Pattern Recognition Letters 171 (2023) 21–27

Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

# Seeing the unseen: Wifi-based 2D human pose estimation via an evolving attentive spatial-Frequency network

Yi-Chung Chen[a,1], Zhi-Kai Huang[a,1], Lu Pang[a], Jian-Yu Jiang-Lin[a], Chia-Han Kuo[a], Hong-Han Shuai[a,*], Wen-Huang Cheng[b]

[a] National Yang Ming Chiao Tung University, No.1001, University Road, Hsinchu 300093, Taiwan (R.O.C.)
[b] National Taiwan University, No. 1, Sec. 4, Roosevelt Rd., Taipei 106216, Taiwan (R.O.C.)

ARTICLE INFO

*Article history:*
Received 20 May 2022
Revised 2 January 2023
Accepted 29 April 2023
Available online 1 May 2023

Edited by: Prof. S. Sarkar

*Keywords:*
CSI
Wifi
2D Pose estimation
Deep learning
Spatial-Frequency network

ABSTRACT

Camera-based human pose estimation has become popular due to its wide applications and easy implementation. However, it is not applicable under several circumstances such as poor illumination, occlusion, and private protection. In this work, we utilize Wifi signals to estimate 2D human poses, which is challenging because Wifi signals are abstract and contain limited information. To address these challenges, we develop an evolving attentive spatial-frequency network to discover the relationship between signal variation and body movement for Wifi-based 2D human pose estimation. By first taking dilated CSI sequences as inputs, a spatial-frequency encoder is then introduced to effectively integrate static spatial information and dynamic frequency information from CSI signals. Finally, we design an evolving attention module to enable our model to attend to certain channels of features. Due to a lack of benchmarks, we propose two Wifi-based human pose estimation datasets, the General Pose Estimation dataset (GPE) and Specific Pose Estimation dataset (SPE), which have been released as a public download at project page. Extensive experiments on the proposed datasets show that our model outperforms the state-of-the-art method by at least 16% in terms of *PCK*@20 (percentage of correct keypoints).



## 1. Introduction

Multi-person 2D pose estimation [1–3] targets and locates anatomical keypoints or "parts" of all people in RGB images. It is a fundamental problem of widespread applications such as crowd estimation [4,5], visual tracking [6], pose tracking [7], hand tracking [8] and virtual try-on [9,10]. However, camera-based 2D human pose estimation methods are not applicable in some application scenarios, such as with poor illumination, and most importantly, camera-based approaches are prohibited in high-privacy environments, such as toilets and changing rooms.

In this paper, we focus on estimating 2D human pose by using Wifi signals. Compared to camera-based approaches, Wifi-based pose estimation approaches are advantageous in poorly-lit and privacy-protected environments. In addition, these approaches are low-cost and convenient for collecting data due to the ubiquity of Wifi devices. Although a few works and surveys have been proposed to estimate poses from Wifi [11–13], they simply apply an off-the-shelf deep neural network backbone, such as ResNet [14] and Unet [15], which are originally designed for computer vision tasks. Without considering the characteristics of Wifi, these backbone models may lead to an inferior performance on Wifi feature extraction.

In fact, the accuracy of the 2D human pose estimation from Wifi signals is limited due to the following two challenges. The first involves **implicit relationships.** Wifi signals are abstract and contain limited information, so how to transform Wifi signals into poses is currently unknown. One possible approach is to use end-to-end training with well-designed deep learning architectures. However, the backbone architectures adopted by previous works [11,12,16] are devised for computer vision tasks, and may not be suitable for this type of transformation. For instance, the rescaling mechanism used in Unet is not suitable for this transformation since Wifi signals do not contain complex spatial relations as video frames do. As pose-related information lies in complicated patterns between subcarriers and transceiver pairs, it is challenging to capture the patterns for reconstructing the

* Corresponding author.
*E-mail addresses:* ck315.ee06@nctu.edu.tw (Y.-C. Chen), r10921059@ntu.edu.tw (Z.-K. Huang), panglu@nctu.edu.tw (L. Pang), epj.ee07@nctu.edu.tw (J.-Y. Jiang-Lin), edwinkuo.ee07@nctu.edu.tw (C.-H. Kuo), hhshuai@nycu.edu.tw (H.-H. Shuai), wenhuang@csie.ntu.edu.tw (W.-H. Cheng).
[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.patrec.2023.04.020
0167-8655/

2D pose image using previous methods. The second challenge involves **environment-sensitive.** According to the prevalent standard of IEEE 802.11n Wifi signals, received signals are composed of 30 subcarriers with orthogonal frequency, and the amount of information hidden in each subcarrier varies with the movements of objects in the environments. Hence, an effective selecting mechanism is required to adaptively leverage the information from each subcarrier. However, this issue has not been well-addressed in previous works. For instance, [17] only selects the top 20 subcarriers with the highest scores, which are more noise resilient and are derived from the covariance between each subcarrier pair[18]. uses the K-means algorithm to divide 30 subcarriers into 3 clusters according to the sensitivity to human activities, and then selects the most obvious subcarriers reflected by the activities. These methods are handcrafted and not robust for all scenarios. Other works like [11] and [12] simply neglect this issue, which may lead to inferior performance.

To address the above challenges, we propose an **Evolving Attentive Spatial-Frequency Network (EASFN)** for Wifi-based 2D human pose estimation. Specifically, our model takes Channel State Information (CSI) as the input because CSI is sensitive to the movements of humans [19]. To collect RGB-CSI data pairs, we construct a hardware system with a camera and a set of receiver antennas and transmitter antennas (three for each). Each data pair contains an RGB frame recorded at 20 fps along with frame-level CSI data including 25 CSI samples sampled at 500Hz. The RGB frames are used for generating ground truth heatmaps through OpenPose [20], and the CSI signals are used as training inputs.

