Scilight

Edge Intelligence and Systems
https://www.sciltp.com/journals/EIS

Article

DT-Pose: Towards Robust and Realistic Human Pose Estimation Using WiFi Signals

Yang Chen [1] and Jingcai Guo [1,2,*]

[1] Department of Computing, Hong Kong Polytechnic University, HongKong SAR, China
[2] Department of Land Surveying and Geo-Informatics, Hong Kong Polytechnic University, HongKong SAR, China
* Correspondence: jc-jingcai.guo@polyu.edu.hk

How To Cite: Chen, Y.; Guo, J. DT-Pose: Towards Robust and Realistic Human Pose Estimation Using WiFi Signals. *Edge Intelligence and Systems* **2026**, *1*(1), 2.

Received: 4 April 2026
Revised: 22 May 2026
Accepted: 4 June 2026
Published: 17 June 2026

Abstract: Robust WiFi-based human pose estimation (HPE) is a challenging task that bridges discrete and subtle WiFi signals to human skeletons. We revisit this problem and reveal two critical yet overlooked issues: (1) cross-domain gap, i.e., due to significant discrepancies in pose distributions between source and target domains; and (2) structural fidelity gap, i.e., predicted skeletal poses manifest distorted topology, usually with misplaced joints and disproportionate bone lengths. This paper fills these gaps by reformulating the task into a novel two-phase framework dubbed **DT-Pose**: **D**omain-consistent representation learning and **T**opology-constrained **Pose** decoding. Concretely, we first propose a temporal consistency contrastive learning strategy with uniformity regularization, integrated into a self-supervised masked pretraining paradigm. This design facilitates robust learning of domain-consistent and sequence-level motion-discriminative WiFi representations while mitigating potential mode collapse caused by signal sparsity. Beyond this, we introduce an effective hybrid decoding architecture that incorporates explicit skeletal topology constraints. By compensating for the inherent absence of spatial priors in WiFi semantic vectors, the decoder enables structured modeling of both adjacent and overarching joint relationships, producing more realistic pose predictions. Extensive experiments conducted on various benchmark datasets highlight the superior performance of our method in tackling these fundamental challenges in 2D/3D WiFi-based HPE tasks.

Keywords: human pose estimation; cross-domain gap; WiFi

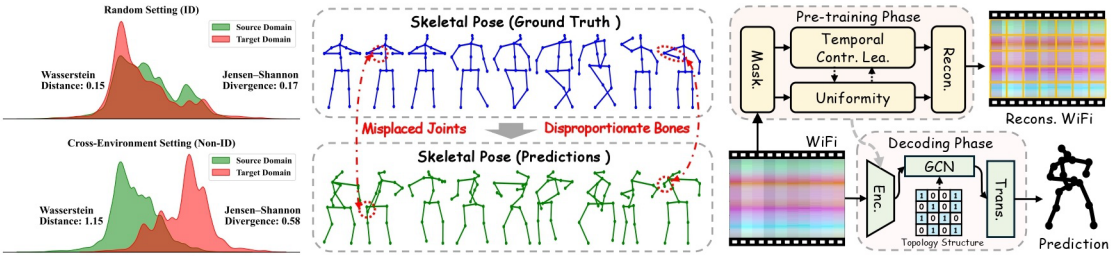
1. Introduction

Image-based human pose estimation (HPE), a highly active and hot topic, has recently achieved remarkable success in both 2D [1,2] and 3D scenarios [3,4], spanning single-person [5] and multi-person settings [6,7]. These advancements have significantly propelled broad applications in virtual reality [8], autonomous driving [9], and the healthcare community [10]. However, those visual-based methods face inherent limitations due to realistic challenges (e.g., lighting intensity, view variations, and occlusions). Furthermore, rising concerns regarding privacy have driven the growing research attention toward non-visual modalities (e.g., WiFi [11,12], RF [13], and wearable sensor [14] signals), which offer significant advantages in privacy protection and resilience to occlusions. Among these, the WiFi modality holds promise due to its widespread deployability and compatibility with Edge AI in the AIoT era.

Tracing the development of the WiFi-based HPE, the field has gradually progressed from single-person 2D to more complex multi-person 3D HPE [11,12,15–22]. Predominantly, these methods rely on supervised learning and focus on designing complex regression networks to map WiFi signals to 2D/3D pose coordinates. However, all of them assume that the training and testing data follow the same distribution, which does not hold in real-

world scenarios due to domain variability. Fortunately, the recently released WiFi dataset (MM-Fi [23]) incorporates cross-domain settings, presenting new challenges for evaluating the generalizability of WiFi-based HPE methods.

Upon analyzing WiFi signals across diverse domains, we observed significant discrepancies in pose coordinate distributions between the source and target domains under cross-environment settings, i.e., ***cross-domain gap***, contrasting with the commonly held assumption of identical distributions (Figure 1a). In such scenarios, existing supervised methods tend to overfit source-domain pose distributions and generalize poorly to target domains. This limitation underscores the inadequacy of supervised learning in capturing intrinsic motion patterns embedded in sparse WiFi signals, leading to the learning of spurious, motion-irrelevant, and noisy features. While AdaPose [24] also recognizes this challenge, its reliance on pre-acquired target domain data for domain adaptation renders it impractical and suboptimal. Thus, we aim to design a WiFi-specific approach that learns domain-consistent and sequence-level motion-discriminative WiFi representations independent of pose coordinates, thereby enhancing cross-domain transferability.



(a) Pose Domain Gap (b) Distorted Predictions (c) Two-Stage Process

Figure 1. (**a**) shows the pose coordinates distribution between the source and target domains. (**b**) represents the predictions of the MetaFi++ method [22] and corresponding ground truth. (**c**) denotes the overview framework of our method.

In addition to domain generalization challenges, we observe that existing methods often produce pose predictions with unrealistic topologies (e.g., misplaced joints and disproportionate bone lengths), resulting in a ***structural fidelity gap*** (Figure 1b). These deficiencies stem from two key factors: (1) prior works typically adopt CNNs combined with MLPs to regress poses in an unconstrained manner, leading to poor modeling of human joint relationships; and (2) unlike the image modality, which provides explicit spatial priors (e.g., human heatmaps), the WiFi modality offers only high-level semantic representations (e.g., global vectors) that lack spatial topological information, making it inherently more difficult to capture valid pose structures. To mitigate these issues, we propose incorporating explicit skeletal topology priors as constraints to better model the non-trivial spatial relationships among human joints.

Building on the above observations, we propose a novel framework, which reformulates WiFi-based HPE as a two-phase process: ***D***omain-consistent WiFi representation learning and ***T***opology-constrained ***Pose*** decoding, as depicted in Figure 1c. In the first phase, we transform raw WiFi signals into image-like inputs and adopt the self-supervised masked prediction strategy of MAE [25] as the main line to learn domain-consistent representations. Considering the temporal continuity of WiFi signals, we treat adjacent WiFi frames within an action sequence as positive pairs and others in the batch as negatives, yielding sequence-level motion-discriminative representations through contrastive objectives. Additionally, uniformity regularization is employed to mitigate potential representational collapse caused by signal sparsity. In the second phase, the pre-trained encoder is frozen to extract domain-consistent WiFi representations. We then introduce task prompts and Graph Convolution layers with spatial topology priors as constraints, enabling localized modeling of adjacent joint relationships. Concurrently, we establish more holistic dependencies among overarching joints within Transformer layers. By exploring these adjacent and overarching spatial correlations, our decoding architecture promotes realistic pose predictions.

The main contributions can be summarized as follows:

(1) We reveal the cross-modal gap issue in WiFi-based HPE and develop a tailored WiFi representation learning method that integrates a temporal consistency contrastive strategy with uniformity regularization, enabling the extraction of domain-consistent and sequence-level motion-discriminative features from sparse signals.

