HPE-Li: WiFi-enabled Lightweight Dual Selective Kernel Convolution for Human Pose Estimation

Toan D. Gian[1], Tien Dac Lai[1], Thien Van Luong[2], Kok-Seng Wong[1], and Van-Dinh Nguyen[1]*

[1] College of Engineering and Computer Science & Center for Environmental Intelligence, VinUniversity, Hanoi, Vietnam (email: {toan.gd, 20tien.ld, wong.ks, dinh.nv2}@vinuni.edu.vn)

[2] Faculty of Computer Science, Phenikaa University, Hanoi, Vietnam (email: thien.luongvan@phenikaa-uni.edu.vn)

Abstract. WiFi-based human pose estimation (HPE) has emerged as a promising alternative to conventional vision-based techniques, yet faces the high computational cost hindering its widespread adoption. This paper introduces a novel HPE-Li approach that harnesses multi-modal sensors (*e.g.* camera and WiFi) to generate accurate 3D skeletal in HPE. We then develop an efficient deep neural network to process raw WiFi signals. Our model incorporates a distinctive multi-branch convolutional neural network (CNN) empowered by a selective kernel attention (SKA) mechanism. Unlike standard CNNs with fixed receptive fields, the SKA mechanism is capable of dynamically adjusting kernel sizes according to input data characteristics, enhancing adaptability without increasing complexity. Extensive experiments conducted on two MM-Fi and WiPose datasets underscore the superiority of our method over state-of-the-art approaches, while ensuring minimal computational overhead, rendering it highly suitable for large-scale scenarios.

Keywords: Attention, convolution neural network, human pose estimation, selective kernel, wireless sensing.

1 Introduction

Human activity monitoring-enabled methods primarily rely on video-based sensors and wearable devices [27], showcasing high accuracy in activity identification but facing several obstacles in the practical deployment [2]. The discomfort of wearing devices during vigorous activity, alongside limitations inherent to camera systems such as fixed angles, occlusions, glare interference and low-light conditions, poses significant barriers. Other concerns regarding personal privacy further impede the widespread adoption of traditional methods in monitoring human tasks.

Multi-modal sensors for HPE tasks have been developed to overcome the limitations of traditional sensors. Although radio frequency (RF)-based solutions

* Corresponding author

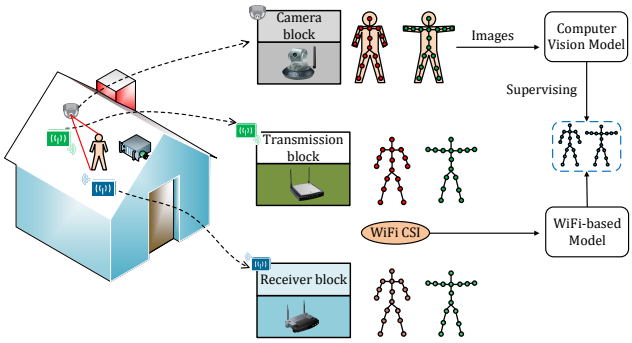


Fig. 1: An example of implementing a human pose estimation application using commercial WiFi in the indoor environment.

have seen extensive adoption [52–54], their potential applications are restricted by hardware constraints, such as the need for a meticulously assembled and synchronized 16+4 T-shaped antenna array, and RF signal limitations, such as the use of frequency modulated continuous wave with a wide signal bandwidth of 1.78 GHz. In contrast, the WiFi-based approach has recently emerged as a promising alternative by leveraging the widespread nature of WiFi signals in practical environments. This approach offers a non-intrusive, privacy-aware solution that mitigates the inherent limitations of traditional methods. The interaction between WiFi signals and the hierarchical structure of the human suggests the feasibility of extracting pose features from diverse WiFi channel state information (CSI) [5], thus enabling HPE tasks. Once WiFi-based approaches achieve reliable performance, they could find widespread implementation in indoor applications, as depicted in Fig. 1.

However, the implementation of WiFi-based HPE encounters several challenges. In particular, WiFi CSI data often lacks the necessary detail for fine-grained tasks, leading to compromised accuracy in pose estimation due to the intertwining of human motions with environmental factors. Moreover, WiFi signals carry unique fingerprints of both the environment and individuals, making the trained models less effective in other contexts. Additionally, synthesized skeleton movements must mirror natural human movement patterns, requiring continuity and smoothness. Recent research has focused on addressing these challenges through various approaches [20, 32, 41, 44, 55, 56]. The resolution and characteristics of WiFi CSI signals are enhanced by utilizing the multi-antenna technique or increasing the number of subcarriers [41, 50, 56]. Meanwhile, deep neural networks, particularly convolutional neural networks (CNNs) with fixed kernels, are developed to capture local features [20, 32, 41, 44] or combined with transformer [39] to obtain global information [55, 56]. However, existing approaches often face a trade-off between performance and computational cost. For instance, the work in [56] is shown to achieve good accuracy but requires high computational complexity (*e.g.* 26.42 million parameters), whereas [9] exhibits low com-

plexity yet with unreliable performance. These challenges naturally give rise to a question: *Is it possible to achieve better performance for WiFi-based HPE tasks, while still guaranteeing a very low computational complexity?*

Our answer is "Yes," by designing a novel multi-model network, *namely* HPE-Li, to generate the lightweight and efficient human pose estimation from WiFi CSI signals. Inspired by the teacher-student network concept [28], the proposed model comprises two distinct networks: a pre-trained teacher network specialized in pose estimation from RGB images and a student network tasked with making pose predictions from raw WiFi signals under the teacher network's supervision. The student network employs dual selective kernel sub-networks (DSKNet) to efficiently learn key characteristics from WiFi CSI signals without increasing complexity. Within DSKNet, we introduce a novel convolutional architecture, called DSKConv, which incorporates a selective kernel attention (SKA) mechanism to dynamically learn diverse features from multiple kernels across both frequency and channel domains, rather than solely from the channel domain as in regular selective kernel (SK) convolution [25]. Unlike basic CNN, DSKConv adaptively selects optimal kernel sizes for each input through a three-step process: *i*) generating multiple paths with different kernel sizes, *ii*) consolidating multiscale features, and *iii*) using an attention mechanism to determine selection weights.

In summary, the main contributions of the paper are three-fold:

1. We introduce a novel HPE-Li system that integrates multi-modal sensors (*e.g.* Camera and WiFi) to accurately generate the skeleton-based HPE from WiFi CSI signals.
2. We propose DSKNet sub-networks to effectively process WiFi CSI signals for HPE tasks. DSKNet employs multi-view feature extraction from signals received by each antenna, enabling a comprehensive understanding of human posture through the fusion of information from all antennas.
3. We develop a novel DSKConv architecture, which is seamlessly integrated into DSKNet. This architecture serves as a versatile block capable of substituting standard convolutional layers in various models. DSKConv dynamically learns diverse features from multiple kernels without imposing additional computational costs.