To deal with the first challenge, a spatial-frequency encoder is proposed to simultaneously reconstruct 2D scene information and extract frequency features. Specifically, as the signal fluctuation is caused by the movements of objects and the frequency is related to the speed of object movements [19], the proposed encoder treats spatial and frequency separately by employing two sets of 3D convolution kernels. One set of 3D convolution kernels learns spatial features from CSI signals, which are based on the environment. Another set of 3D convolution kernels are leveraged to learn multi-scale temporal features of different frequencies. It is worth noting that the proposed spatial-frequency encoder can leverage CSI information more effectively than off-the-shelf backbone networks, such as ResNet [14], used in previous works [11,12].

To address the second challenge, we introduce an evolving attention module to enhance specific spatial and frequency channels across the time. Finally, the enhanced features are fed into a feature decoder to predict human poses. In order to adapt to multi-person pose estimation scenarios, we generate both joint heatmaps (JHMs) and part affinity fields (PAFs) to leverage the joint association for estimating human poses. The experimental results reveal that EASFN outperforms the state-of-the-art method by at least 16% in terms of *PCK*@20.

Our contributions are summarized as follows:

- We propose a novel Wifi feature encoder for pose estimation—spatial-frequency encoder—instead of using conventional CNN backbones. The proposed spatial-frequency encoder is general and can be applied to multiple Wifi-based applications.
- An evolving attentive spatial-frequency network is developed to solve the environment-sensitive issue of Wifi-based 2D human pose estimation, including two crucial components: the spatial-frequency encoder and evolving attention module.
- We contribute the General Pose Estimation dataset (GPE) and Specific Pose Estimation dataset (SPE) for Wifi-based 2D multi-person pose estimation, which contains 344,886 data pairs in total (1 image frame and 1 frame-level CSI data).

## 2. Related works

### 2.1. RGB-Based 2D person pose estimation.

Camera-based 2D person pose estimation aims at detecting keypoints or body segmentation of all persons from an RGB image/video. Most approaches can be categorized into two classes: bottom-up and top-down approaches. The bottom-up approaches [21,22] first detect human keypoints of all persons, and then associate them with each person. On the other hand, the top-down approaches [23,24] use detection algorithms [25] to detect persons first, e.g., using Mask R-CNN [26], and then estimate single-person keypoints or parts. Although top-down approaches usually perform better than bottom-up approaches, they are restricted to the performance of human detectors. Compared to camera-based approaches, the proposed Wifi-based approach can be used for different scenarios, e.g., low-light conditions, privacy-sensitive applications, and through-wall pose estimation.

### 2.2. Non-RGB based pose estimation.

Research works using Wifi signals as probes have received a great amount of attention these years due to the low-cost settings and high availability of Wifi. As a result, many Wifi-based activity interpretation tasks have been explored such as indoor positioning [27,28], action recognition [29], person identification [30] and human pose estimation [11,12,16,31]. Research shows that Wifi signals can be used as a robust data source to reconstruct 2D human poses. Person-in-Wifi (PIW) [12], for example, adopts end-to-end learning by using CNN to map CSI signals to a Segmentation Mask (SM), Joint Heatmaps (JHMs) and Part Affinity Fields (PAFs), which are further used to predict multi-person poses via joint association. Instead of keypoints, Wi-Pose [16] utilizes a skeleton mask as supervision for better robustness. Differing from previous work, our work focuses on finding a better architecture for cross-modal transformation. It is worth noting that a recent work 1) studies 3D human pose estimation from Wifi signals [31], which uses a system composed of 21 infrared cameras to obtain ground truth as supervision, and 2) addresses the cross-domain issue by taking the 3D velocity profile as the input. However, to obtain a 3D velocity profile, subjects are required to perform activities within a small space surrounded by 3 groups of antennas of different heights. Such a complicated device setting limits the applicability of the method. [32] fuses the amplitude and phase of CSI into images, and designs a network to extract features associated with poses from CSI images. Though it appears to achieve satisfying results, its experimental setup, which contains only data from 5 subjects walking in a single environment, is not sufficient to comprehensively evaluate the model performance. Moreover, [31,32] can only achieve single-person pose estimation while our proposed method can be implemented in a multi-person scenario.

In addition to Wifi, other non-RGB based pose estimations have also drawn a lot of attention[33]. is the first work using RF signals to capture through-wall human figures. It utilizes a Frequency Modulated Carrier Wave(FMCW) system that transmits periodic RF signals whose frequency lies between 5.46GHz and 7.24GHz to measure the depth of signal reflection from humans. The consecutive reflection snapshots are then aggregated together to obtain a complete human figure. Based on the success of [33,34] leverages the powerful convolutional neural network to further implement fine-grained human pose estimation. It takes RGB images as the supervision of cross-modal learning and achieves results of high accuracy. Moreover, [35] introduces RFID-Pose, a system for tracking multi-people 3D poses from RFID data with an average error of less than 5 cm, In the proposed system, RFID tags are attached to the target human joints. The movements of the tags are captured

