GRAPH-BASED 3D HUMAN POSE ESTIMATION USING WIFI SIGNALS

Jichao Chen[1]*, YangYang Qu*[2]*, Ruibo Tang*[3]*, Dirk Slock*[1]

[1] Communication Systems Department, EURECOM, France
[2] Digital Security Department, EURECOM, France
[3] RWTH Aachen University, Germany

{jichao.chen, yangyang.qu, dirk.slock}@eurecom.fr, ruibo.tang@rwth-aachen.de

ABSTRACT

WiFi-based human pose estimation (HPE) has attracted increasing attention due to its resilience to occlusion and privacy-preserving compared to camera-based methods. However, existing WiFi-based HPE approaches often employ regression networks that directly map WiFi channel state information (CSI) to 3D joint coordinates, ignoring the inherent topological relationships among human joints. In this paper, we present GraphPose-Fi, a graph-based framework that explicitly models skeletal topology for WiFi-based 3D HPE. Our framework comprises a convolutional neural network (CNN) encoder shared across antennas for subcarrier–time feature extraction, a lightweight attention module that adaptively reweights features over time and across antennas, and a graph-based regression head that combines graph convolutional network (GCN) layers with self-attention to capture local topology and global dependencies. Our proposed method achieves state-of-the-art performance on the MM-Fi dataset in various settings.

***Index Terms*—** Human pose estimation, WiFi sensing, graph convolutional networks

1. INTRODUCTION

Human pose estimation (HPE) is critical in a wide range of applications, including healthcare monitoring [1], augmented reality [2], and human-robot interaction [3]. This problem has been extensively studied using cameras, spanning 2D HPE [4], 3D HPE [5, 6], single-person pose estimation [7], and multi-person scenarios [8, 9]. Despite their progress, camera-based approaches face challenges in handling occlusions, low-light conditions, and privacy concerns. To address these limitations, radio-frequency (RF)-based sensing has emerged as a promising alternative, as RF signals can perceive human activities without direct visual input and are inherently robust to occlusions [10]. Among RF signals, WiFi is particularly appealing for HPE because it is widely available, cost-effective, and energy-efficient [11].

WiFi-based HPE leverages the variation in channel state information (CSI), which characterizes radio propagation in the surrounding environment, to infer human movements. Existing methods often treat CSI as images and apply deep regression networks to map WiFi signals into human pose coordinates. For example, WiSPPN [12] employs convolutional neural networks (CNNs) to estimate the 2D pose of a single person. WiPose [13] and GoPose [14] extend this idea to 3D pose estimation by using CNNs to extract CSI features and long short-term memory (LSTM) networks to capture temporal dynamics, thereby producing smoother pose predictions. More recently, transformer-based methods such as MetaFi++ [15] and Person-in-WiFi 3D [16] have been proposed to estimate 2D and 3D poses for both single- and multi-person scenarios. Moreover, lightweight architectures such as HPE-Li [17] have been introduced to reduce computational complexity by incorporating selective kernel attention modules.

However, directly applying CNNs or transformers to map WiFi CSI into pose coordinates ignores the inherent topological relationships among human body joints. The human body can be naturally represented as a graph, where joints correspond to nodes and bones to edges. In camera-based HPE, this structural prior has been effectively modeled using graph convolutional networks (GCNs) [5], showing strong performance compared to conventional approaches. Motivated by this success, we propose a graph-based architecture for 3D HPE using WiFi signals. Unlike GCN-based 2D-to-3D lifting methods in camera-based HPE [9], which rely on off-the-shelf 2D pose estimators to provide intermediate 2D keypoints, WiFi-based HPE lacks such explicit intermediate representations. To address this limitation, we design two modules to extract discriminative features from CSI and construct a latent graph representation. Specifically, we first employ a shared CNN encoder across all antennas to extract subcarrier–time features and map the subcarrier dimension to a latent space aligned with body joints. We then introduce a lightweight temporal–spatial attention module to adaptively select informative patterns along the compressed time and antenna dimensions. Finally, the resulting joint-level features are processed by a graph-based regression head, which combines GCN layers and self-attention [18] to capture both global dependencies and the topological structures of human joints.