Extensive experiments are conducted on two challenging datasets to demonstrate the effectiveness of HPE-Li. Results confirm the adaptive adjustment of kernel sizes by DSKConv to achieve multiscale features from input data. Comparative analysis showcases the superior accuracy of the HPE-Li model compared to existing approaches, all while maintaining much less complexity. Moreover, experimental results indicate a significant improvement of DSKConv over the state-of-the-art (SOTA) variant CNNs utilizing fixed-size kernels, with only a minimal increase in memory and computational requirements. The source code is available at here for research purposes.

2 Related Work

HPE task: In the field of computer vision (CV), HPE from 2D images has been extensively studied thanks to powerful deep learning techniques and increasingly annotated datasets produced by regular cameras [4,6,10–12,14,17,30,31,46] and specialized equipment [33, 51]. However, vision-based approaches face several challenges such as poor lighting, occlusion, blurry images and privacy concerns. Despite attempts to address privacy concerns with light-based methods [23,24], these methods struggle in low-light conditions or with obstacles. Although LiDAR has been shown to provide high person detection capabilities [26, 45], its cost and power limitations warrant the exploration of alternative, more accessible solutions for daily and household use. The proposed WiFi-based approach overcomes these limitations by ensuring privacy protection and operability regardless of lighting or occlusion. RF-based approaches have emerged as the promising solution to overcome the aforementioned challenges. In particular, RFCapture in [1] can outline human bodies even through walls, while RFPose in [52] and its 3D variant [54] can extract 2D and 3D skeletons from RF signals with the aid of visual data. However, these methods often require advanced hardware and signal conditions, limiting their practical applications.

WiFi-based HPE: In recent years, WiFi-based sensing methods have attracted significant attention, mainly relying on received signal strength (RSS) and channel state information (CSI). Exploiting the ubiquitous presence of WiFi and the widespread use of smart devices in various environments, Zou *et al.* [57] introduced an indoor localization system based on RSS, providing a feasible alternative to traditional global positioning systems (GPS). Conversely, WiFi CSI holds promise for enhancing daily activities like activity recognition and health monitoring. However, extracting high-quality data from raw WiFi signals poses a significant challenge compared to other sensing methods. Moreover, the absence of standardized WiFi CSI setup protocols and the scarcity of publicly available datasets, primarily due to cost constraints, impede the comparison and enhancement of HPE tasks. WiPose [20] achieved a good performance in terms of the 3D human pose estimation using dedicated antenna setups and VICON, albeit at a high cost. WiSPPN [41] employed standard WiFi devices with 30 subcarriers, albeit sacrificing pose resolution. Notably, MetaFi++ [56] utilized widely available WiFi devices for HPE to improve the HPE performance. However, the high computational complexity of its WPFormer network hinders its real-time application. In this work, we showcase the effectiveness of the proposed HPE-Li on three critical factors: accessible, high-performance WiFi systems and lightweight neural network models.

Variants of CNN: The success of CNNs in the CV field has spurred researchers to explore diverse CNN variations, in which the utilization of multi-branch convolution is highlighted as a key aspect [3]. Expanding on this concept, highway networks were introduced in [34] to incorporate bypassing paths to gat-

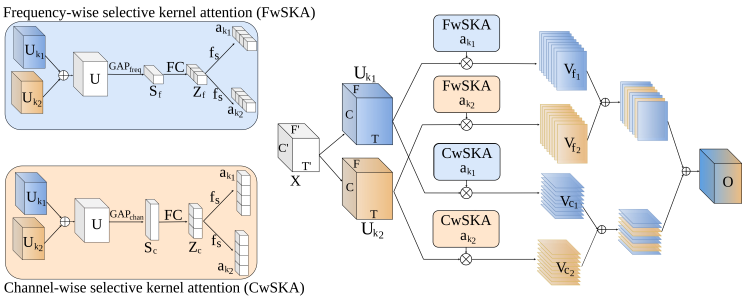


Fig. 2: The proposed DSKConv with two branches consists of the channel-wise selective kernel attention (CwSKA) mechanism (Sec. 3.1) and the frequency-wise selective kernel attention (FwSKA) mechanism (Sec. 3.2).

ing units that facilitate the training of deep networks with hundreds of layers. ResNet [15] adopted a similar bypassing path strategy to preserve original information. Building upon this, BlockDrop [49] introduced additional identical paths for significant transformations. In contrast, InceptionNet [36–38] developed a new approach by amalgamating multiple branches with customized kernel filters, resulting in the extraction of richer and more diverse features. In a recent advancement, attention mechanisms have been seamlessly integrated within convolutional layers. This approach allows the intelligent merging of information from multiple kernels and enable the creation of effective receptive fields with varying sizes within fusion layers [7, 19, 25, 43, 48]. In this work, rather than directly utilizing multi-branch convolution, we develop a novel convolutional architecture by incorporating a SKA mechanism to effectively learn diverse features from multiple kernels in both frequency and channel domains with a very low complexity.

3 The Proposed HPE-Li Method

We now discuss the CwSKA mechanism employed in traditional SK convolution [25] and then present the proposed convolution, which integrates both CwSKA and FwSKA to harness channel and frequency domains. Next, we provide insights into the network architecture and learning objectives of the proposed model.

3.1 Traditional SK Convolution

Let us start by presenting the operation CwSKA mechanism of traditional SK convolution, as illustrated in Fig. 2. The traditional SK convolution is expressed as follows. For a given $\mathbf{X} \in \mathbb{R}^{C' \times F' \times T'}$, with C', F' and T' being the number of channels, height in the frequency domain and width in the time domain, respectively, let the transform $\mathcal{F}_{k_i}$: $\mathbf{X} \rightarrow \mathbf{U}_{k_i} \in \mathbb{R}^{C \times F \times T}$ be the function that splits the input feature map $\mathbf{X}$ into N branches. Each transformation consists of a sequence of efficiently grouped convolutions [22, 25], that has a pre-defined

kernel size, denoted as $\{k_i\}_{i=1}^{N}$, batch normalization (BN) [18] and ReLU activation [13]. The group number G introduced in AlexNet [21] divides model parameters and computational load into G parts to better utilize the GPU resource. The SK convolution integrates the grouped convolution and dilated convolution into branches with larger kernel sizes to reduce model overheads [8, 25, 47]. The dilation factor D expands the receptive field, allowing for a broader view of convolutional networks. The dilated convolution offers lower model complexity while capturing more contextual information and achieving faster runtime.