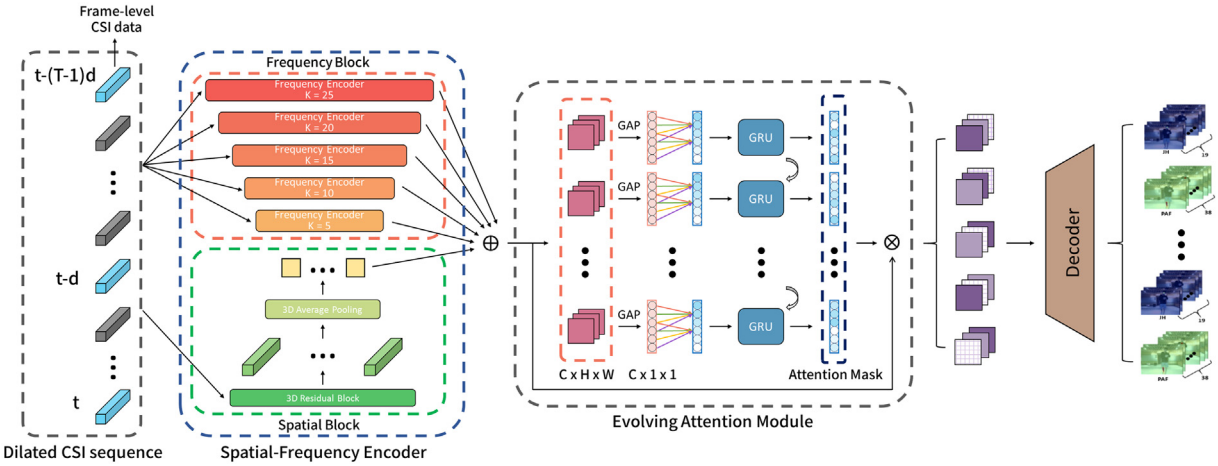


**Fig. 1. The architecture of our model.** We select a dilated CSI sequence as input which is labeled in blue in the figure. After extracting features via a spatial-frequency encoder, we use an evolving attention module to learn attention for each channel of spatial and frequency features. The attentive features are fed into a feature decoder, which is used to predict joint heatmaps (JHMs) and part affinity fields (PAFs). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

by the phase variations in the responses from each tag. The system address the need for complex hardware setting such as the FMCW device and the antenna arrays in RF-based methods. However, it requires user wearing RFID tags, which make it inconvenient to apply in the daily scenario.

## 3. Evolving attentive spatial-Frequency network

### 3.1. Overview

CSI signals [36] record the signal variation between the transmitter and receiver antennas. When users are located between the transmitters and receivers, the received signals are highly affected by human poses. Moreover, the frequency variation in CSI signals reveals human activities and human movement speed [19]. Nevertheless, it is difficult to estimate 2D human poses from Wifi signals due to the limited information. To address this issue, we introduce an evolving attentive spatial-frequency network, which effectively aggregates dynamic frequency information and spatial semantic information. Fig. 1 shows the proposed architecture. Specifically, given a dilated CSI sequence, we put forward a spatial-frequency network to simultaneously extract the spatial semantic information of the scene as well as frequency information related to the movement. Moreover, to better estimate the poses of moving humans, we design an evolving attention module to attend to specific channels of spatial and frequency features across time. In the following, we present the details of these modules.

### 3.2. Wifi signals and CSI

According to the prevalent IEEE 802.11 n/g/ac protocols, the information transmitted through Wifi is carried by 30 subcarriers, which is known as orthogonal frequency division multiplexing (OFDM). The frequency of the subcarriers is located in the band centering at 2.4/5GHz with bandwidth 20/40/80/160MHz, and a multiple-input-multiple-output (MIMO) is also supported under the IEEE 802.11 n/g/ac protocols, which allows packages to be transmitted between multiple antenna pairs. During the transmission, Wifi signals might be changed due to several reasons, such as the multi-path fading effect or Doppler Effect caused by moving objects in the environment.

As CSI signals are a physical layer indicator of the signal variation through channels, we use an open source CSI tool [36] to collect CSI signals and record a CSI sample containing 30 subcarriers with a bandwidth of 20 MHz centering at the standard 5GHz Wifi. As such, we use a laptop equipped with an Intel 5300 wireless network adapter with 3 antennas as the receiver and a personal computer with the same setting as the transmitter, which allows us to manipulate $3 \times 3$ MIMO mode. As a result, we obtain CSI samples $S \in \mathbb{R}^{N_c \times N_{Tx} \times N_{Rx}}$, where $N_{Tx}$ and $N_{Rx}$ respectively denote the number of transmitting and receiving antennas, and $N_c$ represents the number of subcarriers.

Let $N_S$ denote the ratio between the sampling rate of CSI samples and RGB frames. In other words, for every $N_S$ CSI sample, we synchronize with an RGB frame as one frame-level CSI data. The form of frame-level CSI data is commonly-used in previous works such as PIW [12] and WiSPPN [11]. To better capture the dynamic information across frames, we further take the previous $T-1$ frame-level CSI data with a dilated rate $d$ and current frame-level CSI data as the input to predict one RGB frame (as shown on the left-hand side in Fig. 1). By using the dilated rate $d$, the receptive field is greater across time. Let $C_{seq} \in \mathbb{R}^{T \times N_c \times N_S \times N_{Tx} \times N_{Rx}}$ denote the dilated CSI sequence, which is the input of the spatial-frequency encoder.

### 3.3. Spatial-Frequency encoder

The goal of the proposed spatial-frequency encoder is to extract both the spatial semantic information of the scene and human movement information. Since CSI data are abstract and contain limited information, it is necessary to design an effective feature-extracting pipeline for mining the information from CSI signals. On the one hand, each transceiver pair contains different aspects of the environment channel states, so it is possible to reconstruct 2D spatial background information by integrating the 1D amplitude of each pair. On the other hand, existing works on signal processing [19] show that frequency of CSI amplitude fluctuation changes in accordance with the speed of human movement, which implies human motion can be captured by analyzing consecutive CSI samples. Therefore, we design a model that considers both intra-pair and inter-pair CSI information.