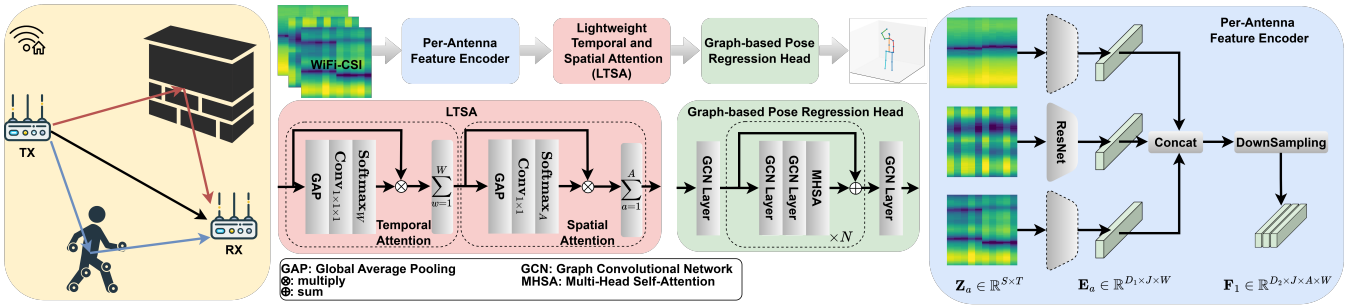

Fig. 1. The overview of our proposed GraphPose-Fi for WiFi-based 3D HPE. It consists of a per-antenna feature encoder, a lightweight temporal and spatial attention module, and a graph-based pose regression head.

In summary, our main contributions are threefold: (i) We propose a novel graph-based framework that explicitly models the topological relationships among human joints for WiFi-based 3D HPE. (ii) We design a feature extraction strategy consisting of a CNN encoder and a lightweight temporal–spatial attention module, which produces effective embeddings for the subsequent graph-based regression head. (iii) We conduct extensive experiments on the MM-Fi public dataset [19] and our method achieves state-of-the-art results across multiple metrics and settings. The source code is available at: `https://github.com/Cirrick/GraphPose-Fi`.

2. METHOD

2.1. Preliminary

Modern WiFi communication systems adopt orthogonal frequency division multiplexing (OFDM) [20], where the transmitted signals are divided into multiple subcarriers to improve spectral efficiency. Signals propagated from the transmitter (TX) to the receiver (RX) antennas experience multipath propagation caused by reflection, diffraction, and scattering in the surrounding environment [16]. These propagation effects are characterized in the CSI, which records the frequency response of each subcarrier. The movements of the human body alter the multipath components, thereby inducing distinctive variations in the CSI measurements. This is the underlying principle of WiFi-based HPE.

Given a sequence of CSI snapshots $\mathbf{H} \in \mathbb{C}^{N_r \times N_t \times S \times T}$, where N_r and N_t denote the number of receive and transmit antennas, respectively, S is the number of subcarriers, and T is the number of CSI samples within a sliding temporal window aligned with one pose label. The goal is to estimate the corresponding 3D human pose coordinates $\mathbf{Y} \in \mathbb{R}^{J\times 3}$, where J represents the number of body joints.

2.2. Proposed Method: GraphPose-Fi

We first apply standard pre-processing to convert the raw complex CSI into a real-valued tensor $\mathbf{Z} \in \mathbb{R}^{A\times S\times T}$, where $A = N_r \times N_t$ is the total number of antenna pairs. We then propose GraphPose-Fi, a graph-based framework for 3D HPE using WiFi CSI, as shown in Fig. 1. It consists of three modules: (1) Per-antenna feature encoder, (2) Lightweight temporal and spatial attention (LTSA) module, and (3) Graph-based pose regression head.