(2) We reveal the structural fidelity gap issue in WiFi-based HPE and propose a hybrid decoding architecture with explicitly incorporated skeletal topology priors as constraints, compensating for the lack of spatial cues in WiFi and enabling effective modeling of joint relationships.
(3) We evaluate the effectiveness of our method through extensive and comprehensive experiments on three mainstream datasets, demonstrating its superior performance in the WiFi-based HPE field.

2. Related Work

2.1. WiFi-Based Human Pose Estimation

WiFi-based HPE is an emerging research topic that has gradually flourished in recent years, encompassing a range of tasks from single-person 2D [12,15] and 3D [17,18,21,22] to multi-person 2D [16] and 3D scenarios [11]. Early work in this field, such as WiSPPN [15,16], pioneers 2D HPE by employing fundamental CNN models [26]. Subsequently, WiPose [17] extends to 3D poses through a combination of CNN and RNN layers, thereby leveraging temporal dynamics to yield smoother skeletal predictions. Differently, both GoPose [18] and Winect [21] methods leverage the 2D angle-of-arrival features of WiFi signals to estimate 3D poses. Recently, such as MetaFi++ [22] and Person-in-WiFi-3D [11], employ the Transformer layers to learn WiFi representations for single-/multi-person 3D HPE. Concurrently, HPE-Li [12] has designed dynamic CNN kernels to predict poses more efficiently. Unfortunately, all these studies overlook cross-domain gaps and rely on pose supervision in the source domain to guide WiFi representation learning from sparse signals, which hinders generalization to target domains with different pose distributions. Even more fundamentally, WiFi signals lack explicit human spatial priors by nature compared to images, making it challenging to perceive body topology when using the abovementioned CNN architectures directly. [**Summary**]: In this work, we are the first to tackle the cross-domain gap challenge by introducing a self-supervised pretraining strategy with WiFi-specific designs, tailored to the sparse and continuous nature of WiFi signals. Simultaneously, we explicitly capture both adjacent and overarching spatial correlations between joints by compensating skeletal topology priors into a hybrid decoding architecture, thereby ensuring structurally faithful pose predictions.

2.2. Masked Pre-Training

Masked pretraining techniques have been widely studied across various data modalities for self-supervised representation learning, leveraging the reconstruction of masked inputs as a core strategy [25,27–34]. Among these modalities, the BERT [27] and GPT [28] are two seminal language models that pioneered the masked modeling paradigm by predicting masked word tokens based on context information. Inspired by them, the computer vision community introduced masked pertaining frameworks for images, giving rise to representative methods like BEiT [29] and MAE [25], while the video community [30,31] subsequently demonstrated that the masked mechanism extends effectively into the temporal dimension. Beyond these, other data modalities, including audio (Audio-MAE [32]), skeleton (SkeletonMAE [33]), and time series (TimeMAE [34]), have similarly validated the feasibility and efficacy of masked modeling for task-agnostic representation learning in a self-supervised manner. [**Summary**]: To the best of our knowledge, this work is the first to employ a self-supervised masked pre-training paradigm in the WiFi modality. Moreover, we incorporate the temporal-consistent contrastive strategy with uniformity regularization to extract sequence-level motion-discriminative representations from sparse and continuous WiFi signals, thereby preserving motion semantic consistency while mitigating potential mode collapse.

2.3. Skeleton-Based Action Recognition

Learning skeleton action representation can be conceptualized as the inverse process of human pose decoding. Typically, skeleton-based action recognition methods can be categorized into CNN-based, GCN-based, and Transformer-based [35–40]. CNN-based methods transform skeleton sequences into image-like formats to extract sequence-level discriminative representations [35,41]. In contrast, GCN-based methods model human joints and bones as graph nodes and edges, explicitly incorporating a learnable adjacent matrix to explore spatial-temporal features, thus improving performance by a large margin [37–39,42]. More recently, Transformer-based methods leverage self-attention mechanisms to capture long-range dependencies among joints [43–45]. [**Summary**]: Inspired by the success of GCNs and Transformers in learning representations directly from the skeleton data, we combine their strengths in a reverse decoding manner to regress poses from WiFi-based representations. Although WiFi semantic vectors inherently lack human skeletal topology priors, our method compensates for this by incorporating skeletal topology, enabling more realistic and faithful pose predictions.

3. Method

3.1. Preliminary

The overall framework of the proposed method is shown in Figure 2. Typically, WiFi signals are captured using multiple transmitters and receivers, with each signal comprising multiple subcarriers operating in orthogonal frequency bands to facilitate inter-device communication. These subcarriers describe the signal propagation process, known technically as Channel State Information (CSI). As shown in Figure 3a, the CSI undergoes various distortions attributed to multipath effects and physical transformations, e.g., reflections, diffraction, and scattering [11,22]. Leveraging these properties, we can record time-continuous WiFi signals that are dynamically influenced by human activities, i.e., action movements, thereby enabling the estimation of corresponding human poses. More specifically, one WiFi sample can be represented as $X \in \mathbb{R}^{E\times R\times A\times S\times T}$, where E, R, A, S denote the numbers of transmitters, receivers, antennas, and subcarriers, respectively. Here, $T = \frac{f_{\text{wifi}}}{f_{\text{video}}}$ represents the temporal resolution, equal to the ratio of the WiFi sampling frequency to that of the corresponding video action sequence. Notably, increasing the number of subcarriers and antennas enhances the resolution of the WiFi signals, capturing more subtle movements and finer variations. We define the ground truth of 3D pose coordinates for each frame as $Y \in \mathbb{R}^{M\times J\times C}$, where M represents the number of humans, J indicates the number of joints, and C specifies the spatial dimensions (coordinates). Hence, the entire dataset can be formalized as $\mathcal{D} = \{X_i \in \mathbb{R}^{E\times R\times A\times S\times T}, Y_i \in \mathbb{R}^{M\times J\times C}\}_{i=1}^{N}$, where N is the total number of samples.

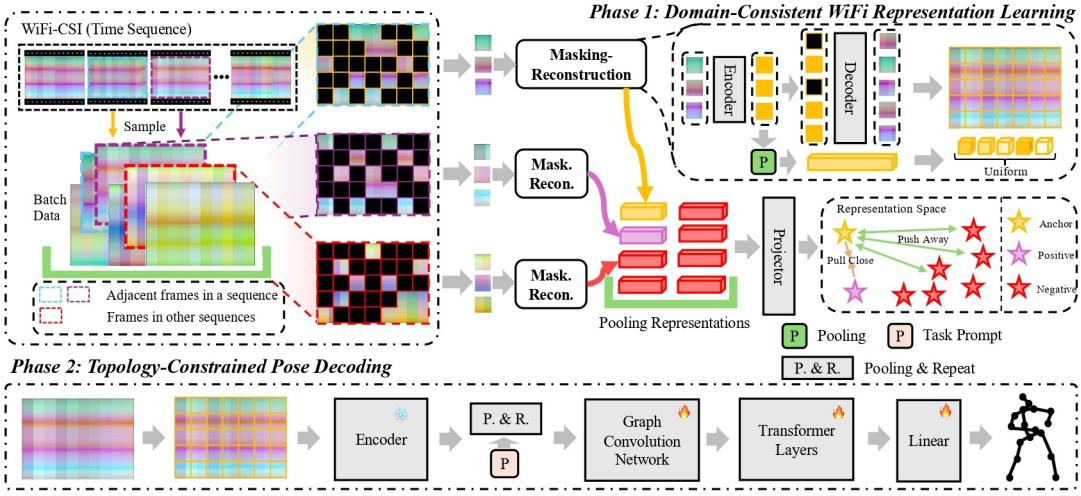
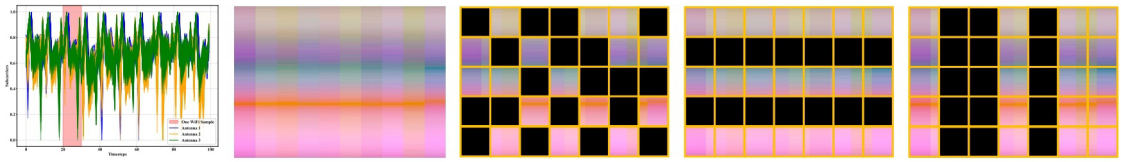


Figure 2. The pipeline of our method, including the pre-training and pose decoding phases.