Specifically, for extracting spatial features from $C_{seq}$, the spatial encoder block is composed of one 3D convolution layer with the kernel size of $1 \times 3 \times 3$, followed by BatchNorm, the ReLU activation function and average pooling along the sample dimension. The convolution layer is first used to extract inter-pair information of $3 \times 3$ Wifi transceiver pairs, and since spatial features should not

vary in a short period, the average pooling layer is then applied to integrate the features of all samples, which reduces the effect of noise, and provides more robust spatial features. Moreover, when extracting frequency features, instead of using the band-pass filters in signal processing, we first apply five 3D convolution layers in parallel branches to extract frequency features. To obtain multi-scale features, the kernel sizes of these five convolution layers are set distinctly to extract features from different scales, i.e., the kernel size is set as $k \times 1 \times 1$ with $k \in \{5, 10, 15, 20, 25\}$. After this, a 3D convolution layer with the kernel size of $N_S \times 1 \times 1$ without padding is then applied to each branch as a role to obtain comprehensive information from consecutive samples. Finally, we concatenate frequency features extracted from the five branches with spatial features as the spatial-frequency features.[2]

### 3.4. Evolving attention module

After deriving the spatial-frequency features, we need to address the second challenge that requires a module able to dynamically identify the associated features from specific channels and eliminate unrelated features since the important channels may be different through time. Existing methods of Wifi-based 2D human pose estimation [11,12,16] ignore temporal information, which leads to a mediocre performance for moving users. In addition, for each channel, the amount of information provided by different sequences is not equal. Based on these observations, we design a new evolving attention module to capture two important aspects: (1) the influence between different channels at the same time, and (2) the interactions among the same channel across a time series.

Specifically, we propose an evolving attention module, which derives the dynamic attention masks along the sequence. The spatial-frequency feature sequence $f_{seq} = [f_{t-\Delta t \times d}, f_{t-(\Delta t-1)\times d} \cdots f_t] \in \mathbb{R}^{T\times H\times W\times C}$, where $d$ represents the dilated rate. We use a channel-wise Global Average Pooling (GAP) to extract the global information from $f_{seq}$. The dimension of $f_t$ in the feature sequence is $1 \times 1 \times C$ after GAP. A 1D convolution with kernel size $k$ is then used to generate channel attention mask:

$$A_t = Conv_k(f_t), \tag{1}$$

where $A_t \in \mathbb{R}^{1\times 1\times C}$ represents the generated channel attention mask at the $t$-th timestamp. The parameters of this 1D convolution are shared between features of $f_{seq}$ at different timestamps. Thereafter, we apply a GRU [37] layer to evolve the attention masks by capturing context information through different times ($t$). Let $h_t$ denote the evolving attention mask generated at the $t$-th timestamp and $G$ denotes the function of GRU. The evolving attention is then computed as follows:

$$h_t = G(A_{t-1}, h_{t-1}) \tag{2}$$

### 3.5. Feature decoder and loss

After deriving the enhanced features by the proposed evolving attention module, we use a feature decoder to map the extracted features into two output feature maps: joint heatmaps (JHMs) and part affinity fields (PAFs). The generated JHMs and PAFs are both used for generating human poses via joint association. Therefore, we optimize the loss functions by minimizing the difference of PAFs and JHMs between the poses predicted by our model and OpenPose [20]. The final loss, denoted by $L_{final}$ is summarized as follows.

$$L_{final} = \lambda_1 L_{JHM} + \lambda_2 L_{PAF} \tag{3}$$

[2] The details of the network including dimensions of feature maps, SPE dataset and GPE dataset are presented on our project page.

$$L_{JHM} = \sum_m \|\hat{m} - m\|_2^2, \qquad m \in M \tag{4}$$

$$L_{PAF} = \sum_n \|\hat{n} - n\|_2^2, \qquad n \in N \tag{5}$$

where $M$ and $N$ represent the sets of JHM pixels and PAF pixels, while $\hat{m}$ and $\hat{n}$ are the predicted values. The scalar weights $\lambda_1$ and $\lambda_2$ are used for balancing the two sub-losses.

## 4. Dataset

Due to the lack of benchmarks, we have collected two datasets, General Pose Estimation (GPE) and Specific Pose Estimation (SPE) datasets, to evaluate the performance of the proposed method. We pair the CSI data and RGB frames by matching the frame-level CSI data consisting of 25 CSI samples recorded at 500Hz with one RGB frame sampled at 800 x 600 resolution and 20 fps. We synchronize each RGB frame with frame-level CSI data according to timestamps, and each RGB frame is fed into OpenPose[20] to obtain human joint heatmaps (JHMs) and part affinity fields (PAFs), which serves as the ground truth for prediction.

### 4.1. GPE Dataset

In this dataset, we follow the experimental settings of the mainstream works, such as PIW [12] and Wi-Pose [16]. Nine volunteers are recruited and asked to perform daily activities such as walking, using smartphones, and waving hands. The GPE dataset contains 106,695 data pairs and is collected at 2 spaces, one classroom and one hall, with depths of 9m and 11m respectively. The number of people in the scene ranges from 1 to 5. Moreover, GPE includes data collected under a dark scenario and a through-wall scenario for testing pose estimation under specific conditions. We manually label the human pose as the ground truth for the dark scenario and set up the through-wall scenario by placing the transmitter right outside the concrete wall of the classroom with the receiver and the camera inside the classroom.

### 4.2. SPE Dataset

As the existing datasets do not provide detailed action information, it is difficult to evaluate the performance variance of different actions. Therefore, we collect the SPE dataset to better evaluate the performance of Wifi-based pose estimation under different conditions. Specifically, the SPE dataset is a single person dataset where each subject is asked to perform four specific actions: 1) walking, 2) waving hands, 3) running and 4) jumping jacks. We choose these four actions because they contain a large variety of moving speeds and displacements. In fact, most properties of daily activities involve these four testing actions, which can demonstrate the performance of different kinds of common actions. We recruit 30 volunteers (19 males and 11 females) with heights ranging from 1.56m to 1.92m to evaluate the robustness of different subjects. All data is collected from 6 classrooms or halls of different sizes. The total size of the dataset is 238,191 data pairs.