2.2.1. *Per-Antenna Feature Encoder*

Each antenna observes channel frequency responses under different multipath conditions, providing a complementary spatial view of human pose. Inspired by [15], we adopt a convolutional encoder based on ResNet [21] shared across antennas. For antenna $a \in \{1, \dots, A\}$, we treat its CSI slice $\mathbf{Z}_a \in \mathbb{R}^{S\times T}$ as a two-dimensional tensor and use the encoder to obtain the features

$$\mathbf{E}_a \;=\; f_\theta(\mathbf{Z}_a), \quad \mathbf{E}_a \in \mathbb{R}^{D_1\times J\times W}, \tag{1}$$

where f_θ denotes the encoder with parameters θ, D_1 represents the feature dimension, J represents the latent dimension aligned with the number of body joints to facilitate subsequent graph modeling, and W represents the downsampled temporal dimension. We then stack per-antenna features $\{\mathbf{E}_a\}_{a=1}^{A}$ and apply a point-wise convolution to reduce the feature dimension, obtaining $\mathbf{F}_1 \in \mathbb{R}^{D_2\times J\times A\times W}$.

2.2.2. *Lightweight Temporal and Spatial Attention*

Different time segments and antennas contribute unequally to pose estimation. We therefore design a lightweight attention module that adaptively reweights features along the temporal and spatial (antenna) dimensions, respectively.

Temporal Attention. For each joint j and antenna a, we compute attention weights over the compressed temporal index $w \in \{1, \dots, W\}$ by first averaging along the feature dimension D_2, then applying a $1\times1\times1$ point-wise convolution, followed by a softmax function along the temporal axis

$$\alpha_{j,a,w} \;=\; \mathrm{Softmax}_W\big(\mathrm{Conv}_{1\times1\times1}\big(\mathrm{Mean}_{D_2}(\mathbf{F}_1)\big)\big)\,. \tag{2}$$

These weights are used to temporally aggregate features

$$\mathbf{F}_t(:,j,a) \;=\; \sum_{w=1}^{W} \alpha_{j,a,w}\,\mathbf{F}_1(:,j,a,w), \quad \mathbf{F}_t \in \mathbb{R}^{D_2\times J\times A}. \tag{3}$$

Spatial Attention. Similarly, we derive attention weights across antennas for each joint by averaging over the feature dimension, and then applying a 1×1 point-wise convolution followed by a softmax function along the antenna axis

$$\beta_{j,a} \;=\; \mathrm{Softmax}_A\big(\mathrm{Conv}_{1\times1}\big(\mathrm{Mean}_{D_2}(\mathbf{F}_t)\big)\big)\,. \quad (4)$$

The features are then aggregated across antennas

$$\mathbf{F}_2(:,j) \;=\; \sum_{a=1}^{A} \beta_{j,a}\,\mathbf{F}_t(:,j,a), \quad \mathbf{F}_2 \in \mathbb{R}^{D_2\times J}. \quad (5)$$

Compared to using multi-head self-attention (MHSA) [18] for aggregating spatiotemporal information at each joint, our method is more lightweight while achieving comparable performance, as demonstrated in the ablation study (Sec. 3.3). Finally, we transpose $\mathbf{F}_2$ to a joint-first representation and apply layer normalization to form the input embedding $\mathbf{F}_3 \in \mathbb{R}^{J\times D_2}$ for the next stage.

2.2.3. Graph-based Pose Regression

While the previous stages extract informative features across subcarriers, time, and antennas, these features lack explicit modeling of the topological relationships among body joints. Motivated by prior graph-based HPE studies [5, 6], we adopt a GCN-based attention architecture that combines Chebyshev graph convolutions (ChebGConv) [22] with self-attention to model both local topology and global dependencies.