(a) Signals (b) Image-like (c) Unstructured (d) Channel (e) Time

Figure 3. Original WiFi CSI signals and different masking strategies on the MM-Fi dataset.

3.2. Domain-Consistent Representation Learning

3.2.1. Masked Operation

Unlike prior supervised WiFi-based HPE methods that tend to learn pose-coordinate-dependent representations, we instead adopt a self-supervised MAE strategy [25] to learn WiFi representations without relying on pose annotations, thereby improving their transferability across different domains. To more closely align the WiFi modality with the image-based framework employed in MAE [25], we first reshape each WiFi sample into an image-like form $\hat{X}_i \in \mathbb{R}^{A\times ERS\times T}$, as illustrated in Figure 3b. Concretely, we treat the antenna dimension A as the image channels, the concatenated subcarriers from all devices (ERS) to image height, and the temporal resolution T as image width. Subsequently, to investigate the best suitable masking strategy tailored to WiFi signals, we consider three distinct approaches at the pre-training stage, including unstructured (i.e., random

masking with grid shape), channel-structured (i.e., random masking subcarriers along the time-frequency axis), time-structured (i.e., random masking timesteps along the subcarrier axis), as shown in Figure 3c–e. Following MAE [25], we divide $\widehat{\boldsymbol{X}}_i$ into non-overlapped regular grid patches and employ convolution layers to embed each patch, obtaining $\tilde{X}_i \in \mathbb{R}^{\mathbb{m}\times\mathbb{d}}$, where n is the patch numbers and d is the embedding dimension. We then incorporate fixed sinusoidal positional embeddings into these embedded patches and apply random masking with a high ratio (80% in our experiments) to enforce robust representation learning. The encoder, comprising 4 stacked Transformer layers, is tasked with learning domain-consistent WiFi representations, while 2 Transformer layers in the decoder strive to reconstruct the original WiFi $\hat{X}_i$, ultimately producing a reconstruction X_i'. The entire procedure is optimized by minimizing MSE between the reconstruction and the original input as follows:

$$\mathcal{L}_{\text{Mask}} = ||X_i' - \hat{X}_i||_2^2. \tag{1}$$

3.2.2. Temporal Consistency Strategy

WiFi signals corresponding to the same action may vary across domains due to differences in signal strength and surrounding layouts. However, the temporal dynamics induced by human motion remain relatively consistent. Since the masked operation applied to individual WiFi samples mainly captures modality-level representations and does not explicitly exploit motion-related temporal patterns, we further introduce a temporal consistency strategy to strengthen the learning of stable and motion-dynamic features that generalize across domains. Firstly, we treat adjacent WiFi frames within the same action sequence as a positive pair due to motion consistency within them. Notice that the WiFi samples in a batch include one-pair adjacent WiFi frames $\left(\hat{X}_t, \hat{X}_{t+1}\right)$ from the same sequence and other non-isomorphic WiFi frames $\{\hat{X}_i\}_{i=1}^{\mathrm{B}-2}$ from all action sequences, where B is the batch size. Thus, other combinations of WiFi samples in a batch should be negative pairs. Following the masked procedure, we pool the encoded visible embedding patches of each sample to derive a representation $e_i \in \mathbb{R}^d$. Next, we project them to obtain positive pair representations $(\mathrm{s}_t, \mathrm{s}_{t+1})$ and other sample representations $\{\mathrm{v}_i\}_{i=1}^{\mathrm{B}-2}$. Then, we pull the positive pair closer and push the negative pairs away in a batch based on InfoNCE loss as follows, where $\rho(\cdot)$ is the cosine similarity, $\phi(\cdot)$ is the $\exp(\cdot)$ function, and τ is the temperature parameter.

$$\mathcal{L}_{\text{CL}} = -\log \frac{\phi(\rho(\mathrm{s}_t, \mathrm{s}_{t+1})/\tau)}{\phi(\rho(\mathrm{s}_t, \mathrm{s}_{t+1})/\tau) + \sum_{i=1}^{\mathrm{B}-2} \phi(\rho(\mathrm{s}_t, \mathrm{v}_i)/\tau)}. \tag{2}$$

3.2.3. Objective with Uniformity Regularization

The masked process and temporal consistency strategy jointly ensure the extraction of domain-agnostic and sequence-level motion-discriminative WiFi representations. However, due to the inherent sparsity and homogeneity of WiFi signals, the learned representations may suffer from dimensional collapse. Here, we introduce explicit uniformity regularization to enhance representation diversity as follows:

$$\mathcal{L}_{\text{unif}} = \frac{1}{\mathrm{B}} \sum_{j=1, j\neq i}^{\mathrm{B}} \left(\hat{\mathbf{e}}_i^{\mathrm{T}} \hat{\mathbf{e}}_j\right)^2, \hat{\mathbf{e}}_i = \frac{\mathbf{e}_i}{\|\mathbf{e}_i\|_2}, \hat{\mathbf{e}}_j = \frac{\mathbf{e}_j}{\|\mathbf{e}_j\|_2}. \tag{3}$$

Overall, the pre-training optimization objective for each WiFi sample can be formulated as follows, where $\mathcal{L}_{\text{CL}}$ and λ_{unif} are trade-off hyperparameters.

$$\mathcal{L} = \mathcal{L}_{\text{Mask}} + \lambda_{\text{CL}} \cdot \mathcal{L}_{\text{CL}} + \lambda_{\text{unif}} \cdot \mathcal{L}_{\text{unif}}. \tag{4}$$

3.3. Topology-Constrained Pose Estimation

3.3.1. Adjacent Joint Local Modeling

After the pre-training phase, we freeze the pretrained encoder to extract WiFi representations $\mathbf{F} \in \mathbb{R}^{\mathrm{n}\times\mathrm{d}}$. To align these representations with the structure of human joints, we first pool all patches into one vector and repeat them into the joint numbers. Next, we add the learnable task prompt on them to obtain $\hat{\mathbf{F}} \in \mathbb{R}^{\mathrm{J}\times\mathrm{d}}$ for structural pose shape learning. Furthermore, we represent the human skeleton as a graph, where each joint is a vertex and each bone is an edge. This structure allows us to incorporate explicit spatial topology prior $\mathbf{A} \in \{0,1\}^{\mathrm{J}\times\mathrm{J}}$, where $\mathbf{A}_{\mathrm{i,j}} = 1$ indicates that the i-th joint and j-th joint are physical connected and $\mathbf{A}_{\mathrm{i,j}} = 0$ otherwise. Following the standard GCN formulation, we augment the adjacency matrix with self-connections as $\widetilde{\mathbf{A}} = \mathbf{A} + \mathbf{I}$. By employing Graph Convolution layers, we leverage $\widetilde{\mathbf{A}}$ as a structural constraint to aggregate information from spatially

connected joints, thereby mitigating the lack of spatial priors in $\hat{\mathbf{F}}$. Formally, the updated representation $\tilde{\mathbf{F}}$ is computed as follows, where $\mathbf{D} \in \mathbb{R}^{J\times J}$ is the degree matrix for normalization, $\mathbf{W}$ is a learnable parameter, σ is the activation function.

$$\tilde{\mathbf{F}} = \sigma\left(\mathbf{D}^{-\frac{1}{2}}\tilde{\mathbf{A}}\mathbf{D}^{-\frac{1}{2}}\hat{\mathbf{F}}\mathbf{W}\right). \tag{5}$$

3.3.2. Overarching Joint Holistic Modeling

Beyond local relationships, it is essential to capture holistic, long-range correlations among overarching joints. To this end, we treat the joints as an ordered sequence and apply Transformer encoder layers to enhance their non-physical interdependencies, such as the potential relationships between head and hand joints in "drinking water" pose. We calculate the attention values among all joints as follows, where W_Q, W_K, W_V are learnable parameters, $\tilde{d}$ is the dimension of K, and $\mathrm{LN}(\cdot)$ denotes the layer normalization.

$$Q = \tilde{F}W_Q, K = \tilde{F}W_K, V = \tilde{F}W_V, \tag{6}$$

$$Z_{\text{attn}} = \mathrm{LN}\left(\tilde{F} + \mathrm{softmax}\left(\frac{QK^{\top}}{\sqrt{\tilde{d}}}\right)V\right). \tag{7}$$

Then, we feed them into the feed-forward network $\mathrm{FFN}(\cdot)$ and regress them into pose coordinates by MLPs $\Psi(\cdot)$, where $\hat{Y}_i$ is the predicted pose.

$$Z = \mathrm{LN}(\mathrm{FFN}(Z_{\text{attn}}) + Z_{\text{attn}}), \hat{Y}_i = \Psi(Z). \tag{8}$$

By jointly capturing local and holistic dependencies, our hybrid decoder produces predicted structurally coherent and realistic poses.