## 5. Experiments

### 5.1. Experimental setup

To train the proposed EASFN, we use Adam as the optimizer with its default parameters. The batch size is set at 32 with the initial learning rate at 0.001, and a learning rate scheduler is applied to decrease the learning rate by 0.5 every 5 epochs. The number of training epochs is 40. We used the OpenPose Python API to

**Table 1**
Comparisons on the proposed benchmarks (PCK@20).

| benchmark | WiSPPN | PIW | Wi-Mose | EASFN |
|---|---|---|---|---|
| SPE | 21.86% | 32.96% | 33.13% | **50.05%** |
| GPE | × | 27.64% | × | **43.98%** |

conduct multi-person joint association given JHMs and PAFs. The output tensor is $p \times 18 \times 3$, where $p$ represents the number of persons that the networks detected, $18 \times 3$ denotes the x-axis, y-axis, and confidences of 18 body joints.

For the evaluation metric, Percentage of Correct Keypoints (PCK) is a widely-used metric for evaluating the performance of pose estimation, which measures the probability that a model correctly predicts the location of human joints within a tolerated error. Since the human body size varies in video frames due to displacement, the tolerated error should change with different human body sizes. Therefore, we take the distance from the right shoulder to the left hip as torso length and also correlate tolerated error with torso length using $\alpha$ as follows.

$$PCK_i@\alpha = \frac{1}{N}\sum_{n}^{N} \Phi\left(\frac{\|\hat{p}_{(i,n)} - p_{(i,n)}\|_2}{\|p_{(5,n)} - p_{(8,n)}\|_2}\right) \tag{6}$$

$$\Phi(x) = \begin{cases} 1, & \text{if } x < \alpha\% \\ 0, & \text{otherwise,} \end{cases} \tag{7}$$

where $i$ denotes the index of human joints, i.e., $i \in \{1,2,...,18\}$ for the COCO_18 model in OpenPose [20] and $N$ represents the number of people in the test frames; $p_{(i,n)}$ and $\hat{p}_{(i,n)}$ refer to the coordinates of $i$-th joints of $n$-th person obtained from ground truth and prediction, respectively; $\|p_{(5,n)} - p_{(8,n)}\|_2$ is torso length since the index of right shoulder and left hip are 5 and 8. We compare our method with other methods with different $\alpha$, *i.e.*, $\alpha$ is set from 10 to 50. Note that PIW normalizes the error by a fraction of the bounding box size instead of torso length, and thus the PCK value reported in their work is greater.

Our train-test split strategy follows PIW, by taking the first 80% of samples for training, and the last 20% for testing.

### 5.2. Quantitative results

We compare our approach with two state-of-the-art methods for Wifi-based 2D human pose estimation. The first is WiSPPN [11], which is an end-to-end CNN network supervised by a joint coordinate adjacency matrix. The second is PIW [12], which leverages a well-developed CNN to map Wifi signals into three supervision components: SM, JHMs, and PAFs. The third is Wi-Mose [32], which fuses the amplitude and phase of CSI signals into CSI images and uses a neural network to extract features from them.

Table 1 compares EASFN with WiSPPN and PIW in terms of *PCK*@20. Please note that WiSPPN can only estimate poses in single-person scenes due to the limitation of the fixed dimension ($18 \times 18 \times 3$) of the supervision matrix. Therefore, we only train and test WiSPPN on the SPE dataset. The results show that our model outperforms other two approaches in both scenarios. In terms of *PCK*@20, our model outperforms WiSPPN and PIW on the SPE dataset by 28% and 17%, respectively. On the GPE dataset, the proposed EASFN outperforms PIW by 16% in terms of *PCK*@20. This improvement is attributed to our spatial-frequency encoder, which captures the temporal relationship between consecutive CSI signals by multiple frequency encoders of different kernel sizes while previous works ignore such temporal information. For instance, Wi-Mose treats consecutive CSI signals as images and uses kernels of the small receptive field ($3 \times 3$) to extract features, which performs poorly at capturing long-term patterns. Also, WiSPPN and PIW simply aggregate both subcarriers and consecutive samples by channel dimension of convolutional layers, which loses the sequential relationship. Therefore, without long-term consecutive information, previous works perform inferior to EASFN. The PCK curves of all approaches are provided in Fig. 2, which illustrates that EASFN outperforms other methods by a large margin with respect to all PCK thresholds.

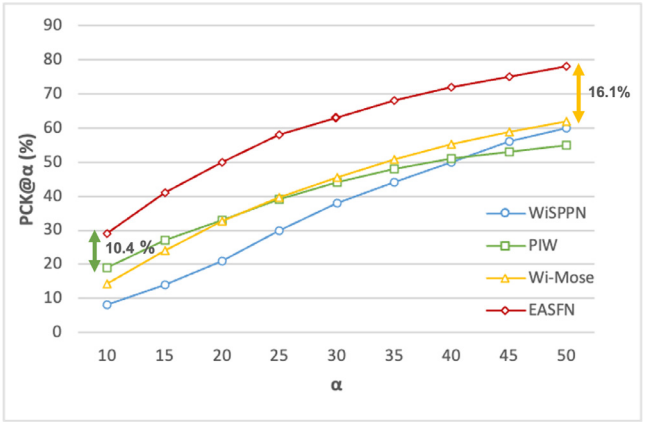


**Fig. 2.** PCK curve of three models evaluated on SPE benchmark.

**Table 2**
Comparative results against other methods on our SPE benchmark (*PCK*@20 for different actions).

| Action | WiSPPN | PIW | Wi-Mose | EASFN |
|---|---|---|---|---|
| Walking | 23.58% | 39.87% | 42.97% | **61.14%** |
| Waving | 25.92% | 33.06% | 33.49% | **45.81%** |
| Running | 22.12% | 37.91% | 35.71% | **58.11%** |
| Jumping | 15.82% | 20.99% | 21.97% | **35.15%** |

### 5.3. Detailed analysis on each action

We further investigate the performance of our model with other methods with different human actions on the SPE dataset. Table 2 presents the comparative results, revealing the advantage of EASFN for predicting moving subjects. Specifically, for the actions of Walking and Running, EASFN significantly improves the baselines, i.e., by about 21% over PIW and 37% over WiSPPN, because the frequency features are sensitive to moving subjects, while the evolving attention module adaptively adjusts the weights of each channel to focus on specific frequency features. The lowest improvement is in the cases of waving, i.e., only 12% over PIW and 20% over WiSPPN. This is because volunteers only wave their hands or arms while keeping their torso static, and the performance on static keypoints is similar for the different methods.