We first map the joint-wise features $\mathbf{F}_3$ to a latent embedding $\mathbf{X} \in \mathbb{R}^{J\times D_3}$ using a ChebGConv layer. Let $\mathbf{A}$ denote the skeleton adjacency matrix, $\mathbf{D}$ its degree matrix, $\mathbf{L} = \mathbf{I} - \mathbf{D}^{-1/2}\mathbf{A}\mathbf{D}^{-1/2}$ the normalized Laplacian and $\mathbf{I}$ the identity matrix. With the rescaled Laplacian $\tilde{\mathbf{L}} = 2\mathbf{L}/\lambda_{\max} - \mathbf{I}$, the Chebyshev convolution [22] updates features from layer l to $l+1$ as

$$\mathbf{X}^{\,l+1} \;=\; \sum_{k=0}^{K-1} \mathbf{T}_k(\tilde{\mathbf{L}})\,\mathbf{X}^{\,l}\,\mathbf{\Theta}_k, \quad (6)$$

where $\mathbf{T}_k(\tilde{\mathbf{L}}) = 2\tilde{\mathbf{L}}\mathbf{T}_{k-1}(\tilde{\mathbf{L}}) - \mathbf{T}_{k-2}(\tilde{\mathbf{L}})$ is the Chebyshev polynomial of degree k ($\mathbf{T}_0 = \mathbf{I}$, $\mathbf{T}_1 = \tilde{\mathbf{L}}$), $\lambda_{\max}$ is the largest eigenvalue of $\mathbf{L}$, and $\mathbf{\Theta}_k$ are learnable parameters. This expands the receptive field to K-hop neighbors.

Each GCN-based attention block contains two ChebGConv layers and an MHSA layer to aggregate global context across all joints. After N such blocks, the features are mapped by a final ChebGConv layer to the predicted pose $\hat{\mathbf{Y}} \in \mathbb{R}^{J\times3}$.

2.3. Training Objective

We train the networks using the mean squared error (MSE) between the predicted and ground-truth joint coordinates

$$\mathcal{L} \;=\; \frac{1}{J}\sum_{j=1}^{J} \left\|\hat{\mathbf{Y}}_j - \mathbf{Y}_j\right\|_2^2. \quad (7)$$

3. EXPERIMENTS

3.1. Experimental Setup

Dataset. We evaluate our approach on Protocol 1 (P1) of the MM-Fi dataset [19], which comprises 14 daily activities performed by 40 subjects across 4 environments. The WiFi system operates at 5 GHz with a 40 MHz bandwidth, consisting of one transmitter with a single antenna and one receiver with three antennas. CSI is collected over 114 subcarriers, and consecutive measurements are aggregated into a sample of size $3 \times 114 \times 10$. Each pose annotation contains 17 joints in 3D coordinates. To evaluate generalization, we follow three data split strategies: (i) Random split (S1), where data are randomly divided into training and testing sets with a 3:1 ratio; (ii) Cross-subject split (S2), where 32 subjects are used for training and the remaining 8 subjects for testing; and (iii) Cross-environment split (S3), where data from three environments are used for training and one for testing.

Evaluation Metrics. We adopt three commonly used metrics for 3D HPE [9, 23]. Mean per joint position error (MPJPE (mm)) measures the average Euclidean distance between predicted and ground-truth joint coordinates. Procrustes aligned MPJPE (PA-MPJPE (mm)) applies a similarity transformation (translation, rotation, and scaling) before computing the error, focusing on the structural accuracy of poses. Percentage of correct keypoints (PCK) reports the proportion of joints whose prediction error falls within a specified threshold relative to body size, reflecting the accuracy at the joint level.

Benchmarks. We compare the performance of our approach with several existing methods, including MetaFi++ [15] used as the baseline solution in MM-Fi dataset, HPE-Li [17], and DT-Pose [24], which are most recent and shown to achieve the best performance in MM-Fi dataset.

Implementation Details. We train our model using the AdamW optimizer [25] with an initial learning rate 3×10^{-4} and a weight decay 0.02. The learning rate follows a cosine decay over 50 epochs with a batch size of 256. All experiments are conducted on an NVIDIA GH200 GPU using PyTorch. Our model comprises 21.94 M parameters and 14.86 GFLOPs. It achieves an inference latency of 30.38 ms (32.9 FPS), demonstrating its suitability for real-time applications.