Incorporating various levels of information into the subsequent stage involves merging N branches through an element-wise summation, *i.e.* $\mathbf{U} = \sum_{i=1}^{N} \mathbf{U}_{k_i}$. Subsequently, the 2D global average pooling (GAP) operation encapsulates the global information into the channel-wise feature vector $\mathbf{s}_c \in \mathbb{R}^C$, $\mathrm{GAP}_{\mathtt{chan}}$, as follows:

$$\mathbf{s}_c = \mathrm{GAP}_{\mathtt{chan}}(\mathbf{U}) = \frac{1}{H \times W} \sum_{h=1}^{H} \sum_{\mathrm{w}=1}^{W} \mathbf{U}(h, w). \tag{1}$$

Subsequently, a fully connected (FC) layer is used to generate a more compact feature map $\mathbf{z}_c \in \mathbb{R}^{d \times 1}$, which enables the model to perform the SK operation efficiently. During the dimension reduction, the value of d is controlled by a reduction ratio r. The compact feature map $\mathbf{z}_c$ and the reduction ratio r can be computed as

$$\mathbf{z}_c = \mathrm{FC}(\mathbf{s}_c) = \mathrm{ReLU}(\beta(\mathbf{W}\mathbf{s}_c)) \tag{2}$$

where β refers to the BN operation, $\mathbf{W} \in \mathbb{R}^{C \times d}$ denotes the weight matrix of a FC layer and d is the dimension of $\mathbf{z}_c$; Herein, $d = C/r$ with r represents the reduction ratio of the compact feature map $\mathbf{z}_c$.

The SK convolution then applies a soft attention mechanism for the compact feature map $\mathbf{z}_c$ passed down from the previous layer, which can guide the model to adaptively extract multiscale information across the channel axis [7, 25, 42]. The softmax attention can focus on the important branches and play a key role in the adaptive kernel selection. The soft attention weights $\mathbf{a}_{k_i} \triangleq [a_{k_i,1}, \cdots, a_{k_i,C}]^T \in \mathbb{R}^C$ are calculated via a softmax function f_{s} as

$$a_{k_i,j} = f_{\mathrm{s}}(\mathbf{Z}_c) = \frac{\exp(A_{k_i,j}\mathbf{z}_c)}{\sum_{l=1}^{N} \exp(A_{k_l,j}\mathbf{z}_c)} \tag{3}$$

where $A_{k_i,j} \in \mathbb{R}^d$ is the j-th row of $\mathbf{A}_{k_i} = [A_{k_i,1}^T, \cdots, A_{k_i,C}^T]^T \in \mathbb{R}^{C \times d}$. Finally, the output features map $\mathbf{V}_c \in \mathbb{R}^{C \times H \times W}$ is computed as the weighted summation over the different branches:

$$V_{c_j} = \sum\nolimits_{i=1}^{N} a_{k_i,j} U_{k_i,j}, \text{ with } \sum\nolimits_{i=1}^{N} a_{k_i,j} = 1 \tag{4}$$

where V_{c_j} and $U_{k_i,j} \in \mathbb{R}^{H \times W}$ are the j-th components of $\mathbf{V}_c$ and $\mathbf{U}_{k_i}$, respectively.

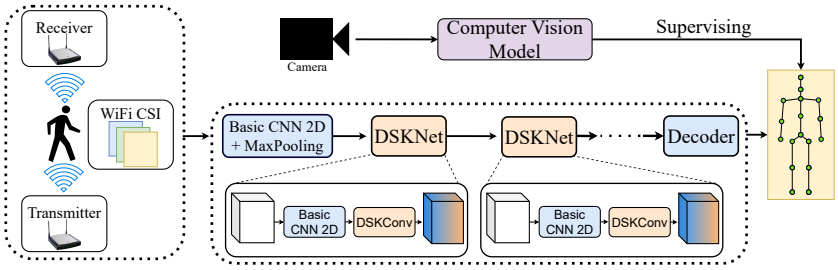


Fig. 3: The teacher-student network: 1) The CV model acts as the teacher network to produce ground truth from the camera images, and 2) the proposed model (in the dashed rectangle) serves as the student network to predict human pose under the monitoring of the teacher network, as discussed in Sec. 3.3.

3.2 Dual SK Convolution (DSKConv)

We notice that the conventional SK method is not tailored to HPE tasks since each frame of WiFi-CSI data is also present in both frequency and channel domains (*i.e.* representing pose information from one specific viewpoint). As a result, the channel-wise recalibration is comprehensively captured. Thus, we introduce an SK convolution-based approach specifically designed to process raw WiFi signals. The proposed DSKConv employs both CwSKA and FwSKA mechanisms to obtain the adaptive kernel in both channel and frequency domains, as shown in Fig. 2. FwSKA utilizes the CwSKA technique similarly, but with a distinction: $\mathbf{s}_f$ represents a frequency-wise feature vector instead of a channel-wise feature vector, which is obtained from the transformation, $\text{GAP}_{\texttt{freq}}$, as

$$\mathbf{s}_f = \text{GAP}_{\texttt{freq}}(\mathbf{U}) = \frac{1}{C \times T} \sum\nolimits_{c=1}^{C} \sum\nolimits_{\text{t}=1}^{T} \mathbf{U}(c, t). \tag{5}$$

The compact feature $\mathbf{z}_f$, attention weight $\mathbf{a}$ and output feature map $\mathbf{V}_f$ are derived in the same manner with CwSKA. This combination ensures that the different channels and the characteristics of distinctive frequencies are considered when selecting the suitable kernel size for convolution. Finally, we concatenate the output features $\mathbf{V}_{f_i}$ and $\mathbf{V}_{c_i}$ at the i-th branches to generate the output feature $\mathbf{O}$ as

$$\mathbf{O} = \sum\nolimits_{i=1}^{N} \mathbf{V}_{c_i} + \sum\nolimits_{i=1}^{N} \mathbf{V}_{f_i}\,. \tag{6}$$

The resulting feature $\mathbf{O}$ contains crucial information extracted from both frequency and channel domains, aligning well with the inherent characteristics of WiFi CSI data.

3.3 Network Architecture

The teacher-student network architecture [28] is illustrated in Fig. 3. This architecture consists of two separate networks: a pre-trained teacher network for pose

estimation from RGB images and a student network to generate pose predictions from raw WiFi signals under the guidance of the teacher network. The teacher network utilizes a computer vision model to obtain 2D key points of frames from the multi-views, such as HRNet-w48 [35, 50]. The 2D key points serve as the ground truth $\mathbf{y}$ to estimate human poses from various modalities, including coordinates of pixels (a, b) written as $\mathbf{y} = \{(a_i, b_i)|i \in [1, \cdots, P]\}$ with P being the number of key points. The student network primarily consists of a stack of repeated sub-networks, so-called "DSKNet", to construct the feature transformation network. The features obtained from DSKNet are finally fed into the decoder module to generate key points prediction $\hat{\mathbf{y}} = \{(\hat{a}_i, \hat{b}_i)|i \in [1, \cdots, P]\}$ of the human body.

DSKConv relies on three essential hyperparameters to dictate the final configuration of SK convolutions. Precisely, the number of branches N decides the variety of kernels, the group number G controls the cardinality of each branch, and the reduction ratio (r) impacts the number of parameters in (2). We elaborate on the impact of these parameters in Sec. 4.3.

3.4 Learning Objective

We consider the mean-squared error (MSE) between the prediction $\hat{\mathbf{y}}$ and ground truth $\mathbf{y}$ generated by the visual pose estimation model as the loss function, which is given as:

$$\mathcal{L}_{\texttt{MSE}} = \|\hat{\mathbf{y}} - \mathbf{y}\|_2^2. \tag{7}$$

Rather than employing the pose adjacency matrix [41], we have found that utilizing the MSE loss function yields better performance. This phenomenon is attributed to the granularity of CSI data. Both MM-Fi and WiPose datasets contain abundant subcarriers of CSI data for each antenna pair, resulting in higher-resolution data. As a result, the task can be successfully processed without necessitating a more complex loss function design.