From Table 2, we can also observe that EASFN achieves the worst result when handling the pose of jumping. Accurately estimating the human pose of jumping is an arduous task. It requires fine time granularity to capture the thin and fast-moving limbs. Specifically, jumping can be regarded as high-speed waving, and therefore using CSI sampled from the same time granularity to estimate the movement of higher speed would lead to a performance drop. In that case, one possible solution to improve EASFN's ability to handle jumping is to increase the sampling rate of CSI samples, which can refine the time granularity of data. To investigate the relation between the sampling rate of CSI and pose estimation result, we need to collect a larger dataset including multiple sampling rates, which is left as our future work.

**Table 3**
Ablation study on our SPE benchmark.

| Combination | Module | | | | Metric |
|---|---|---|---|---|---|
| | SFE | AM | EM | d | *PCK*@20 |
| Baseline | - | - | - | - | 35.13% |
| C1 | √ | - | - | - | 44.99% |
| C2 | - | √ | - | - | 45.95% |
| C3 | - | - | √ | 1 | 48.08% |
| **C4** | **-** | **-** | √ | **3** | **50.05%** |
| C5 | - | - | √ | 5 | 47.95% |

### 5.4. Ablation study

To demonstrate the advantage of each component in EASFN, we conduct an ablation study on the SPE dataset. Five combinations are conducted for comparison. (Baseline): taking the ResNet as feature encoder same as PIW and WiSPPN; (C1): using spatial-frequency encoder; (C2): applying the static attention module; (C3-C5): proposed EASFN with different dilated rates.

Table 3 shows the results of different models in terms of *PCK*@20, which verifies our assumptions. First of all, the baseline has better performance than PIW, demonstrating that Unet is not suitable for Wifi signal and thus leads to worse performance. In contrast, C1 outperforms the baseline by a large margin since the spatial-frequency encoder is designed for Wifi and can extract both the spatial semantic information of the environment and human movement information. Moreover, the attention mechanism (C2) is beneficial for pose estimation by focusing on certain channels. However, if the attention masks are derived simply from single frame-level features, the performance only slightly increases by around 1% in terms of *PCK*@20. On the other hand, the proposed evolving attention module (C3-C5) can increase the performance at least by 2%. The different outcomes between C3 and C4 imply that increasing the receptive field across time can improve performance by around 2%. Nevertheless, a comparison of the results of C4 and C5 shows that if the receptive field is too large across time, it can deteriorate the performance. This is why we take EASFN with d=3 as the proposed method.

### 5.5. Qualitative results

Fig. 3 illustrates the examples of human poses generated from CSI signals by EASFN and the other methods. We choose samples from three subjects with different actions (walking, waving, and complex stretching actions). The results show that EASFN generates more accurate poses for different actions since it effectively learns the static spatial features and dynamic frequency features across continuous frames. The other contenders can approximately detect the human torso but perform poorly at estimating joints in human arms and legs. These joints are difficult to predict because arms and legs usually have complicated actions across the time. However, with the evolving attention module, EASFN can pay more attention to the specific channels of moving arms and legs. By learning the dependency between channels across the time, EASFN estimates the joints in arms or legs accurately.

### 5.6. Through-wall and low-light scenarios

In order to demonstrate that the proposed approach offers strong advantages under specific scenarios, we conduct experiments under a through-wall and a low-light scenario. For the through-wall scenario, the transmitter is placed right outside the concrete wall of the classroom, and signals would pass through the wall, reflect off the human body, and then be received by the receiver. The task is challenging because the CSI would be dominated by the through-wall degradation. From Table 4, we can observe that PIW suffers a severe performance drop under the through-wall scenario. On the flip side, though having a perceptible degradation, EASFN still remains an acceptable performance under such harsh conditions. The result indicates that EASFN is more robust compared with PIW when capturing useful information from through-wall signals.

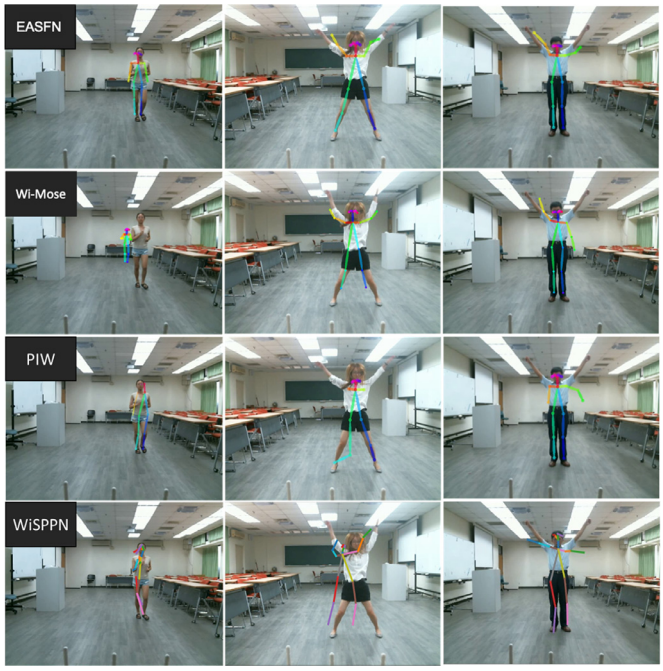


**Fig. 3.** Comparative visualization of EASFN and other methods.

**Table 4**
Comparative results under through-wall and low-light scenarios(PCK@20).

| scenario | PIW | Openpose | EASFN |
|---|---|---|---|
| Wall | 8.92% | - | **30.45%** |
| Dark | - | 42.89% | **46.97%** |