3.2. Comparison with State-of-the-art Methods

Table 1 summarizes the performance of different methods on MM-Fi dataset (Protocol 1). Our approach consistently outperforms existing baselines across all three data split settings. Under the random split (S1), it achieves the best PCK scores at all thresholds and reduces MPJPE to 160.6 mm, surpassing all the baselines. In the more challenging cross-subject split (S2), our method again achieves the highest accuracy on most metrics. For the cross-environment split (S3), where domain shift is most severe, our framework outperforms all compared methods by a clear margin across all metrics. These

Table 1. State-of-the-art performance comparisons on MM-Fi dataset (Protocol 1). ↑ higher is better; ↓ lower is better. Best values are bolded.

Method	PCK@10↑	20↑	30↑	40↑	50↑	MPJPE↓	PA-MPJPE↓
Setting 1 (Random Split):							
MetaFi++ [15]	24.6	56.2	72.9	82.4	88.1	174.2	112.4
HPE-Li [17]	26.4	56.4	72.6	82.0	87.8	172.6	**102.0**
DT-Pose [24]	26.3	57.2	74.1	83.7	88.9	168.0	102.4
Ours	**33.3**	**61.1**	**75.6**	**83.9**	**89.3**	**160.6**	105.0
Setting 2 (Cross-Subject):							
MetaFi++ [15]	9.5	39.9	64.0	78.6	**86.7**	215.7	118.2
HPE-Li [17]	11.1	40.4	62.6	75.9	84.3	221.4	**104.4**
DT-Pose [24]	11.6	40.2	62.1	76.1	84.8	221.1	105.8
Ours	**13.1**	**44.2**	**66.4**	**78.8**	86.3	**210.5**	105.5
Setting 3 (Cross-Environment):							
MetaFi++ [15]	1.7	10.5	25.6	44.0	60.3	329.9	108.3
HPE-Li [17]	0.5	5.8	18.1	35.6	52.3	361.1	104.4
DT-Pose [24]	0.8	7.9	23.7	43.5	61.0	326.9	104.7
Ours	**2.7**	**12.9**	**29.2**	**49.6**	**67.2**	**302.7**	**103.0**

Table 2. Per-joint MPJPE (mm) ↓ comparisons on MM-Fi dataset (P1–S1).

Joint	MetaFi++ [15]	HPE-Li [17]	DT-Pose [24]	**Ours**
Bot Torso	107.8	108.3	105.8	**93.9**
L.Hip	111.7	111.6	109.5	**100.2**
L.Knee	111.0	111.1	111.1	**99.7**
L.Foot	109.0	109.6	110.3	**102.3**
R.Hip	111.8	114.7	112.5	**101.2**
R.Knee	109.4	113.1	113.6	**99.6**
R.Foot	114.2	116.2	119.4	**107.0**
Center Torso	115.8	117.7	116.3	**100.4**
Upper Torso	142.5	142.0	141.5	**123.7**
Neck Base	168.2	165.6	163.7	**150.9**
Center Head	169.5	166.0	167.6	**150.8**
R.Shoulder	155.8	153.4	153.4	**141.0**
R.Elbow	264.4	260.7	**246.0**	251.7
R.Hand	382.8	381.2	**359.7**	360.4
L.Shoulder	155.7	148.3	149.4	**137.4**
L.Elbow	255.4	244.0	**230.4**	244.6
L.Hand	376.6	371.2	**346.4**	365.2
Average	174.2	172.6	168.0	**160.6**

results demonstrate both the effectiveness and robustness of the proposed method across different experimental settings.

Table 2 further reports per-joint MPJPE on MM-Fi dataset (P1–S1). Our approach achieves significant improvements on most joints, particularly in torso and head regions (e.g., bot torso, upper torso, neck base), where errors are reduced by 10–20 mm compared to the baselines. This demonstrates the effectiveness of GCN-based attention in modeling global skeletal topology. Nevertheless, all methods still exhibit higher errors on hands and elbows, which is likely due to the limited spatial resolution of WiFi signals [24].