4 Experiment Results

All experiments are conducted on a computer with an Intel 13-core i9-13900k CPU (3 GHz) and a GeForce RTX 4070. Network models are trained for 50 epochs using the stochastic gradient descent with momentum (SGDM) algorithm, employing a batch size of 32, a learning rate of 0.001, and a momentum of 0.9.

4.1 Dataset and Evaluation Metrics

We evaluate the performance of the HPE-Li framework on two challenging datasets (MM-Fi [50], WiPose [55]). These are briefly provided in Table 1 with more detailed descriptions below.

MM-Fi [50] includes 17 skeleton points of pose annotations from the camera sensor and WiFi CSI data, collected from 40 human subjects with 27 action

Table 1: Detail of MM-Fi and WiPose datasets.

Dataset	Activity×Subject	Tx×Rx×Subcarrier×Time	Frequency	Packets	Training	Testing
MM-Fi	27×40	$1 \times 3 \times 314 \times 10$	5 GHz	320K	70%	30%
WiPose	12×12	$3 \times 3 \times 30 \times 5$	5 GHz	166K	70%	30%

categories over 14 daily activities and 13 rehabilitation exercises. In addition, MM-Fi utilizes three protocols aligned with the benchmark setup presented in [50]. Specifically, Protocol 1 (P1) involves 14 daily activities performed freely in space, such as picking up objects and raising arms. Protocol 2 (P2) comprises 13 activities conducted in a fixed location, such as limb extension. Protocol 3 (P3) includes all 27 aforementioned activities. Each protocol employs two data splitting strategies: *i*) Setting 1 (S1-Random Split) randomly divides all video samples into training and testing sets with a ratio of 3:1, and *ii*) Setting 2 (S2-Cross-Subject Split) splits the data by subject, allocating 32 subjects for training and 8 for testing.

WiPose [55] comprises 166k packets, featuring pose annotations with 18 skeleton points and WiFi CSI of 12 different actions (i.e. wave, walk, throw, run, push, pull, jump, crouch, circle, sit down, stand up and bend) performed by 12 volunteers. All WiPose's data are randomly split into training and testing sets.

To evaluate pose estimation, we consider three key performance metrics, which are the Percentage of Correct Keypoints (PCK), Mean Per Joint Position Error (MPJPE), and Procrustes Analysis MPJPE (PA-MPJPE) in millimeters. The PCK metric evaluates the accuracy of predicted keypoints by calculating the percentage of keypoints correctly localized within a predefined threshold a from their ground truth positions, denoted as PCK_a. The MPJPE metric calculates the Euclidean distance between the predicted joint positions and ground truth positions, while PA-MPJPE is actually the MPJPE after aligning the predicted results to the ground truth through a procrustes transformation.

Benchmarks: To showcase the effectiveness of HPE-Li, we conduct a comprehensive comparison with several SOTA models. We will mainly consider MetaFi++ [56] and PerUnet [55], which are shown to achieve the best performance on MM-Fi and WiPose datasets. Additionally, we extend the comparison to include results from previously well-established methods for the HPE task, such as Wi-Pose [20], Wi-Mose [44], WiLDAR [9] and WiSPPN [41].

4.2 Results and Discussion

Results on MM-Fi: We begin by evaluating the proposed HPE-Li's performance in the **P3-S1** scenario using PCK_a for each body part. As shown in Table 2, HPE-Li demonstrates reliable overall pose estimation, achieving an average score of 85.12% at PCK_{50}. Notably, the accuracy of these particular body joints remains around 50% even under more stringent conditions at PCK_{20}. These insights stem from analyzing MM-Fi activities, focusing primarily on upper body movement, posing challenges, particularly for intense hand movements.

Table 2: The PCK of effectiveness for each body part on MM-Fi with **P3-S1**.

Keypoint	PCK_{20}	PCK_{30}	PCK_{40}	$PCK_{50}\uparrow$
Bot Torso	70.25	84.25	90.61	94.08
L.Hip	69.01	83.04	89.84	93.61
L.Knee	67.98	82.81	89.87	93.62
L.Foot	61.19	81.25	89.38	93.51
R.Hip	59.79	83.88	90.36	93.96
R.Knee	66.86	81.84	89.33	93.81
R.Foot	58.95	77.71	86.72	91.56
Center Torso	66.45	81.56	88.41	92.23
Upper Torso	54.52	75.86	85.21	90.09
Neck Base	46.41	69.25	81.61	87.44
Center Head	46.02	68.86	84.59	87.57
R.Shoulder	55.58	75.75	85.33	90.42
R.Elbow	34.67	54.41	67.56	76.44
R.Hand	5.08	14.34	29.59	46.83
L.Shoulder	55.27	75.42	85.13	90.37
L.Elbow	35.23	55.37	68.38	76.97
L.hand	5.03	14.12	30.21	48.97
Average	**52.07**	**68.22**	**78.18**	**85.12**

Table 3: The PCK effectiveness for each body part on WiPose.

Keypoint	PCK_5	PCK_{10}	PCK_{20}	$PCK_{50}\uparrow$
Nose	93.75	96.87	96.87	96.87
Neck	90.62	93.75	96.87	96.87
R.Shoulder	90.62	96.87	96.87	96.87
R.Elbow	68.75	90.62	96.87	96.87
R.Wrist	71.87	84.37	96.87	96.87
L.Shoulder	81.25	90.62	96.87	96.87
L.Elbow	46.87	71.87	81.25	96.87
L.Wrist	56.25	71.87	78.12	90.62
R.Hip	68.75	93.75	96.87	100.00
R.Knee	71.87	93.75	100.00	100.00
R.ankle	75.00	84.37	96.87	96.87
L.Hip	50.00	78.12	87.50	100.00
L.Knee	59.37	75.00	93.75	100.00
L.ankle	71.87	84.37	93.75	93.75
R.Eye	96.87	96.87	96.87	96.87
L.Eye	43.75	59.37	78.12	93.75
R.Ear	93.75	96.87	96.87	96.87
L.Ear	25.00	25.00	28.12	50.00
Average	**69.79**	**82.46**	**89.41**	**94.27**

Table 4: The PCK_a performance with different protocols and settings on the MM-Fi dataset: Best in **bold** and second best in underlined.

Protocol	Setting 1				Setting 2			
	PCK_{20}	PCK_{30}	PCK_{40}	$PCK_{50}\uparrow$	PCK_{20}	PCK_{30}	PCK_{40}	$PCK_{50}\uparrow$
1	51.08	68.12	78.12	84.41	36.42	56.91	70.51	79.59
2	40.59	61.27	74.89	83.59	34.17	55.07	70.23	80.34
3	**52.07**	**68.22**	**78.18**	**85.12**	**38.45**	**59.15**	**72.92**	**81.57**

Table 5: The MPJPE and PA-MPJPE results of HPE-Li with different protocols and settings on the MM-Fi dataset: Best in **bold** and second best in underlined.

