

HuPR: A Benchmark for Human Pose Estimation Using Millimeter Wave Radar

Shih-Po Lee^{†,*}, Niraj Prakash Kini[†], Wen-Hsiao Peng[†], Ching-Wen Ma[†], Jenq-Neng Hwang[‡]

[†]National Yang Ming Chiao Tung University, Taiwan

[‡]University of Washington, USA

{map10756051.cs07g, nirajnctu.cs06g, machingwen}@nctu.edu.tw, wpeng@cs.nctu.edu.tw, hwang@uw.edu

Abstract

This paper introduces a novel human pose estimation benchmark, Human Pose with Millimeter Wave Radar (HuPR), that includes synchronized vision and radio signal components. This dataset is created using cross-calibrated mmWave radar sensors and a monocular RGB camera for cross-modality training of radar-based human pose estimation. There are two advantages of using mmWave radar to perform human pose estimation. First, it is robust to dark and low-light conditions. Second, it is not visually perceivable by humans and thus, can be widely applied to applications with privacy concerns, e.g., surveillance systems in patient rooms. In addition to the benchmark, we propose a cross-modality training framework that leverages the ground-truth 2D keypoints representing human body joints for training, which are systematically generated from the pre-trained 2D pose estimation network based on a monocular camera input image, avoiding laborious manual label annotation efforts. The framework consists of a new radar pre-processing method that better extracts the velocity information from radar data, Cross- and Self-Attention Module (CSAM), to fuse multi-scale radar features, and Pose Refinement Graph Convolutional Networks (PRGCN), to refine the predicted keypoint confidence heatmaps. Our intensive experiments on the HuPR benchmark show that the proposed scheme achieves better human pose estimation performance with only radar data, as compared to traditional pre-processing solutions and previous radio-frequency-based methods. Our code is available at [here](https://github.com/robert80203/HuPR-A-Benchmark-for-Human-Pose-Estimation-Using-Millimeter-Wave-Radar)¹

1. Introduction

Human pose estimation (HPE) is one of the widely studied traditional tasks in computer vision. Given the RGB

^{*}This work is supported by National Center for High-performance Computing, Taiwan.

¹<https://github.com/robert80203/HuPR-A-Benchmark-for-Human-Pose-Estimation-Using-Millimeter-Wave-Radar>

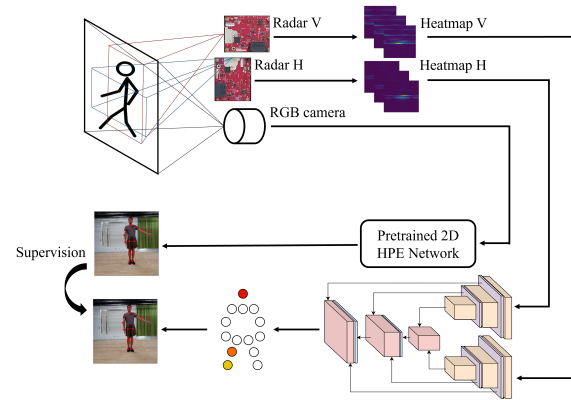


Figure 1: Illustration of our overall system setup. The upper branch represents mmWave Data Collection and Pre-processing. The lower branch represents our proposed Cross and Self Attention Module and Pose Refinement Graph Convolutional Networks. H: Horizontal, V: Vertical.

images under a single or multiple camera view, it predicts 2D/3D human skeletons, in terms of estimating human body keypoints. Though promising results have been demonstrated by previous HPE solutions, the natural properties of an RGB image undoubtedly constrain the advancement of HPE. In particular, the RGB images captured in the dark and a low-light conditions can hardly show a person's pose, leading to an inferior quality of pose estimation. In addition, using such vision-based inputs consequently raise the concern of the personal privacy. For example, the surveillance systems installed in the patient rooms monitor personal activities by analyzing their poses while at the same time, the personal appearance is inevitably disclosed to the systems. Therefore, predicting human poses using vision-based input encounters adverse lighting and privacy invasion issues.

To address the above issues, a new type of HPE task has been proposed. Several radio frequency (RF) datasets [30, 19, 9] are built to predict human skeletons. Such RF signals are robust to lighting conditions and barely visually perceivable by humans. RF signals can be catego-

rized by the frequency bands with varying characteristics. Zhao *et al.* [30] adopt Wi-Fi signals (2.4 GHz), which possess a unique ability that it is able to capture a person's pose even when he or she is standing behind a wall. In spite of showing excellent results on 2D pose estimation, the Wi-Fi sensors used by [30] is proprietorially-designed. On the other hand, Sengupta *et al.* [19] introduces another type of RF signal, Frequency Modulated Continuous Wave (FMCW) radar, with frequency band periodically changing from 77GHz to 81GHz, which can precisely detect the depth (range) and the velocity of an object. Comparing to the Wi-Fi sensor used by [30], the mmWave radar sensor is more economical and accessible, as well as commercially available from many instrument providers [8]. The 3D HPE results shown in [19] seem promising; however, it ignores the human body keypoints with high uncertainty, such as wrists, due to their low prediction accuracy, showing an inferior capability of capturing human poses using radar. Most importantly, the datasets of both [30, 19] remain inaccessible to the public, restricting the further development of an HPE in terms of RF data.

