54               | 0.50               | 0.78               | 0.77               | 0.59    | 0.89 |
| ResNet-18 [18] | 0.92        | 0.90      | 0.85          | 0.79      | 0.77             | 0.68              | 0.60               | 0.82               | 0.80               | 0.60               | 0.65    | 0.76 |
| SqueezeNet [19] | 0.95        | 0.93      | 0.83          | 0.76      | 0.70             | 0.66              | 0.62               | 0.78               | 0.79               | 0.60               | 0.67    | 0.76 |
| E2EDLF [20] | 0.96        | 0.92      | 0.89          | 0.84      | 0.86             | 0.78              | 0.82               | 0.85               | 0.90               | 0.73               | 0.80    | 0.85 |
| SVM          | 0.99        | 0.96      | 0.90          | 0.83      | 0.82             | 0.73              | 0.79               | 0.69               | 0.62               | 0.74               | 0.77    | 0.74 |
| 2DWNN        | 0.93        | 0.89      | 0.93          | 0.88      | 0.67             | 0.99              | 0.99               | 0.89               | 0.94               | 0.99               | 0.99    | 0.88 |
| STWNN        | 0.99        | 0.99      | 0.93          | 0.96      | 0.85             | 0.84              | 0.83               | 0.89               | 0.92               | 0.76               | 0.75    | 0.87 |

Table 4. Classification accuracy on the dataset CSLOS

| Scenes | Methods      | avg_move | falling | walking | sitting | standing | turning | picking | OA   |
|--------|--------------|----------|---------|---------|---------|----------|---------|---------|------|
| E1     | SVM [23]     | 0.95     | 0.86    | 1.00    | 0.91    | 0.90     | 0.92    | 0.94    |      |
|        | 2DWNN        | 0.89     | 0.60    | 0.73    | 0.86    | 0.67     | 0.94    | 0.91    |      |
|        | STWNN        | 0.86     | 0.96    | 0.93    | 0.99    | 0.92     | 0.99    | 0.96    |      |
| E2     | SVM [23]     | 0.95     | 0.82    | 0.99    | 0.82    | 0.81     | 0.82    | 0.89    |      |
|        | 2DWNN        | 0.84     | 0.78    | 0.75    | 0.83    | 0.69     | 0.84    | 0.79    |      |
|        | STWNN        | 0.95     | 0.94    | 0.94    | 0.95    | 0.90     | 0.88    | 0.92    |      |

Qualitative Results In this section, we show the effectiveness of CSI data on WVAR. Besides, we demonstrate the spatiotemporal scheme and the attention module of STWNN are meaningful at the feature level.

Skeleton visualization is further to show the effectiveness of WVAR as mentioned in section 4.1. As seen in Fig. 4(a)-(h), our STWNN yields robust skeletons close to Alphapose. Particularly, these partially covered actions such as kick are also well-estimated. This demonstrates that our CSI data on WVAR has a good efficiency in these two scenarios.

In Fig. 5, we show gSOM [24] projections of the features of the WAR dataset by 2DWNN and 3DWNN before and after using the attention module, respectively. Features extracted from the same action tend to be near each other, and vice versa. Compared Fig. 5(a) with Fig. 5(c), we can find the features from the same action of 3DWNN are more compact than that of 2DWNN in terms of lie-down, walk, and stand-up. It proves that 3DWNN has better potential than 2DWNN in exploring the effective spatiotemporal features. Perceptually, comparing Fig. 5(c) with (d), features from the same category after using the attention module are more clustered than before, such as run and pick-up. It further indicates that the attention module improves the efficiency of the features to a certain extent.

In this paper, an end-to-end spatiotemporal WiFi-based neural network STWNN was proposed to enhance the performance of privacy-preserving WiFi-based HAR. Its strength lays in the ability for the effective exploitation of the multi-scale spatiotemporal features and explicit maintenance of self-attention features. Moreover, we collected synchronous video and WiFi datasets WVAR to enable STWNN to see the skeleton in complex visual conditions like partial and full occlusions scenarios. In addition, we have compared the results of our proposed STWNN with the results of SVM, 2DWNN, and state-of-the-art competitors. The experiments on four benchmark datasets WVAR, WAR, HHI, and CSLOS showed that STWNN compares favorably against competitive baselines.
5. REFERENCES

[1] Mariko Isogawa, Ye Yuan, Matthew O’Toole, and Kris M Kitani, “Optical non-line-of-sight physics-based 3D human pose estimation,” in CVPR, 2020, pp. 7013–7022.

[2] Yiyue Luo, Yunzhu Li, Michael Foshey, Wan Shou, Pratyusha Sharma, Tomas Palacios, Antonio Torralba, and Wojciech Matusik, “Intelligent Carpet: Inferring 3D human pose from tactile signals,” in CVPR, 2021, pp. 11255–11265.