3.3. Ablation Study

Impact of LTSA Module. We investigate different strategies to aggregate the temporal and spatial information while fixing the encoder and the graph regression head. Specifically, we compare: (i) Global average pooling (GAP) over the temporal and antenna dimensions, and (ii) Per-joint multi-head self-attention (PJ-MHSA): we treat the antenna-time dimensions as tokens with features at the first dimension as embeddings. For each joint, MHSA is applied to capture intra-temporal and inter-antenna dependencies, followed by token pooling to obtain the joint embeddings. Results in Table 3 show that our proposed LTSA module outperforms the simple GAP baseline and achieves slightly better accuracy than PJ-MHSA while requiring substantially lower computational cost.

Table 3. Ablation study for LTSA module design.

Method	MPJPE (mm)
GAP	161.6
PJ-MHSA	160.8
LTSA (ours)	**160.6**

Table 4. Ablation study for graph-structure design.

Method	MPJPE (mm)
MLP regression head	167.8
Graph-based regression head (N = 2)	161.8
Graph-based regression head (N = 4)	**160.6**
Graph-based regression head (N = 6)	161.5

Graph-Structure Design. To demonstrate the effectiveness of the graph-based method and test its configuration, we replace the graph regression head with a multi-layer perceptron (MLP) and vary the number of GCN-based attention blocks (N from 2 to 6), while keeping the encoder and LTSA module fixed. Results in Table 4 show that our graph-based architecture outperforms the MLP baseline. Moreover, performance improves when increasing the number of blocks from 2 to 4, but further stacking yield no additional gains. Consequently, we adopt 4 GCN-based attention blocks in our model, which are effective and efficient to model the human joint topology.

4. CONCLUSION

We propose GraphPose-Fi, a graph-based framework for WiFi-based 3D HPE. The model couples a CNN encoder shared by antennas for subcarrier–time feature extraction with a lightweight temporal–spatial attention module, and a GCN-based attention regression head that encodes skeletal topology while capturing global dependencies. Experimental results on the MM-Fi dataset demonstrate that GraphPose-Fi outperforms existing state-of-the-art methods across various settings and metrics. While WiFi sensing enhances privacy by avoiding direct imagery, its deployment must ensure user consent and secure data management. Future work will explore multi-sensor fusion and enhance domain generalization across diverse environments.

Acknowledgements. EURECOM's research is partially supported by its industrial members: ORANGE, BMW, SAP, iABG, Norton LifeLock, by the Franco-German project 5GOPERA (BPI), the French project YACARI (PEPR-5G), the EU INFRA project CONVERGE, and by a Huawei France funded Chair towards Future Wireless Networks.

5. REFERENCES

[1] Tian He, Yang Chen, Ling Wang, and Hong Cheng, "An expert-knowledge-based graph convolutional network for skeleton-based physical rehabilitation exercises assessment," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 32, pp. 1916–1925, 2024.

[2] Haotian Zhang, Cristobal Sciutto, Maneesh Agrawala, and Kayvon Fatahalian, "Vid2player: Controllable video sprites that behave and appear like professional tennis players," *ACM Transactions on Graphics (TOG)*, vol. 40, no. 3, pp. 1–16, 2021.

[3] Qing Gao, Zhaojie Ju, Yongquan Chen, Qiwen Wang, Yinan Zhao, and Shiwu Lai, "Parallel dual-hand detection by using hand and body features for robot teleoperation," *IEEE Transactions on Human-Machine Systems*, vol. 53, no. 2, pp. 417–426, 2023.

[4] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh, "Real-time multi-person 2d pose estimation using part affinity fields," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 7291–7299.

[5] Weixi Zhao, Weiqiang Wang, and Yunjie Tian, "Graformer: Graph-oriented transformer for 3d pose estimation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 20438–20447.

[6] Jia Gong, Lin Geng Foo, Zhipeng Fan, Qiuhong Ke, Hossein Rahmani, and Jun Liu, "Diffpose: Toward more reliable 3d pose estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 13041–13051.

[7] Wenhao Li, Hong Liu, Hao Tang, Pichao Wang, and Luc Van Gool, "Mhformer: Multi-hypothesis transformer for 3d human pose estimation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 13147–13156.

[8] Jianfeng Zhang, Yujun Cai, Shuicheng Yan, Jiashi Feng, et al., "Direct multi-view multi-person 3d pose estimation," *Advances in Neural Information Processing Systems*, vol. 34, pp. 13153–13164, 2021.