3.3.3. Objective

For training the pose decoder, we adopt the MSE loss for each sample to regress the pose as follows:

$$\mathcal{L} = ||\hat{Y}_i - Y_i||_2^2. \tag{9}$$

4. Experiments

To evaluate the effectiveness of DT-Pose, we conduct comprehensive experiments across three mainstream datasets: MM-Fi [23], WiPose [19], and Person-in-WiFi-3D [11].

4.1. Datasets

4.1.1. MM-Fi

It comprises 27 distinct action categories performed by 40 volunteers across four different rooms, resulting in approximately 320.76 k single-person synchronized frames. One transmitter with one antenna and one receiver with three antennas capture all WiFi signals. Each skeletal pose consists of 17 joints encoded with 3D coordinates. To rigorously assess robustness, the dataset introduces three protocols and three settings for data splitting. Protocol 3 (P3) encompasses all 27 action categories, while Protocol 1 (P1) and Protocol 2 (P2) focus on 14 daily activities and 13 rehabilitation exercises, respectively. Setting 1 (S1 Random Split) randomly divides all data into training and testing sets with a 3:1 ratio. Setting 2 (S2 Cross-Subject Split) employs 32 subjects for training and the remaining 8 subjects for testing. Setting 3 (S3 Cross-Environment Split) selects 3 rooms randomly for training and the others for testing.

4.1.2. WiPose

It contains 12 action categories performed by 12 volunteers. WiFi signals are captured by one transmitter with three antennas and one receiver with three antennas. Each pose annotation comprises 18 joints in 2D coordinates. The official split provides 132,847 WiFi samples for training and 33,753 for testing.

4.1.3. Person-in-WiFi-3D

It includes 8 daily actions performed by 7 volunteers at three distinct locations. A single transmitter with one antenna and three receivers with three antennas capture all WiFi signals. Each skeleton pose features 14 joints with 3D coordinates. The dataset has been officially partitioned into training and test sets, with 89,946 WiFi samples allocated for training and 7824 for testing.

4.2. Experimental Setup

4.2.1. Implementation Details

In the pre-training phase, the encoder-decoder is trained for 400 epochs using AdamW, employing a batch size of 256, a learning rate of 1.5×10^{-4} with a cosine annealing schedule, a warm-up epoch of 40, and a weight decay of 0.05. The mask ratio is set as 80%. For the MM-Fi dataset, we train the pose decoder for 50 epochs using the SGD optimizer with a weight decay of 0.01. For the WiPose dataset, we train the pose decoder with the AdamW optimizer for 50 epochs. For the Person-in-WiFi-3D dataset, we train the pose decoder for 200 epochs using the AdamW optimizer. All the learning rates are 1×10^{-3}, and the batch size is 32. All the experiments are finished using the PyTorch platform on a GeForce RTX 4090 GPU (NVIDIA Corporation, Santa Clara, CA, USA).

4.2.2. Evaluation Metric

Three evaluation metrics are adopted following mainstream methods [12,22,23]. MPJPE (mm) measure the average Euclidean distance between ground truth and predictions, which is widely used to evaluate absolute positional accuracy. PA-MPJPE (mm) measure the MPJPE after aligning the predictions and ground truth using rigid transformations (translation, rotation, and scaling) by Procrustes analysis. Typically, it can be used to reflect the similarity in human shape and structure. PCK@α(%) measure the percentage of predictions that fall within a certain threshold distance from the ground truth. The threshold is set as a fraction α of the torso length following the previous works [12,22]. It is widely used to evaluate the local accuracy.

4.3. 2D & 3D HPE Performance Comparison

As shown in Tables 1–4, our framework outperforms existing methods in both 2D and 3D WiFi-based HPE tasks, illustrating its versatility and robust generalization. In particular, superior PA-MPJPE results highlight the plausibility of the predicted poses and the structural coherence of the generated skeletons. Notably, the remarkable gains under cross-domain settings in Table 1 confirm that our pretraining strategy successfully captures generalizable representations. Furthermore, our method also delivers competitive performance in terms of efficiency in Table 2.

Table 1. 3D HPE results on MM-Fi dataset. P refers to PCK@. M refers to MPJPE. PM refers to PA-MPJPE.

	Protocol 1				**Protocol 2**				**Protocol 3**			
Method	**P20**	**P50**	**M**	**PM**	**P20**	**P50**	**M**	**PM**	**P20**	**P50**	**M**	**PM**
Setting 1 (Random Split)												
MetaFi++ [22]	49.1	86.5	186.9	120.7	32.2	81.7	213.5	121.4	43.9	85.0	197.1	121.2
HPE-Li [12]	56.2	87.6	173.4	104.5	36.9	81.9	206.1	102.7	49.6	85.6	184.3	106.4
DT-Pose (Ours)	**59.4**	**88.9**	**165.3**	**101.0**	**41.4**	**83.5**	**195.6**	**101.2**	**51.7**	**86.5**	**178.5**	**104.5**
Setting 2 (Cross-Subject)												
MetaFi++ [22]	36.4	85.5	222.3	125.4	24.0	77.5	247.0	122.7	32.3	81.9	231.1	124.0
HPE-Li [12]	38.2	82.8	228.6	106.8	26.9	78.0	242.6	101.9	36.5	80.8	228.6	107.7
DT-Pose (Ours)	**41.9**	**86.7**	**213.0**	**105.6**	**28.5**	**78.5**	**238.3**	**101.1**	**37.7**	**82.6**	**221.6**	**106.2**
Setting 3 (Cross-Environment)												
MetaFi++ [22]	9.3	55.1	367.8	121.0	5.3	45.9	360.2	117.2	6.4	49.1	369.5	116.0
HPE-Li [12]	4.3	47.8	381.1	110.3	4.2	40.3	378.2	104.0	3.4	41.9	388.4	107.9
DT-Pose (Ours)	**10.7**	**58.8**	**332.7**	**105.1**	**4.4**	**49.7**	**338.3**	**102.0**	**9.8**	**61.2**	**316.8**	**104.2**

Bold values indicate the best results.

Table 2. 2D HPE results and efficiency comparison on WiPose.

Method	**MPJPE**	**PA-MPJPE**	**Params (M)**	**Flops (G)**
MetaFi++ [22]	49.2	30.1	25.6	502.3
HPE-Li [12]	40.9	25.9	**3.5**	5.2
DT-Pose (Ours)	**34.3**	**23.1**	3.8	**1.5**

Bold values indicate the best results.

Table 3. 3D HPE results on Person-in-WiFi-3D (1 Person).

Method	MPJPE	PA-MPJPE
MetaFi++ [22]	132.0	75.8
HPE-Li [12]	120.2	69.5
Wi-Pose [17]	101.8	-
PiW3D [11]	91.7	-
DT-Pose (Ours)	**90.0**	**58.7**

Bold values indicate the best results.

Table 4. 2D HPE results on MM-Fi (P3-S1).