For exhibiting the power of EASFN under low-light scenarios, we design a comparative experiment in a dark environment. We separately perform human pose estimation from CSI data via

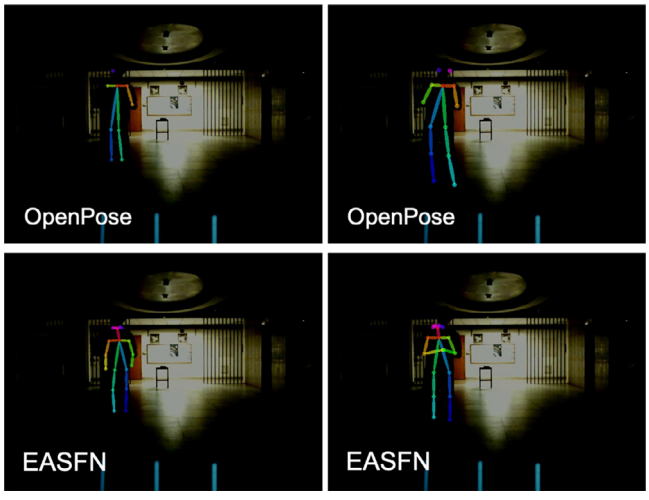


**Fig. 4.** Comparison of results of EASFN and OpenPose in the dark.

EASFN and RGB images via Openpose and evaluate the performance with the manually labeled human pose. Table 4 shows that Openpose behaves poorly when the lights are turned off while EASFN is not affected at all. Fig. 4 provides a better understanding of the impact of turning off the lights. Under poor illumination, Openpose is likely to lose some keypoints, such as wrists, ankles, and so forth. However, EASFN can correctly predict all the keypoints. Therefore, EASFN can be effectively applied in some suboptimal scenarios, such as in poorly-lit conditions or even in the dark.

## 6. Conclusion

We propose an evolving attentive spatial-frequency network for Wifi-based 2D human pose estimation. To learn effective representations from abstract CSI signals, we extract static spatial and dynamic frequency information simultaneously via a spatial-frequency encoder. We also leverage an evolving attention module to learn channel weights and temporal dependency between consecutively dilated frames. In addition, two datasets are introduced and publicly released for further research. Extensive experiments show that our model outperforms state-of-the-art approaches by a large margin. Though we have achieved considerable progress, there are still challenges left unsolved, which are crucial for applying Wifi-based human pose estimation in a real-world application. In this work, we conduct experiments under well-controlled environments. However, in a real-world scenario, other moving objects, i.e. oscillating fans, or house pets, may lead to a more complicated Wifi response. How to adapt Wifi-based human pose estimation to such a challenging scenario is left as our future work.

## Declaration of Competing Interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

Hong-Han Shuai reports financial support was provided by Ministry of Science and Technology of Taiwan.

## Data availability

I have shared the link to my data and codes in the abstract.

## References


[1] A. Kamel, B. Sheng, P. Li, J. Kim, D. Feng, Hybrid refinement-correction heatmaps for human pose estimation, IEEE Trans. Multimedia 23 (2021) 1330–1342.

[2] X. Wang, X. Hu, Y. Li, C. Jiang, Multi-modal human pose estimation based on probability distribution perception on a depth convolution neural network, Pattern. Recognit. Lett. 153 (2022) 36–43.

[3] H. Yu, C. Du, L. Yu, Scale-aware heatmap representation for human pose estimation, Pattern. Recognit. Lett. 154 (2022) 1–6.

[4] Y.-J. Ma, H.-H. Shuai, W.-H. Cheng, Spatiotemporal dilated convolution with uncertain matching for video-based crowd estimation, IEEE Trans. Multimedia 24 (2022) 261–273.

[5] M.-C. Hu, W.-H. Cheng, C.-S. Hu, J.-L. Wu, J.-W. Li, Efficient human detection in crowded environment, Multimedia Syst. 21 (2014) 177–187.

[6] C.-H. Sio, Y.-J. Ma, H.-H. Shuai, J.-C. Chen, W.-H. Cheng, S2siamFC: self-supervised fully convolutional siamese network for visual tracking, Proceedings of the 28th ACM International Conference on Multimedia (2020).

[7] Q. Bao, W. Liu, Y. Cheng, B. Zhou, T. Mei, Pose-guided tracking-by-detection: robust multi-person pose tracking, IEEE Trans. Multimedia 23 (2021) 161–175.

[8] J. Sanchez-Riera, K. Srinivasan, K.-L. Hua, W.-H. Cheng, M.A. Hossain, M.F. Alhamid, Robust RGB-d hand tracking using deep learning priors, IEEE Trans. Circuits Syst. Video Technol. 28 (9) (2018) 2289–2301.

[9] C.W. Hsieh, C.Y. Chen, C.L. Chou, H.H. Shuai, J. Liu, W.H. Cheng, Fashionon: semantic-guided image-based virtual try-on with detailed human and clothing information, in: Proceedings of the 27th ACM International Conference on Multimedia, 2019, pp. 275–283.

[10] C.-L. Chou, C.-Y. Chen, C.-W. Hsieh, H.-H. Shuai, J. Liu, W.-H. Cheng, Template-free try-on image synthesis via semantic-guided optimization, IEEE Trans. Neural Netw. Learn. Syst. (2021) 1–14.