P	Seting 1					Seting 2				
	Packet	Training	Testing	MPJPE↓	PA-MPJPE↓	Packet	Training	Testing	MPJPE↓	PA-MPJPE↓
1	166K	116K	50K	152.71	94.39	166K	133K	33K	189.16	93.22
2	154K	108K	46K	164.42	89.18	154K	123K	31K	190.98	**89.18**
3	320K	224K	96K	**149.43**	**92.52**	320K	256K	64K	**182.24**	93.16

In Tables 4 and 5, we further present comprehensive results, including PCK_a, MPJPE, and PA-MPJPE metrics across various protocols. Analyzing Table 5, our approach achieves optimal performance in **P3-S1** and weakest in **P2-S2**. HPE-Li demonstrates impressive outcomes in **P3-S1** with MPJPE at 149.43 mm, while exceeding by 89.18 mm in favorable PA-MPJPE results on **P2-S2**. These findings highlight the significant impact of protocol configurations on the HPE-Li's performance.

Results on WiPose: We proceed to evaluate the effectiveness of HPE-Li on WiPose, focusing on individual body parts. Results given in Table 3 demonstrate remarkable performance in overall pose estimation, with HPE-Li achieving notable average scores of 94.27% for PCK_{50}. Notably, even under stricter criteria such as PCK_5 and PCK_{10}, HPE-Li maintains commendable accuracy levels of

Table 6: Performance comparison between different schemes on both MM-Fi **P3-S1** and WiPose datasets (M: Million, G: Giga).

	MM-Fi						
	PCK_{20}	PCK_{30}	PCK_{40}	$PCK_{50}\uparrow$	MPJPE↓	PA-MPJPE↓	Params & FLOPs
Wi-Pose	48.55	65.06	75.58	82.441	158.21	97.72	5.34M & 84.31G
Wi-Mose	48.67	66.58	77.34	83.87	155.76	95.35	36.20M & 245.64G
WiLDAR	44.12	62.58	72.64	79.26	170.38	115.64	**1.63M** & 4.91G
WiSPPN	45.41	63.21	74.08	80.97	166.59	110.03	26.78M & 159.81G
PerUnet	50.12	67.34	77.59	83.56	154.66	98.67	34.51M & 168.52G
MetaFi++	45.46	64.44	75.13	81.75	164.45	106.31	26.42M & 507.89G
HPE-Li	**52.07**	**68.22**	**78.18**	**85.12**	**149.43**	**92.52**	1.66M & **2.42G**
	WiPose						
	PCK_5	PCK_{10}	PCK_{20}	$PCK_{50}\uparrow$	MPJPE↓	PA-MPJPE↓	Params & FLOPs
Wi-Pose	46.23	62.78	74.21	85.69	34.36	40.12	6.76M & 38.49G
Wi-Mose	54.65	66.74	77.12	88.54	26.48	31.19	35.75M & 116.47G
WiLDAR	36.26	54.38	72.16	84.32	55.63	62.60	**1.63M** & **4.90G**
WiSPPN	52.95	64.16	75.46	86.26	30.37	36.42	26.33M & 75.81G
PerUnet	63.07	71.77	79.50	88.74	17.12	22.64	33.85M & 167.51G
MetaFi++	53.64	66.72	76.68	88.62	28.62	33.72	25.58M & 502.32G
HPE-Li	**69.79**	**82.46**	**89.41**	**94.27**	**15.85**	**19.21**	3.49M & 5.18G

approximately 69.79% and 82.46%, respectively, underscoring its robust performance on specific body joints. This achievement surpasses that of MM-Fi, attributed to differences in raw signal acquisition methodologies between the two datasets. The WiPose system, with its simpler setup and lower-intensity human activities, yields more reliable accuracy in HPE tasks compared to MM-Fi.

Comparison with SOTA: Table 6 provides a comparison between HPE-Li and existing HPE approaches. Our method primarily focuses on relative pose accuracy, and it significantly outperforms the SOTA approaches for all evaluation criteria, particularly excelling in the low PCK_a as well as in MPJPE and PA-MPJPE metrics. These superior results are achieved with remarkably low computational costs. Notably, PerUnet also achieves good performance on both datasets, but with the cost of high complexity, (*e.g.* approximately 34M parameters). Conversely, our method on MM-Fi dataset has a similar level of complexity with WiLDAR, known as the most lightweight model in human recognition tasks (*e.g.* about 1.6M parameters). However, HPE-Li demonstrates superior performance, requiring only half the number of FLOPs compared to WiLDAR.

Qualitative Results: In Fig. 4, we present qualitative results to show visualizations of human skeletons in diverse environments from the MM-Fi dataset. Specifically, we focus on generating 2D poses and then transforming them into 3D poses by adding a constant vector as the third dimension. Our method consistently and reliably generates human poses across daily activities and rehabilitation exercises.

4.3 Ablation Study

To evaluate the role of each component in HPE-Li, we now conduct ablation experiments on the MM-Fi dataset.

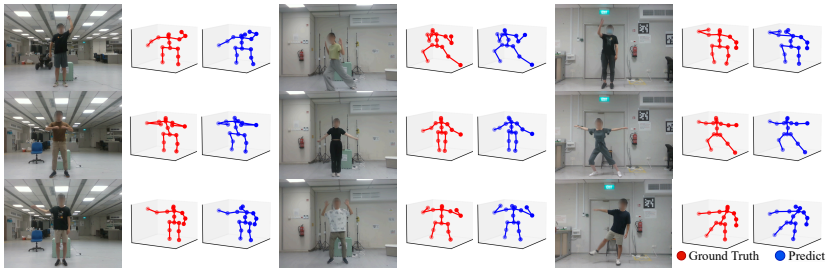


Fig. 4: Visualization of the human pose landmarks generated by the vision model (red) and WiFi model (blue) on the MM-Fi **P3-S1** dataset.

Table 7: Comparison of the DSKNet with other selective attention mechanisms on the MM-Fi **P3-S2** dataset.

General Models	MPJPE↓	Params	FLOPs
CNN+CNN	236.59	1.545M	2.256G
CNN+SENet	224.73	2.138M	20.434G
CNN+DConv	218.92	1.759M	4.277G
CNN+FDConv	<u>210.23</u>	1.814M	4.319G
CNN+SKConv	<u>210.45</u>	1.607M	2.376G
CNN+DSKConv	**182.62**	1.668M	2.427G

Table 8: Results of the DSKNet with different group numbers and dilation rate on the MM-Fi **P3-S1** dataset.