Method	PCK@20	PCK@50	MPJPE	PA-MPJPE
Wi-Pose [17]	48.6	82.4	158.2	97.7
Wi-Mose [46]	48.7	83.9	155.8	95.4
WiLDAR [47]	44.1	79.3	170.3	115.6
WiSPPN [15]	45.4	81.0	166.5	110.0
PerUnet [19]	50.1	83.6	154.6	98.6
MetaFi++ [22]	45.5	81.8	164.4	106.3
HPE-Li [12]	52.1	85.1	149.4	92.5
DT-Pose (Ours)	**65.8**	**89.8**	**137.0**	**92.3**

Bold values indicate the best results.

4.4. Cross-Dataset Evaluation

We further evaluate the cross-dataset generalization ability of our method. As shown in Table 5, we pre-train the model on MM-Fi and transfer it to the other two datasets, WiPose and Person-in-WiFi-3D, to evaluate whether the representations learned from MM-Fi can generalize beyond the source dataset. Training with pose supervision from scratch on each target dataset yields unsatisfactory in-dataset performance (row 1), indicating strong domain pose dependency. In contrast, our self-supervised pretraining (row 2), which does not rely on pose annotations, achieves the best performance, demonstrating its effectiveness in learning robust representations. Furthermore, cross-dataset pretraining on MM-Fi followed by transfer to other datasets also outperforms in-dataset supervised training from scratch (row 3), and even surpasses prior supervised methods (rows 4–5). These results suggest that our SSL paradigm alleviates domain bias and enables better generalization across environments, although a gap still remains under severe domain shifts.

Table 5. Cross-dataset Transfer Evaluation of DT-Pose. SL refers to supervised learning, and SSL indicates self-supervised learning.

Training Strategy	WiPose		Person-in-WiFi-3D	
	MPJPE	PA-MPJPE	MPJPE	PA-MPJPE
In-dataset SL form Scratch	43.9	27.6	120.1	63.8
In-dataset SSL Pretraining	34.3	23.1	90.0	58.7
Cross-dataset (MM-Fi) SSL Pretraining	36.3	23.9	111.6	63.0
MetaFi++ [22] (In-dataset SL form Scratch)	49.2	30.1	132.0	75.8
HPE-Li [12] (In-dataset SL form Scratch)	40.9	25.9	120.2	69.5

4.5. Ablation Study

4.5.1. Influence of Masking Ratios

In Figure 4, the performance improves with higher ratios but drops beyond 80% due to the excessive reconstruction difficulty. Thus, we use the default ratio at 80%.

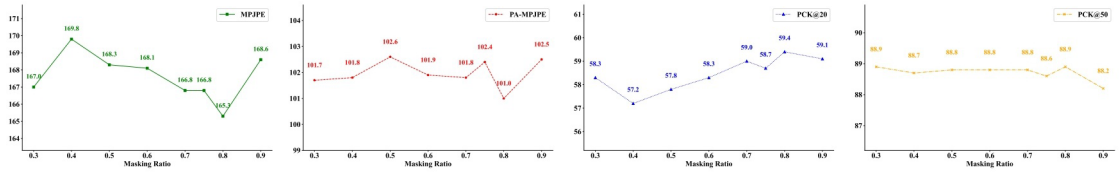


Figure 4. Performance on the MM-Fi (P1-S1) with different masking ratios.

4.5.2. Influence of Masking Strategies

Table 6 shows that the unstructured masking strategy yields the best performance during pretraining, primarily because it captures contextual cues across both time and channel levels.

Table 6. Influence of masking strategies on the MM-Fi (P1-S1).

Masking Strategy	MPJPE	PA-MPJPE
Channel-Structured	175.2	103.9
Time-Structured	180.7	107.1
Unstructured (Ours)	**165.3**	**101.0**

Bold values indicate the best results.

4.5.3. Influence of Pre-training Components

Table 7 shows that our pretraining phase learns more general and robust representations than training from scratch with pose supervision. Specifically, by capturing implicit motion and spatial localization patterns from WiFi signals, the pretrained encoder consistently reduces MPJPE under different settings, indicating improved absolute localization ability. Since this stage does not explicitly model skeletal topology, its influence on PA-MPJPE mainly comes from the improved quality of the learned WiFi representations. Moreover, the temporal consistency strategy further enhances performance, highlighting the importance of motion-aware learning in WiFi-based HPE.

Table 7. Pretraining component analysis on the MM-Fi (P1-S1 & P1-S3) dataset.

Masked Operation	Temporal Consistency	Uniformity	Protocol 1—Setting 1		Protocol 1—Setting 3	
			MPJPE	PA-MPJPE	MPJPE	PA-MPJPE
×	×	×	198.6	**100.9**	370.6	107.6
✓	×	×	183.1	102.0	351.1	106.5
✓	✓	×	173.1	102.7	341.0	105.8
✓	×	✓	181.8	101.9	339.9	106.7
✓	✓	✓	**165.3**	101.0	**332.7**	**105.1**

Bold values indicate the best results.

4.5.4. Influence of Pose Decoding Components

Table 8 shows that replacing the simple MLP decoder with our adjacent-overarching joint modeling with explicit priors significantly improves absolute joint localization, underscoring the value of capturing both local and global skeletal dependencies. The simple MLP decoder performs an unstructured patch-to-joint projection, while the task prompt provides joint-specific anchors by differentiating the repeated global WiFi representation into stable joint-aware tokens. Built upon this initialization, GCN and Transformer further model local physical connections and global joint dependencies, respectively, leading to the best overall performance.

Table 8. Pose decoder component analysis on the MM-Fi (P1-S1 & P1-S3) dataset.

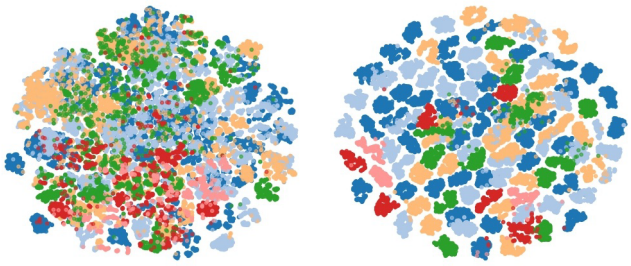
Task Prompt	Graph Convolution	Transformer	Protocol 1—Setting 1		Protocol 1—Setting 3	
			MPJPE	PA-MPJPE	MPJPE	PA-MPJPE
×	×	×	197.4 †	103.5 †	376.8 †	109.7 †
✓	×	×	174.1	101.3	365.6	106.1
×	✓	×	179.8 ‡	107.0 ‡	353.6 ‡	108.3 ‡
×	×	✓	181.4 ‡	103.0 ‡	352.1 ‡	107.5 ‡
✓	✓	×	166.7	103.2	349.4	106.2
✓	×	✓	167.1	101.1	348.5	106.0
×	✓	✓	167.0	103.3	343.5	107.2
✓	✓	✓	**165.3**	**101.0**	**332.7**	**105.1**

†: MLPs as the pose decoder; ‡: we transform all patches into the number of joints by MLPs. Bold values indicate the best results.

4.6. Qualitative Analysis

4.6.1. Temporal Consistency Strategy

As shown in Figure 5, incorporating the temporal consistency strategy enhances inter-sequence separability and intra-sequence compactness, thereby strengthening temporal coherence and improving sequence-level motion discrimination across action sequences.



(a) w/o TCS (b) w/ TCS

Figure 5. t-SNE visualization of WiFi representations on MM-Fi (P1-S1) with and without the temporal consistency strategy (TCS) in the pretraining phase. Each color indicates an action class.

4.6.2. Dimension Collapse Phenomenon

In Figure 6a, we calculate the covariance values of each dimension of the WiFi representations. A more compact distribution is clearly visible when the uniformity term is included, implying that it enriches dimensional diversity and improves inter-dimensional dependencies. Additionally, Figure 6b represents the singular values of WiFi representations. More large singular values emerge upon introducing the uniformity term, suggesting that the embedding space mitigates the dimension collapse and preserves richer information.