[9] Ce Zheng, Wenhan Wu, Chen Chen, Taojiannan Yang, Sijie Zhu, Ju Shen, Nasser Kehtarnavaz, and Mubarak Shah, "Deep learning-based human pose estimation: A survey," *ACM computing surveys*, vol. 56, no. 1, pp. 1–37, 2023.

[10] Yongsen Ma, Gang Zhou, and Shuangquan Wang, "Wifi sensing with channel state information: A survey," *ACM Computing Surveys (CSUR)*, vol. 52, no. 3, pp. 1–36, 2019.

[11] Iftikhar Ahmad, Arif Ullah, and Wooyeol Choi, "Wifi-based human sensing with deep learning: Recent advances, challenges, and opportunities," *IEEE Open Journal of the Communications Society*, vol. 5, pp. 3595–3623, 2024.

[12] Fei Wang, Stanislav Panev, Ziyi Dai, Jinsong Han, and Dong Huang, "Can wifi estimate person pose?," *arXiv preprint arXiv:1904.00277*, 2019.

[13] Wenjun Jiang, Hongfei Xue, Chenglin Miao, Shiyang Wang, Sen Lin, Chong Tian, Srinivasan Murali, Haochen Hu, Zhi Sun, and Lu 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.

[14] Yili Ren, Zi Wang, Yichao Wang, Sheng Tan, Yingying Chen, and Jie Yang, "Gopose: 3d human pose estimation using wifi," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 6, no. 2, pp. 1–25, 2022.

[15] Yunjiao Zhou, He Huang, Shenghai Yuan, Han Zou, Lihua Xie, and Jianfei Yang, "Metafi++: Wifi-enabled transformer-based human pose estimation for metaverse avatar simulation," *IEEE Internet of Things Journal*, vol. 10, no. 16, pp. 14128–14136, 2023.

[16] Kangwei Yan, Fei Wang, Bo Qian, Han Ding, Jinsong Han, and Xing Wei, "Person-in-wifi 3d: End-to-end multi-person 3d pose estimation with wi-fi," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 969–978.

[17] Toan D. Gian, Tien Dac Lai, Thien Van Luong, Kok-Seng Wong, and Van-Dinh Nguyen, "Hpe-li: Wifi-enabled lightweight dual selective kernel convolution for human pose estimation," in *European Conference on Computer Vision*. Springer, 2024, pp. 93–111.

[18] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[19] Jianfei Yang, He Huang, Yunjiao Zhou, Xinyan Chen, Yuecong Xu, Shenghai Yuan, Han Zou, Chris Xiaoxuan Lu, and Lihua Xie, "Mm-fi: Multi-modal non-intrusive 4d human dataset for versatile wireless sensing," *Advances in Neural Information Processing Systems*, vol. 36, pp. 18756–18768, 2023.

[20] "Ieee standard for information technology– local and metropolitan area networks– specific requirements– part 11: Wireless lan medium access control (mac)and physical layer (phy) specifications amendment 5: Enhancements for higher throughput," *IEEE Std 802.11n-2009 (Amendment to IEEE Std 802.11-2007 as amended by IEEE Std 802.11k-2008, IEEE Std 802.11r-2008, IEEE Std 802.11y-2008, and IEEE Std 802.11w-2009)*, pp. 1–565, 2009.

[21] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[22] Michaël Defferrard, Xavier Bresson, and Pierre Vandergheynst, "Convolutional neural networks on graphs with fast localized spectral filtering," *Advances in neural information processing systems*, vol. 29, 2016.

[23] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu, "Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments," *IEEE transactions on pattern analysis and machine intelligence*, vol. 36, no. 7, pp. 1325–1339, 2013.

[24] Yang Chen, Jingcai Guo, Song Guo, Jingren Zhou, and Dacheng Tao, "Towards robust and realistic human pose estimation via wifi signals," *arXiv preprint arXiv:2501.09411*, 2025.

[25] Ilya Loshchilov and Frank Hutter, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017.