Model Setting	MPJPE↓	Params	FLOPs	Kernel
3×3,D=1,G=32	150.57	1.668M	2.426G	3×3
3×3,D=1,G=64	151.35	1.669M	2.427G	3×3
3×3,D=1,G=128	150.97	1.701M	2.429G	3×3
3×3,D=2,G=32	**149.43**	1.668M	2.426G	**5×5**
5×5,D=1,G=32	<u>149.71</u>	1.669M	2.429G	5×5
3×3,D=3,G=32	151.08	1.668M	2.426G	7×7

Impact of DSKConv: We conduct a comparative analysis using models that combine standard CNNs with CNNs incorporating attention mechanisms in cross-subject settings **P3-S2**. This analysis accounts for differences in frequency information among individuals in the same posture, aiming to demonstrate the ability of these variant CNNs to handle fluctuations in information frequency and signal strength. As seen from Table 7, HPE-Li encompasses various models, including the regular CNN, SENet [16], dynamic convolution (DConv) [7], frequency dynamic convolution (FDConv) [29] and traditional SK convolution (SKConv) [25], which server as benchmarks. SENet calculates kernel attention using global feature information through GAP and two FC layers with softmax activation. In contrast, DConv and FDConv compute attention across multiple kernels of the same size. Performance comparisons consistently demonstrate that SKConv outperforms SENet. However, SKConv exclusively evaluates attention in the channel domain, potentially leading to the loss of critical information. Consequently, DSKConv achieves superior performance compared to SKConv, surpassing 28 mm in the MPJPE metric. Furthermore, a comparison involving DSKConv, DConv and FDConv on MM-Fi reveals that DSKConv achieves more favorable results due to its adaptive receptive field size and diverse feature extraction, outperforming DConv by 36 mm and FDConv by 28 mm in the MPJPE metric. This confirms its ability to perceive contextual information in the frequency and channel domains, thereby facilitating accurate human pose generation.

Table 9: Performance of the DSKNet with different branches on the MM-Fi **P3-S1** dataset.

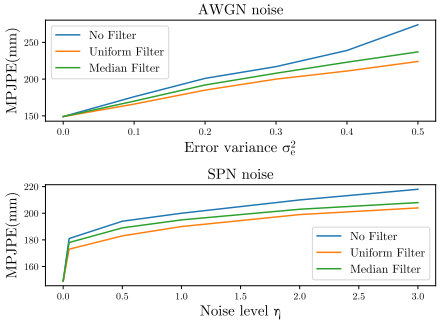
Model Settings	MPJPE↓	PCK_{20}↑	Params
K_1	152.67	50.28	1.62M
K_2	152.43	50.85	1.62M
K_3	152.89	50.15	1.62M
$K_1 + K_2$	149.49	51.82	1.63M
$K_1 + K_3$	149.54	51.76	1.63M
$K_2 + K_3$	149.56	51.42	1.63M
$K_1 + K_2 + K_3$	<u>149.48</u>	<u>51.92</u>	1.65M
$K_1 + K_2 + K_3 + K_4$	**149.43**	**52.07**	1.67M
$K_1 + K_2 + K_3 + K_4 + K_5$	150.21	51.54	1.69M

Table 10: Performance of the DSKNet with different reduction ratios on the MM-Fi **P3-S1** dataset.

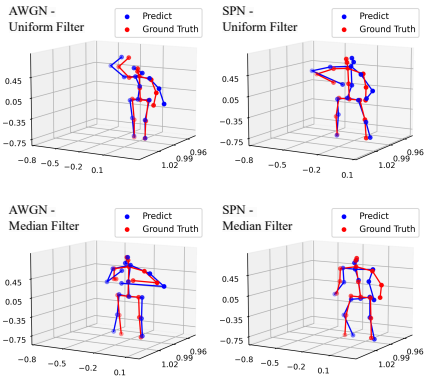
r	MPJPE↓	PCK_{20}↑	Params	FLOPs
1	151.85	51.62	1.676M	2.438G
4	151.65	51.74	1.674M	2.437G
8	151.03	51.85	1.673M	2.435G
16	<u>150.16</u>	<u>51.89</u>	1.671M	2.431G
32	**149.43**	**52.07**	1.668M	2.427G
64	150.32	41.96	1.664M	2.423G
96	151.68	51.73	1.663M	2.421G

Impact of Dilation and Group: We assess the impact of parameters in the DSKConv block by varying the dilation D and group G values. In Table 8, the term "kernel" denotes the approximate kernel size derived from dilated convolution. The best result, achieving an MPJPE metric of 149.43 mm, is obtained with 3×3 kernel, $D = 2$, and $G = 32$. The second-best result is 149.71 mm with 5×5 kernel, $D = 1$, and $G = 32$. These findings highlight the advantage of employing different kernel sizes to facilitate the aggregation of multiscale features. Furthermore, the results also suggest that utilizing the same kernel size in both branches may compromise outcomes. Comparing the two optimal configurations, *namely* the 5×5 kernel, $D = 1$ and the 3×3 kernel, $D = 2$, the latter exhibits slightly lower model complexity. Despite sharing the same receptive field, the smaller kernel with various dilations demonstrates significant performance and model complexity compared to the larger kernel without dilation.

Impact of Number Branches: We show the impact of the number of branches by incorporating two or more kernels (*e.g.* larger than 3×3). Due to limitations in the search space, we focus on a scenario with five branches, employing kernel sizes of K_1 (3×3), K_2 (5×5), K_3 (7×7), K_4 (9×9) and K_5 (11×11). Dilated convolution is employed for larger kernels, where G is set to 32. Table 9 illustrates that performance initially increases and then decreases with an increase in the number of branches, N. The one-branch case ($N = 1$) yields the poorest results. Through the utilization of multiple kernels, the spatial kernel achieves favorable outcomes by adaptively selecting among various branches. The optimal outcome is 149.43 mm with $N = 4$, while the second-best result is 149.48 mm associated with $N = 3$. The performance gap from $N = 3$ to $N = 4$ is negligible. When $N = 5$ the accuracy quickly decreases to 150.21 mm. To strike a balance between accuracy and performance, the case with $N = 3$ is preferable. Additionally, the dimension of the compact feature map (*i.e.* Eq. (2)) can be modified by r. Results in Table 10, with different r values, $N = 4$ and $G = 32$, show that accuracy does not monotonically decrease with an increase in r. The optimal performance is observed at $r = 32$ to balance between accuracy and overfitting due to channel interdependencies during training.



(a) The MPJPE accuracy under the impact of noises.



(b) Visualization of the human pose landmarks under the impact of noise (AWGN with σ_e^2=0.4 and SPN with $\eta = 2.5$).

Fig. 5: Impact of noise on the proposed HPE-Li model with MM-Fi **P3-S1** dataset.

4.4 Limitation

WiFi signals are susceptible to noise, interference, and multipath propagation, resulting in signal strength and quality fluctuations. To evaluate its impact, we consider two common types of noise: Additive white Gaussian noise (AWGN) characterized by a zero-mean and variance of σ_e^2 and Salt and Pepper Noise (SPN) with a noise level of η. These noises emulate the natural randomness encountered in real-world scenarios. To tackle this limitation, we integrate two denoising methods, *namely* Uniform and Median filters, into HPE-Li without changing the network architecture. The MPJPE accuracy of HPE-Li is depicted in Fig. 5a. As seen, the MPJPE accuracy is decreased when increasing σ_e^2 and η. However, the results remain reliable when employing filter methods, as illustrated in Fig. 5b. Future endeavors will focus on developing a robust model, possibly incorporating a stacked autoencoder [40].