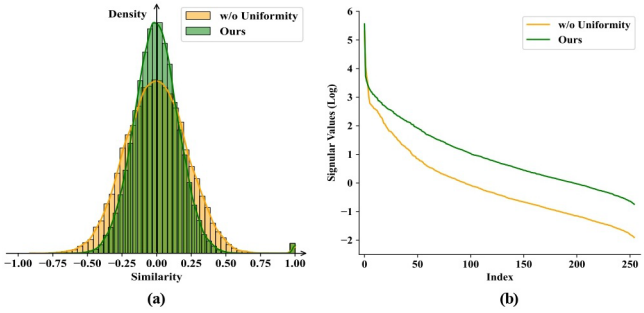
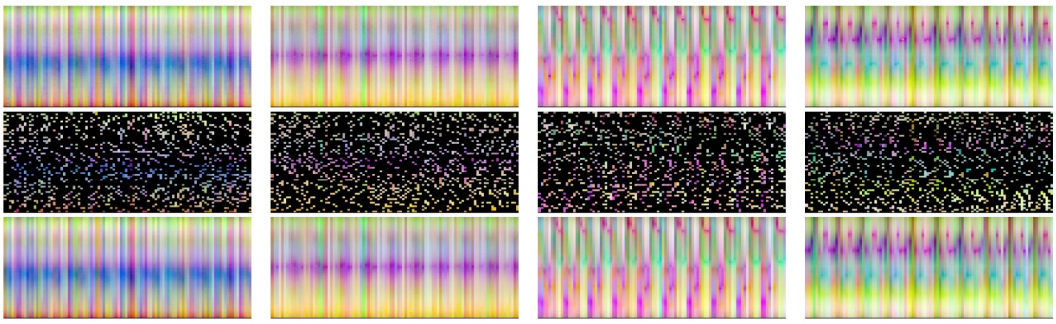


Figure 6. Dimension collapse. (**a**) represents the statistics of the covariance values of the WiFi representation dimensions. (**b**) compares the singular values of WiFi representations.

4.6.3. Masking-Reconstruction Visualization

In Figure 7, we plot the raw, masked, and reconstructed WiFi signals, selecting four actions to highlight variations in WiFi signal patterns. Our method faithfully reconstructs the original WiFi signals, underscoring its ability to capture domain-consistent and sequence-level motion-discriminative WiFi representations effectively.

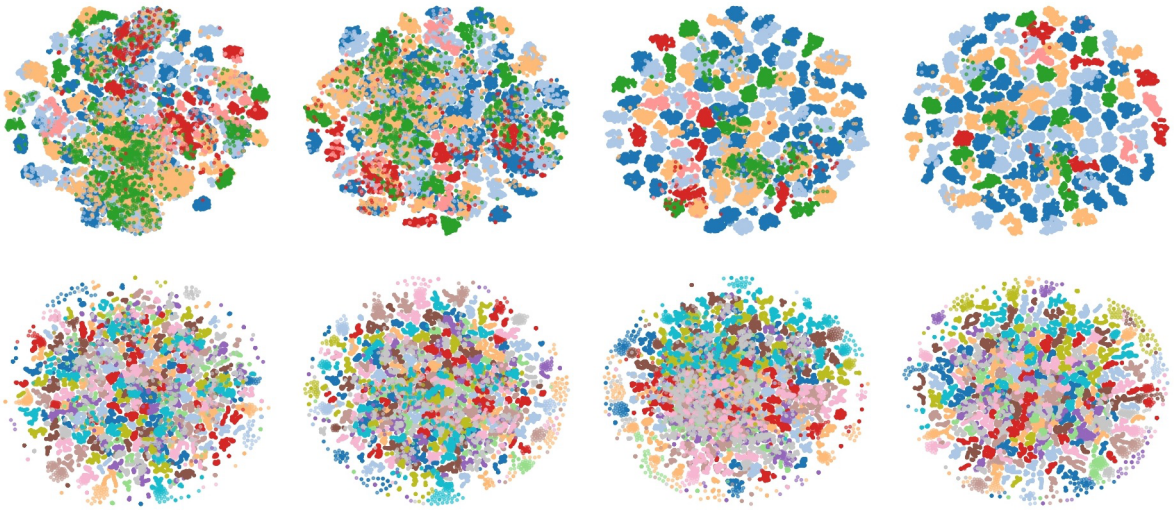


(a) (b) (c) (d)

Figure 7. WiFi visualizations on the MM-Fi (P3-S1). The first row represents the raw WiFi signals, the second row represents the masked WiFI input, and the third row denotes the reconstructed WiFi output. All of them contain ten continuous frames. (**a**) refers to Chest Expanding Horizontally action, (**b**) refers to Chest Expanding Vertically action, (**c**) refers to Raising Right Arm action, and (**d**) refers to Waving Left Arm action.

4.6.4. WiFi Representations Comparison

Figure 8 provides the t-SNE visualization of WiFi representations across multiple models and datasets. In contrast to both the raw WiFi signals and other existing methods, the learned WiFi representations of our proposed method exhibit excellent inter-sequence separability and intra-sequence compactness. Consequently, this sequence-level motion-discriminative representation space benefits subsequent pose estimation tasks, whereas prior methods tend to learn spurious or motion-irrelevant features that can undermine estimation accuracy.



(a) Original WiFi (b) HPE-Li [12] (c)MetaFi++ [22] (d) Ours

Figure 8. t-SNE visualization of WiFi representations. The first row denotes the representations extracted on the MM-Fi (P1-S1) testing set, and the second row represents the representations obtained on the WiPose testing set. Each color corresponds to an action category.

4.6.5. Pose Realistic

Figure 9 compares predicted poses across various methods and datasets. Our predictions exhibit a more consistent motion tendency in the MM-Fi dataset, highlighted by the green circles. Moreover, as the resolution of WiFi signals increases in the other two datasets, the predicted poses of our method become more coherent and precise. Notably, our predicted skeletal structures adhere closely to the human topology.

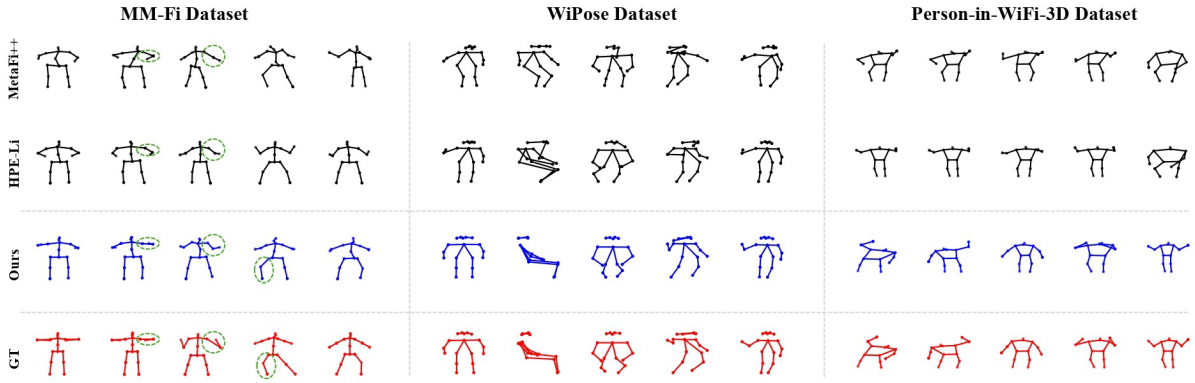


Figure 9. Predicted poses of MetaFi++, HPE-Li, and our method among all datasets.

4.7. Limitation and Discussion

4.7.1. Joints Analysis

To evaluate the joint-level accuracy of our method, we calculate the pose estimation error for each joint, as shown in Tables 9–11. Our method performs superiorly on coarse-grained body parts like the torso. However, the hands and elbows exhibit the highest errors. These results stem from the limited resolution of current WiFi signals, which hinders the capture of fine-grained actions, e.g., hand movements. In Tables 10 and 11, hand and elbow joints errors decrease when an increased number of antennas and receivers is employed for WiFi signal capture. Consequently, fine-grained pose estimation with WiFi necessitates higher-resolution signals, whereas images can achieve it even at lower resolutions.