[3] Bang Wu, Zixiang Ma, Stefan Poslad, and Yidong Li, “WiFi fingerprint based, indoor, location-driven activities of daily living recognition,” in BESC. IEEE, 2018, pp. 148–151.

[4] Bo Tan, Qingchao Chen, Kevin Chetty, Karl Woodbridge, Wenda Li, and Robert Piechocki, “Exploiting WiFi channel state information for residential healthcare informatics,” IEEE Commun. Mag., vol. 56, no. 5, pp. 130–137, 2018.

[5] Fei Wang, Jinsong Han, Shiyuan Zhang, Xu He, and Dong Huang, “CSI-net: Unified human body characterization and action recognition,” arXiv:1810.03064, 2018.

[6] Tianhong Li, Lijie Fan, Mingmin Zhao, Yingcheng Liu, and Dina Katabi, “Making the invisible visible: Action recognition through walls and occlusions,” in ICCV, 2019, pp. 872–881.

[7] Djamila Romaissa Beddiar, Brahim Nini, Mohammad Sabokrou, and Abdenour Hadid, “Vision-based human activity recognition: a survey,” Multimed. Tools. Appl., vol. 79, no. 41, pp. 30509–30555, 2020.

[8] Chenning Li, Zhichao Cao, and Yunhao Liu, “Deep AI enabled ubiquitous wireless sensing: A survey,” CSUR, vol. 54, no. 2, pp. 1–35, 2021.

[9] Shiwei Fang, Sirajum Munir, and Shahrirar Nirjon, “Person tracking and identification using cameras and Wi-Fi channel state information (CSI) from smartphones: dataset,” in SenSys, 2020, pp. 26–30.

[10] Rami Alazraei, Ali Awad, Alsaify Baha’a, Mohammad Hababeh, and Mohammad I Daoud, “A dataset for Wi-Fi-based human-to-human interaction recognition,” Data Brief, vol. 31, pp. 105668, 2020.

[11] Hao Wang, Daqing Zhang, Yasha Wang, Junyi Ma, Yuxiang Wang, and Shengjie Li, “RT-Fall: A real-time and contactless fall detection system with commodity WiFi devices,” IEEE Trans. Mobile Comput., vol. 16, no. 2, pp. 511–526, 2016.

[12] Alsaify Baha’a, Mahmoud M Almazari, Rami Alazraei, and Mohammad I Daoud, “A dataset for Wi-Fi-based human activity recognition in line-of-sight and non-line-of-sight indoor environments,” Data in Brief, vol. 33, pp. 106534, 2020.

[13] Siamak Yousefi, Hirokazu Narui, Sankalp Dayal, Stefano Ermon, and Shahrokh Valaee, “A survey on behavioral recognition using WiFi channel state information,” IEEE Commun. Mag., vol. 55, no. 10, pp. 98–104, 2017.

[14] Koby Crammer and Yoram Singer, “On the algorithmic implementation of multiclass kernel-based vector machines,” J. Mach. Learn Res., vol. 2, no. Dec, pp. 265–292, 2001.

[15] Tin Kam Ho, “Random decision forests,” in ICDAR. IEEE, 1995, vol. 1, pp. 278–282.

[16] Sean R Eddy, “What is a hidden Markov model?,” Nat. Biotechnol., vol. 22, no. 10, pp. 1315–1316, 2004.

[17] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich, “Going deeper with convolutions,” in CVPR, 2015, pp. 1–9.

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, “Deep residual learning for image recognition,” in CVPR, 2016, pp. 770–778.

[19] Forrest N Iandola, Song Han, Matthew W Moskewicz, Khalid Ashraf, William J Dally, and Kurt Keutzer, “SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and¡ 0.5 mb model size,” arXiv:1602.07360, 2016.

[20] Rami Alazraei, Mohammad Hababeh, Alsaify Baha’a, Mostafa Z Ali, and Mohammad I Daoud, “An end-to-end deep learning framework for recognizing human-to-human interactions using Wi-Fi signals,” IEEE Access, vol. 8, pp. 197695–197710, 2020.

[21] Mingmin Zhao, Tianhong Li, Mohammad Abu Alsheik, Yonglong Tian, Hang Zhao, Antonio Torralba, and Dina Katabi, “Through-wall human pose estimation using radio signals,” in CVPR, 2018.

[22] Rami Alazraei, Mohammad Hababeh, Alsaify Baha’a, Mostafa Z Ali, and Mohammad I Daoud, “An end-to-end deep learning framework for recognizing human-to-human interactions using Wi-Fi signals,” IEEE Access, vol. 8, pp. 197695–197710, 2020.

[23] Alsaify Baha’a, Mahmoud M Almazari, Rami Alazraei, and Mohammad I Daoud, “Exploiting Wi-Fi signals for human activity recognition,” in ICICS. IEEE, 2021, pp. 245–250.

[24] Nejc Ilc and Andrej Dobnikar, “Generation of a clustering ensemble based on a gravitational self-organising map,” Neurocomputing, vol. 96, pp. 47–56, 2012.