5 Conclusions

We have introduced HPE-Li, a multi-model network designed to predict human pose landmarks by interpreting raw WiFi signals. Our method has attained state-of-the-art accuracy while upholding a lightweight model architecture, boasting significantly fewer parameters compared to its counterparts in the literature. Through validation on the MM-Fi and WiPose datasets, we have demonstrated the robustness and generalizability of HPE-Li across diverse and challenging scenarios. These achievements pave the way for the practical deployment of our application, aimed at enhancing the daily lives of individuals.

Acknowledgement. This work was supported in part by the VinUniversity Seed Grant Program.

References

1. Adib, F., Hsu, C.Y., Mao, H., Katabi, D., Durand, F.: Capturing the human figure through a wall. ACM Trans. Graph. **34**(6) (2015). https://doi.org/10.1145/2816795.2818072
2. Ahad, M.A.R., Antar, A.D., Ahmed, M.: Sensor-based human activity recognition: Challenges ahead. IoT Sensor-Based Activity Recognition (2020), https://api.semanticscholar.org/CorpusID:224963916
3. Aslani, S., Dayan, M., Storelli, L., Filippi, M., Murino, V., Rocca, M.A., Sona, D.: Multi-branch convolutional neural network for multiple sclerosis lesion segmentation. NeuroImage **196**, 1–15 (2019). https://doi.org/https://doi.org/10.1016/j.neuroimage.2019.03.068
4. Cao, Z., Hidalgo, G., Simon, T., Wei, S.E., Sheikh, Y.: Openpose: Realtime multi-person 2D pose estimation using part affinity fields. IEEE Transactions on Pattern Analysis and Machine Intelligence **43**, 172–186 (2018)
5. Che, R., Chen, H.: Channel state information based indoor fingerprinting localization. Sensors **23**(13) (2023)
6. Chen, X., Yuille, A.L.: Articulated pose estimation by a graphical model with image dependent pairwise relations (2014)
7. Chen, Y., Dai, X., Liu, M., Chen, D., Yuan, L., Liu, Z.: Dynamic convolution: Attention over convolution kernels. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 11027–11036 (2019), https://api.semanticscholar.org/CorpusID:208910380
8. Cohen, T.S., Welling, M.: Group equivariant convolutional networks. CoRR **abs/1602.07576** (2016)
9. Deng, F., Jovanov, E., Song, H., Shi, W., Zhang, Y., Xu, W.: Wildar: WiFi signal-based lightweight deep learning model for human activity recognition. IEEE Internet of Things Journal **11**(2), 2899–2908 (2024). https://doi.org/10.1109/JIOT.2023.3294004
10. Fan, X., Zheng, K., Lin, Y., Wang, S.: Combining local appearance and holistic view: Dual-source deep neural networks for human pose estimation. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1347–1355 (2015)
11. Fang, H.S., Xie, S., Tai, Y.W., Lu, C.: Rmpe: Regional multi-person pose estimation. In: 2017 IEEE International Conference on Computer Vision (ICCV). pp. 2353–2362 (2017)
12. Gkioxari, G., Hariharan, B., Girshick, R.B., Malik, J.: Using k-poselets for detecting people and localizing their keypoints. 2014 IEEE Conference on Computer Vision and Pattern Recognition pp. 3582–3589 (2014)
13. Glorot, X., Bordes, A., Bengio, Y.: Deep sparse rectifier neural networks. In: International Conference on Artificial Intelligence and Statistics (2011), https://api.semanticscholar.org/CorpusID:2239473
14. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask R-CNN. In: 2017 IEEE International Conference on Computer Vision (ICCV). pp. 2980–2988 (2017). https://doi.org/10.1109/ICCV.2017.322
15. He, K., Zhang, X., Ren, S., Sun, J.: Identity mappings in deep residual networks. In: Computer Vision – ECCV 2016. pp. 630–645. Springer International Publishing, Cham (2016)
16. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7132–7141 (2018). https://doi.org/10.1109/CVPR.2018.00745

17. Insafutdinov, E., Pishchulin, L., Andres, B., Andriluka, M., Schiele, B.: DeeperCut: A deeper, stronger, and faster multi-person pose estimation model. In: Computer Vision – ECCV 2016. pp. 34–50. Springer International Publishing (2016)
18. Ioffe, S., Szegedy, C.: Batch normalization: accelerating deep network training by reducing internal covariate shift. In: Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37. p. 448–456. ICML'15, JMLR.org (2015)
19. Jia, X., De Brabandere, B., Tuytelaars, T., Gool, L.V.: Dynamic filter networks. In: Advances in Neural Information Processing Systems. vol. 29. Curran Associates, Inc. (2016), https://proceedings.neurips.cc/paper_files/paper/2016/file/8bf1211fd4b7b94528899de0a43b9fb3-Paper.pdf
20. Jiang, W., Xue, H., Miao, C., Wang, S., Lin, S., Tian, C., Murali, S., Hu, H., Sun, Z., Su, L.: Towards 3D human pose construction using wifi. Proceedings of the 26th Annual International Conference on Mobile Computing and Networking (2020), https://api.semanticscholar.org/CorpusID:214796512
21. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems. vol. 25. Curran Associates, Inc. (2012), https://proceedings.neurips.cc/paper_files/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf
22. Kwapisz, J.R., Weiss, G.M., Moore, S.A.: Activity recognition using cell phone accelerometers. SIGKDD Explor. Newsl. **12**(2), 74–82 (2011). https://doi.org/10.1145/1964897.1964918, https://doi.org/10.1145/1964897.1964918
23. Li, T., An, C., Zhao, T., Campbell, A.T., Zhou, X.: Human sensing using visible light communication. Proceedings of the 21st Annual International Conference on Mobile Computing and Networking (2015), https://api.semanticscholar.org/CorpusID:7473648
24. Li, T., Liu, Q., Zhou, X.: Practical human sensing in the light. In: Proceedings of the 14th Annual International Conference on Mobile Systems, Applications, and Services. p. 71–84. MobiSys '16, Association for Computing Machinery, New York, USA (2016). https://doi.org/10.1145/2906388.2906401
25. Li, X., Wang, W., Hu, X., Yang, J.: Selective kernel networks. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 510–519 (2019)
26. Maturana, D., Scherer, S.: VoxNet: A 3D convolutional neural network for real-time object recognition. In: 2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 922–928 (2015). https://doi.org/10.1109/IROS.2015.7353481
27. Minh Dang, L., Min, K., Wang, H., Jalil Piran, M., Hee Lee, C., Moon, H.: Sensor-based and vision-based human activity recognition: A comprehensive survey. Pattern Recognition **108**, 107561 (2020). https://doi.org/https://doi.org/10.1016/j.patcog.2020.107561
28. Mirzaei, A., Pourahmadi, V., Soltani, M., Sheikhzadeh, H.: Deep feature selection using a teacher-student network. Neurocomputing **383**, 396–408 (2020)
29. Nam, H., Kim, S., Ko, B., Park, Y.: Frequency dynamic convolution: Frequency-adaptive pattern recognition for sound event detection. In: Interspeech 2022, 23rd Annual Conference of the International Speech Communication Association, Incheon, Korea, 18-22 September 2022. pp. 2763–2767. ISCA (2022). https://doi.org/10.21437/INTERSPEECH.2022-10127