Table 9. Per-joint performance on MM-Fi (P1-S1). L. and R. denote left and right, respectively.

Joints	MPJPE	PA-MPJPE
Bot Torso	102.7	56.7
L. Hip	106.9	63.7
L. Knee	105.7	66.1
L. Foot	104.1	88.6
R. Hip	108.3	64.3
R. Knee	105.3	67.4
R. Foot	109.5	90.8
Center Torso	112.3	44.5
Upper Torso	135.8	54.2
Neck Base	158.2	66.2
Center Head	160.5	70.9
R. Shoulder	147.7	73.6
R. Elbow	249.1	140.5
R. Hand	364.5	284.0
L. Shoulder	141.8	77.2
L. Elbow	235.8	132.4
L. Head	362.4	277.0
Average	165.3	101.0

Underlined values indicate human body regions with larger errors.

Table 10. Per-joint performance on WiPose. L. and R. denote left and right, respectively.

Joints	MPJPE	PA-MPJPE
Nose	30.9	14.9
Neck	27.3	11.7
R. Shoulder	28.7	13.4
R. Elbow	38.7	25.6
R. Wrist	48.2	35.8
L. Shoulder	29.8	16.6
L. Elbow	37.2	24.9
L. Wrist	43.3	30.6
R. Hip	24.6	17.2
R. Knee	21.0	19.3
R. Ankle	22.6	21.7
L. Hip	25.6	17.9
L. Knee	22.4	19.2

Table 10. *Cont.*

Joints	MPJPE	PA-MPJPE
L. Ankle	26.0	22.2
R. Eye	31.6	15.5
L. Eye	32.4	16.3
R. Ear	30.8	14.9
L. Ear	96.8	77.7
Average	34.3	23.1

Underlined values indicate human body regions with larger errors.

Table 11. Per-joint performance on Person-in-WiFi-3D. L. and R. denote left and right, respectively.

Joints	MPJPE	PA-MPJPE
Neck	71.6	36.2
Head	77.6	43.1
L. Shoulder	80.5	37.8
R. Shoulder	81.1	37.8
L. Elbow	107.4	54.1
L. Hip	57.7	41.5
R. Elbow	114.2	55.1
R. Hip	58.6	42.0
L. Hand	164.8	117.9
L. Knee	65.6	52.2
R. Hand	179.2	122.4
R. Knee	64.3	52.8
L. Ankle	69.8	65.8
R. Ankle	67.4	62.5
Average	90.0	58.7

Underlined values indicate human body regions with larger errors.

4.7.2. Modality Comparison

Table 12 compares the HPE performance between images and WiFi, highlighting a notable performance gap that can be primarily attributed to two factors: (1) images inherently encode human spatial priors, which are absent in WiFi signals; and (2) the spatial resolution of existing WiFi signals remains limited. Nevertheless, the two modalities can complement each other: WiFi for low-light or occluded scenarios, while images provide high-resolution spatial details in well-lit environments.

Table 12. Modality comparison on MM-Fi.

Modality	MPJPE	PA-MPJPE
Protocol 3—Setting 1 (Random Split)		
Image [23]	279.0	**81.2**
WiFi (Ours)	**178.5**	104.5
Protocol 3—Setting 2 (Cross-Subject)		
Image [23]	285.3	**81.9**
WiFi (Ours)	**221.6**	106.2
Protocol 3—Setting 3 (Cross-Environment)		
Image [23]	**288.6**	**84.1**
WiFi (Ours)	316.8	104.2

Bold values indicate the best results.

4.7.3. Other Limitations and Future Work

Besides the aforementioned limitations inherent to WiFi signals, our method still has several aspects that warrant further discussion. First, its performance under cross-environment and cross-dataset settings, as shown in Table 5, remains limited. Although the proposed self-supervised learning paradigm alleviates the domain gap, severe multipath interference still leads to relatively high absolute errors. Few-shot adaptation to new environments may be a promising solution, which we plan to explore in future work. Second, our two-stage framework relies on the compatibility between the pretrained representation and the topology-aware decoder. When the pretrained WiFi representation lacks sufficient structure-related information, the decoder may still struggle to recover precise local skeletal relationships. Thus, jointly optimizing representation learning and structural decoding is another important direction for improving robustness.

5. Conclusions

In this paper, we revisit and highlight two critical challenges in WiFi-based human pose estimation (HPE): (1) the cross-domain gap and (2) the structural fidelity gap. To tackle these issues, we introduce a two-phase framework, ***DT-Pose***, and evaluate its effectiveness through extensive experiments on both 2D and 3D WiFi-based HPE tasks. Furthermore, we discuss the limitations and advantages of WiFi signals, emphasizing their suitability for Edge AI applications in the AIoT era.

Author Contributions

Y.C.: conceptualization, methodology, experiments, and writing; J.G.: conceptualization, reviewing and editing, supervision. All authors have read and agreed to the published version of the manuscript.

Funding

This research was funded by Hong Kong RGC General Research Fund (Grant Nos. 15221123, 15216424, and 15211525) and the Hong Kong PolyU Internal Research Fund (Grant Nos. P0058468 and P0056171).

Institutional Review Board Statement

Not applicable.

Informed Consent Statement

Not applicable.

Data Availability Statement

The associated code is released at https://github.com/cseeyangchen/DT-Pose.

Conflicts of Interest

Given the role as the Executive Editor-in-Chief of Edge Intelligence and Systems, Jingcai Guo had no involvement in the peer review of this paper and had no access to information regarding its peer-review process. Full responsibility for the editorial process of this paper was delegated to another editor of the journal. The Hong Kong RGC General Research Fund and the Hong Kong PolyU Internal Research Fund had involvement in all stages of the research process. The authors take full responsibility for the content of the published article.

Use of AI and AI-Assisted Technologies

During the preparation of this work, the author(s) used ChatGPT to assist with language editing and sentence polishing. After using this tool, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the published article.

References

1. Cao, Z.; Simon, T.; Wei, S.E.; et al. Realtime multi-person 2D pose estimation using part affinity fields. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 7291–7299.
2. Wang, Y.; Li, M.; Cai, H.; et al. Lite pose: Efficient architecture design for 2D human pose estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 19–24 June 2022; pp. 13126–13136.
3. Li, W.; Liu, H.; Ding, R.; et al. Exploiting temporal contexts with strided transformer for 3D human pose estimation. *IEEE Trans. Multimed.* **2022**, *25*, 1282–1293.
4. Gong, J.; Foo, L.G.; Fan, Z.; et al. Diffpose: Toward more reliable 3D pose estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, Canada, 18–22 June 2023; pp. 13041–13051.
5. Zhang, F.; Zhu, X.; Wang, C. Single person pose estimation: A survey. *arXiv* **2021**, *preprint*, arXiv:2109.10056.
6. Shi, D.; Wei, X.; Li, L.; et al. End-to-end multi-person pose estimation with transformers. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 19–24 June 2022; pp. 11069–11078.