[11] F. Wang, S. Panev, Z. Dai, J. Han, D. Huang, Can wifi estimate person pose? arXiv preprint arXiv:1904.00277 (2019).

[12] F. Wang, S. Zhou, S. Panev, J. Han, D. Huang, Person-in-wifi: fine-grained person perception using wifi, in: 2019 IEEE International Conference on Computer Vision, 2019, pp. 5451–5460.

[13] J. Wang, Y. Chen, S. Hao, X. Peng, L. Hu, Deep learning for sensor-based activity recognition: a survey, Pattern. Recognit. Lett. 119 (2019) 3–11.

[14] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.

[15] O. Ronneberger, P. Fischer, T. Brox, U-net: convolutional networks for biomedical image segmentation, in: MICCAI, 2015, pp. 234–241.

[16] L. Guo, Z. Lu, X. Wen, S. Zhou, Z. Han, From signal to image: capturing fine-grained human poses with commodity wi-fi, IEEE Commun. Lett. 24 (2020) 802–806.

[17] C. Shi, J. Liu, H. Liu, Y. Chen, Smart user authentication through actuation of daily activities leveraging wifi-enabled iot, Proc. 18th ACM Int. Sympos. Mobile Ad Hoc Network. Comput. (2017).

[18] L. Guo, L. Wang, J. Liu, W. Zhou, B.L.T. Liu, G. Li, C. Li, A novel benchmark on human activity recognition using wifi signals, in: 2017 IEEE 19th International Conference on e-Health Networking, Applications and Services, 2017, pp. 1–6.

[19] W. Wang, A.X. Liu, M. Shahzad, K. Ling, S. Lu, Understanding and modeling of wifi signal based human activity recognition, in: Proceedings of the 21st Annual International Conference on Mobile Computing and Networking, Association for Computing Machinery, 2015, pp. 65–76.

[20] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, Y. Sheikh, Openpose: realtime multi-person 2d pose estimation using part affinity fields, IEEE Trans. Pattern Anal. Mach. Intell. 43 (1) (2021) 172–186.

[21] S. Kreiss, L. Bertoni, A. Alahi, Pifpaf: composite fields for human pose estimation, in: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, 2019, pp. 11969–11978.

[22] Z. Geng, K. Sun, B. Xiao, Z. Zhang, J. Wang, Bottom-up human pose estimation via disentangled keypoint regression, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2021, pp. 14676–14686.

[23] K. Su, D. Yu, Z. Xu, X. Geng, C. Wang, Multi-person pose estimation with enhanced channel-wise and spatial information, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 5674–5682.

[24] Z. Liu, H. Chen, R. Feng, S. Wu, S. Ji, B. Yang, X. Wang, Deep dual consecutive network for human pose estimation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2021, pp. 525–534.

[25] C. Peng, T. Xiao, Z. Li, Y. Jiang, X. Zhang, K. Jia, G. Yu, J. Sun, Megdet: a large mini-batch object detector, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 6181–6189.

[26] K. He, G. Gkioxari, P. Dollár, R. Girshick, Mask r-cnn, in: Proceedings of the IEEE international conference on computer vision, 2017, pp. 2961–2969.

[27] X. Wang, X. Wang, S. Mao, Deep convolutional neural networks for indoor localization with CSI images, IEEE Trans. Network Sci. Eng. 7 (2020) 316–327.

[28] Y. Jing, J. Hao, P. Li, Learning spatiotemporal features of CSI for indoor localization with dual-stream 3d convolutional neural networks, IEEE Access 7 (2019) 147571–147585.

[29] J.-Y. Chang, K.-Y. Lee, Y.-L. Wei, K.C.-J. Lin, W. Hsu, Location-independent wifi action recognition via vision-based methods, in: Proceedings of the 24th ACM International Conference on Multimedia, in: MM '16, 2016, pp. 162–166.

[30] H. Zou, Y. Zhou, J. Yang, W. Gu, L. Xie, C.J. Spanos, Wifi-based human identification via convex tensor shapelet learning, in: Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence„ 2018, pp. 1711–1719.

[31] W. Jiang, H. Xue, C. Miao, S. Wang, S. Lin, C. Tian, S. Murali, H. Hu, Z. Sun, L. Su, Towards 3d human pose construction using wifi, in: Proceedings of the 26th Annual International Conference on Mobile Computing and Networking, 2020, pp. 1–14.

[32] Y. Wang, L. Guo, Z. Lu, X. Wen, S. Zhou, W. Meng, From point to space: 3d moving human pose estimation using commodity wifi, IEEE Commun. Lett. 25 (2020) 2235–2239.

[33] F. Adib, C.-Y. Hsu, H. Mao, D. Katabi, F. Durand, Capturing the human figure through a wall, ACM Trans. Graph. 34 (2015) 1–13.

[34] M. Zhao, T. Li, M.A. Alsheikh, Y. Tian, H. Zhao, A. Torralba, D. Katabi, Through-wall human pose estimation using radio signals, in: 2018 IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7356–7365, doi:10.1109/CVPR.2018.00768.

[35] C. Yang, X. Wang, S. Mao, Rfid-pose: vision-aided three-dimensional human pose estimation with radio-frequency identification, IEEE Trans. Reliab. 70 (2020) 1218–1231.

[36] D. Halperin, W. Hu, A. Sheth, D. Wetherall, Tool release: gathering 802.11 n traces with channel state inforicroation, ACM SIGCOMM Comput. Commun. Rev. 41 (1) (2011). 53–53

[37] K. Cho, B. van Merrienboer, Ç. Gülçehre, D. Bahdanau, F. Bougares, H. Schwenk, Y. Bengio, Learning phrase representations using RNN encoder-decoder for statistical machine translation, in: EMNLP, 2014, pp. 1724–1734.