30. Papandreou, G., Zhu, T., Kanazawa, N., Toshev, A., Tompson, J., Bregler, C., Murphy, K.: Towards accurate multi-person pose estimation in the wild. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3711–3719 (2017). https://doi.org/10.1109/CVPR.2017.395
31. Pishchulin, L., Insafutdinov, E., Tang, S., Andres, B., Andriluka, M., Gehler, P., Schiele, B.: Deepcut: Joint subset partition and labeling for multi person pose estimation. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) pp. 4929–4937 (2015), https://api.semanticscholar.org/CorpusID:5264846
32. Ren, Y., Wang, Z., Wang, Y., Tan, S., Chen, Y., Yang, J.: GoPose: 3D human pose estimation using WiFi. Association for Computing Machinery **6**(2) (2022)
33. Sigal, L., Balan, A., Black, M.J.: HumanEva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. International Journal of Computer Vision **87**(1), 4–27 (Mar 2010)
34. Srivastava, R.K., Greff, K., Schmidhuber, J.: Highway networks. CoRR **abs/1505.00387** (2015)
35. Sun, K., Xiao, B., Liu, D., Wang, J.: Deep high-resolution representation learning for human pose estimation. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 5686–5696 (2019), https://api.semanticscholar.org/CorpusID:67856425
36. Szegedy, C., Ioffe, S., Vanhoucke, V., Alemi, A.A.: Inception-v4, inception-resnet and the impact of residual connections on learning. In: Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence. p. 4278–4284. AAAI'17, AAAI Press (2017)
37. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S.E., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) pp. 1–9 (2014), https://api.semanticscholar.org/CorpusID:206592484
38. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) pp. 2818–2826 (2015), https://api.semanticscholar.org/CorpusID:206593880
39. Vaswani, A., Shazeer, N.M., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: Neural Information Processing Systems (2017), https://api.semanticscholar.org/CorpusID:13756489
40. Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y., Manzagol, P.A.: Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. J. Mach. Learn. Res. **11**, 3371–3408 (2010), https://api.semanticscholar.org/CorpusID:17804904
41. Wang, F., Panev, S., Dai, Z., Han, J., Huang, D.: Can WiFi estimate person pose? Clinical Orthopaedics and Related Research(CORR) **abs/1904.00277** (2019)
42. Wang, F., Zhou, S., Panev, S., Han, J., Huang, D.: Person-in-WiFi: Fine-grained person perception using WiFi. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 5451–5460 (2019). https://doi.org/10.1109/ICCV.2019.00555
43. Wang, W., Shen, J.: Deep visual attention prediction. IEEE Transactions on Image Processing **27**, 2368–2378 (2017), https://api.semanticscholar.org/CorpusID:3446654
44. Wang, Y., Guo, L., Lu, Z., Wen, X., Zhou, S., Meng, W.: From point to space: 3d moving human pose estimation using commodity WiFi. IEEE Communications Letters **25**(7), 2235–2239 (2021)

45. Wang, Z., Liu, Y., Liao, Q., Ye, H., Liu, M., Wang, L.: Characterization of a rs-lidar for 3d perception. In: 2018 IEEE 8th Annual International Conference on CYBER Technology in Automation, Control, and Intelligent Systems (CYBER). pp. 564–569 (2018). https://doi.org/10.1109/CYBER.2018.8688235
46. Wei, S., Ramakrishna, V., Kanade, T., Sheikh, Y.: Convolutional pose machines. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4724–4732. IEEE Computer Society (2016). https://doi.org/10.1109/CVPR.2016.511
47. Wei, Y., Xiao, H., Shi, H., Jie, Z., Feng, J., Huang, T.S.: Revisiting dilated convolution: A simple approach for weakly- and semi-supervised semantic segmentation. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition pp. 7268–7277 (2018), https://api.semanticscholar.org/CorpusID:44097132
48. Wu, F., Fan, A., Baevski, A., Dauphin, Y.N., Auli, M.: Pay less attention with lightweight and dynamic convolutions. CoRR **abs/1901.10430** (2019)
49. Wu, Z., Nagarajan, T., Kumar, A., Rennie, S., Davis, L.S., Grauman, K., Feris, R.S.: Blockdrop: Dynamic inference paths in residual networks. In: 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018. pp. 8817–8826. Computer Vision Foundation / IEEE Computer Society (2018). https://doi.org/10.1109/CVPR.2018.00919
50. Yang, J., Huang, H., Zhou, Y., Chen, X., Xu, Y., Yuan, S., Zou, H., Lu, C.X., Xie, L.: MM-fi: Multi-modal non-intrusive 4D human dataset for versatile wireless sensing. In: Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track (2023), https://openreview.net/forum?id=1uAsASS1th
51. Zhang, Z.: Microsoft kinect sensor and its effect. IEEE MultiMedia **19**(2), 4–10 (2012). https://doi.org/10.1109/MMUL.2012.24
52. Zhao, M., Li, T., Alsheikh, M.A., Tian, Y., Zhao, H., Torralba, A., Katabi, D.: Through-wall human pose estimation using radio signals. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition pp. 7356–7365 (2018)
53. Zhao, M., Liu, Y., Raghu, A., Zhao, H., Li, T., Torralba, A., Katabi, D.: Through-wall human mesh recovery using radio signals. 2019 IEEE/CVF International Conference on Computer Vision (ICCV) pp. 10112–10121 (2019)
54. Zhao, M., Tian, Y., Zhao, H., Alsheikh, M.A., Li, T., Hristov, R., Kabelac, Z., Katabi, D., Torralba, A.: RF-based 3D skeletons. Proceedings of the 2018 Conference of the ACM Special Interest Group on Data Communication (2018)
55. Zhou, Y., Zhu, A., Xu, C., Hu, F., Li, Y.: Perunet: Deep signal channel attention in unet for WiFi-based human pose estimation. IEEE Sensors Journal **22**(20), 19750–19760 (2022)
56. Zhou, Y., Huang, H., Yuan, S., Zou, H., Xie, L., Yang, J.: MetaFi++: WiFi-enabled transformer-based human pose estimation for metaverse avatar simulation. IEEE Internet of Things Journal **10**(16), 14128–14136 (2023)
57. Zou, H., Chen, C.L., Li, M., Yang, J., Zhou, Y., Xie, L., Spanos, C.J.: Adversarial learning-enabled automatic WiFi indoor radio map construction and adaptation with mobile robot. IEEE Internet of Things Journal **7**(8), 6946–6954 (2020). https://doi.org/10.1109/JIOT.2020.2979413