7. Liu, H.; Chen, Q.; Tan, Z.; et al. Group pose: A simple baseline for end-to-end multi-person pose estimation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 2–6 October 2023; pp. 15029–15038.
8. Zheng, C.; Wu, W.; Chen, C.; et al. Deep learning-based human pose estimation: A survey. *ACM Comput. Surv.* **2023**, *1*, 1–37.
9. Zheng, J.; Shi, X.; Gorban, A.; et al. Multi-modal 3D human pose estimation with 2D weak supervision in autonomous driving. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 19–24 June 2022; pp. 4478–4487.
10. He, T.; Chen, Y.; Wang, L.; et al. An expert-knowledge-based graph convolutional network for skeleton-based physical rehabilitation exercises assessment. *IEEE Trans. Neural Syst. Rehabil. Eng.* **2024**, *32*, 1916–1925.
11. Yan, K.; Wang, F.; Qian, B.; et al. Person-in-WIFI 3D: End-to-end multi-person 3D pose estimation with WIFI. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 17–21 June 2024; pp. 969–978.
12. Gian, T.D.; Lai, T.D.; Luong, T.V.; et al. Hpe-li: Wifi-enabled lightweight dual selective kernel convolution for human pose estimation. In *European Conference on Computer Vision*; Springer Nature: Cham, Switzerland, 2024; pp. 93–111.
13. Fan, J.; Yang, J.; Xu, Y.; et al. Diffusion model is a good pose estimator from 3D rf-vision. In *European Conference on Computer Vision*; Springer Nature: Cham, Switzerland, 2024; pp. 1–18.
14. Chen, W.; Yu, C.; Tu, C.; et al. A survey on hand pose estimation with wearable sensors and computer-vision-based methods. *Sensors* **2020**, *20*, 1074.
15. Wang, F.; Panev, S.; Dai, Z.; et al. Can WiFi estimate person pose? *arXiv* **2019**, *preprint*, arXiv:1904.00277.
16. Wang, F.; Zhou, S.; Panev, S.; et al. Person-in-WiFi: Fine-grained person perception using WiFi. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, South Korea, 27 October–2 November 2019; pp. 5452–5461.
17. Jiang, W.; Xue, H.; Miao, C.; et al. Towards 3D human pose construction using WiFi. In Proceedings of the 26th Annual International Conference on Mobile Computing and Networking, London, UK, 21–25 September 2020; pp. 1–14.
18. Ren, Y.; Wang, Z.; Wang, Y.; et al. GoPose: 3D human pose estimation using WiFi. In *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*; Association for Computing Machinery: New York, NY, USA, 2022; Vol. 6, pp. 1–25.
19. Zhou, Y.; Zhu, A.; Xu, C.; et al. PerUnet: Deep signal channel attention in UNET for WiFi-based human pose estimation. *IEEE Sens. J.* **2022**, *20*, 19750–19760.
20. Yang, J.; Chen, X.; Zou, H.; et al. AutoFi: Toward automatic Wi-Fi human sensing via geometric self-supervised learning. *IEEE Internet Things J.* **2022**, *8*, 7416–7425.
21. Ren, Y.; Wang, Z.; Tan, S.; et al. Winect: 3D human pose tracking for free-form activity using commodity WIFI. In *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*; Association for Computing Machinery: New York, NY, USA, 2021; Vol. 4, pp. 1–29.
22. Zhou, Y.; Huang, H.; Yuan, S.; et al. MetaFi++: WiFi-enabled transformer-based human pose estimation for metaverse avatar simulation. *IEEE Internet Things J.* **2023**, *16*, 14128–14136.
23. Yang, J.; Huang, H.; Zhou, Y.; et al. Mm-fi: Multi-modal non-intrusive 4D human dataset for versatile wireless sensing. *Adv. Neural Inf. Process. Syst.* **2023**, *36*, 18756–18768.
24. Zhou, Y.; Yang, J.; Huang, H.; et al. AdaPose: Toward cross-site device-free human pose estimation with commodity WiFi. *IEEE Internet Things J.* **2024**, *24*, 40255–40267.
25. He, K.; Chen, X.; Xie, S.; et al. Masked autoencoders are scalable vision learners. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 19–24 June 2022; pp. 16000–16009.
26. He, K.; Zhang, X.; Ren, S.; et al. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 770–778.
27. Devlin, J.; Chang, M.W.; Lee, K.; et al. Bert: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Minneapolis, MN, USA, 2–7 June 2019; pp. 4171–4186.
28. Radford, A.; Narasimhan, K.; Salimans, T.; et al. Improving Language Understanding by Generative Pre-Training. 2018. Available from: https://www.mikecaptain.com/resources/pdf/GPT-1.pdf (accessed on 11 June 2018).
29. Bao, H.; Dong, L.; Piao, S.; et al. Beit: Bert pre-training of image transformers. *arXiv* **2021**, *preprint*, arXiv:2106.08254.
30. Tong, Z.; Song, Y.; Wang, J.; et al. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 10078–10093.
31. Wang, L.; Huang, B.; Zhao, Z.; et al. Videomae v2: Scaling video masked autoencoders with dual masking. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, Canada, 18–22 June 2023; pp. 14549–14560.
32. Huang, P.Y.; Xu, H.; Li, J.; et al. Masked autoencoders that listen. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 28708–28720.

33. Yan, H.; Liu, Y.; Wei, Y.; et al. Skeletonmae: Graph-based masked autoencoder for skeleton sequence pre-training. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 2–6 October 2023; pp. 5606–5618.
34. Cheng, M.; Tao, X.; Liu, Z.; et al. TimeMAE: Self-Supervised Representations of Time Series with Decoupled Masked Autoencoders. In Proceedings of the Nineteenth ACM International Conference on Web Search and Data Mining, Houston, TX, USA, 26–30 February 2026; pp. 498–508.
35. Wang, P.; Li, Z.; Hou, Y.; et al. Action recognition based on joint trajectory maps using convolutional neural networks. In Proceedings of the 24th ACM International Conference on Multimedia, Amsterdam, Netherlands, 15–19 October 2016; pp. 102–106.
36. Chen, Y.; Guo, J.; He, T.; et al. Fine-grained side information guided dual-prompts for zero-shot skeleton action recognition. In Proceedings of the 32nd ACM International Conference on Multimedia, Melbourne, Australia, 28 October–1 November 2024; pp. 778–786.
37. Chen, Y.; Zhang, Z.; Yuan, C.; et al. Channel-wise topology refinement graph convolution for skeleton-based action recognition. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 11–17 October 2021; pp. 13359–13368.
38. Song, Y.F.; Zhang, Z.; Shan, C.; et al. Constructing stronger and faster baselines for skeleton-based action recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *2*, 1474–1488.
39. Chi, H.; Ha, M.H.; Chi, S.; et al. Infogcn: Representation learning for human skeleton-based action recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 19–24 June 2022; pp. 20186–20196.
40. Chen, Y.; Guo, J.; Guo, S.; et al. Neuron: Learning context-aware evolving representations for zero-shot skeleton action recognition. In Proceedings of the Computer Vision and Pattern Recognition Conference, Nashville, TN, USA, 11–15 June 2025; pp. 8721–8730.
41. Liu, M.; Liu, H.; Chen, C. Enhanced skeleton visualization for view invariant human action recognition. *Pattern Recognit.* **2017**, *68*, 346–362.
42. Chen, Y.; He, T.; Fu, J.; et al. Vision-language meets the skeleton: Progressively distillation with cross-modal knowledge for 3d action representation learning. *IEEE Trans. Multimed.* **2024**, *27*, 2293–2303.
43. Plizzari, C.; Cannici, M.; Matteucci, M. Spatial temporal transformer network for skeleton-based action recognition. In *International Conference on Pattern Recognition*; Springer International Publishing: Cham, Switzerland, 2021; pp. 694–701.
44. Gao, Z.; Wang, P.; Lv, P.; et al. Focal and global spatial-temporal transformer for skeleton-based action recognition. In Proceedings of the Asian Conference on Computer Vision, Macao, China, 4–8 December 2022; pp. 382–398.
45. He, T.; Chen, Y.; Gao, X.; et al. Enhancing skeleton-based action recognition with language descriptions from pre-trained large multimodal models. *IEEE Trans. Circuits Syst. Video Technol.* **2024**, *3*, 2118–2132.
46. Wang, Y.; Guo, L.; Lu, Z.; et al. From point to space: 3D moving human pose estimation using commodity WiFi. *IEEE Commun. Lett.* **2021**, *7*, 2235–2239.
47. Deng, F.; Jovanov, E.; Song, H.; et al. WiLDAR: WiFi signal-based lightweight deep learning model for human activity recognition. *IEEE Internet Things J.* **2023**, *2*, 2899–2908.