To overcome the challenging issues encountered in RGB-based and RF-based HPE, we introduce a new benchmark, Human Pose with mmWave Radar (HuPR). Unlike [19], we additionally incorporate velocity information in our dataset, since radar sensors can provide a highly precise velocity information. Meanwhile, we propose a Cross-and-Self-Attention Module (CSAM) to better fuse the multi-scale features from horizontal and vertical radars and a 2D pose confidence refinement network based on Graph Convolutional Network (PRGCN) to refine the confidence in the output pose heatmaps. Our framework consists of two cascading components, 1) Multi-Scale Spatio-Temporal Radar Feature Fusion with CSAM, which contains two branches to encode temporal range-azimuth and range-elevation information respectively, followed by a decoder to decode the fused features at every scale and predict 2D pose heatmaps and 2) PRGCN which is applied to the output heatmaps to refine the confidence of each keypoint based on a pre-defined graph of human skeletons. Our contributions are threefold:

- We introduce a novel RF-based HPE benchmark, HuPR, which features privacy-preserving data, economical and accessible radar sensors, and handy hardware setup. The dataset and implementation code will be released upon paper acceptance.
- We propose a new radar pre-processing method that better extracts velocity information from radar signals to help RF-based HPE.
- We propose CSAM to relate the features from two different radars for better feature fusion and PRGCN to refine the confidence of each keypoint, especially to

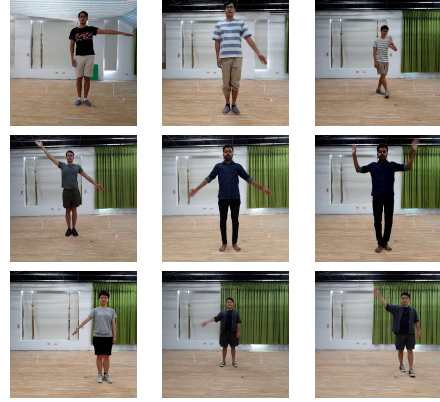


Figure 2: Examples of actions in our dataset, including standing with fixed actions, standing with waving hands, and walking with waving hands.

improve the precision of the faster moving edge keypoints, such as wrists. Experimental results and ablation studies show that our proposed method makes significant improvement over RF-based 2D HPE methods and 3D pointcloud-based methods.

2. Related Work

2.1. RGB-based HPE

There have been extensive studies on RGB-based HPE. In general, these works can be split into two categories: regression-based methods and heatmap-based methods. Traditional regression-based methods [15, 27, 10] map input sources to the coordinates of body keypoints via an end-to-end neural network. The regression-based solutions are straightforward but less attractive since it is more difficult for a neural network to map image features into just several keypoint coordinates. On the other hand, heatmap-based methods [22, 11, 14, 6, 3] generally outperforming regression-based methods and dominate the field of HPE. Heatmap-based HPEs produce likelihood heatmaps for each keypoint as the target of pose estimation.

2.2. RF-based HPE

RF-based data are often used to deal with simpler human sensing tasks, such as activity recognition [20, 21], gesture recognition [24, 13] and human object detection [5, 23]. Channel State Information (CSI) data are the main RF sources in early days, but they do not provide range or distance information. With the development of economical radio sensors, the estimation of range and angle of arrival becomes feasible with affordable devices, allowing more detailed and complicated tasks like HPE to be conducted on RF-based data [30, 19, 18, 4]. Zhao *et al.* [30] utilize WiFi-ranged FMCW signals with the ability to generate the 2D

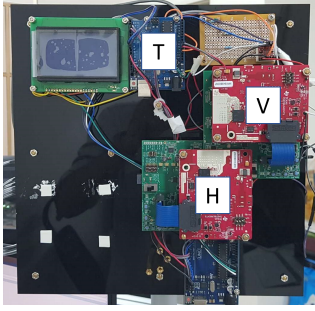


Figure 3: Our mmWave radar sensor board setup. H: Horizontal Radar Module, V: Vertical Radar Module, T: Trigger Generator

human skeletons through walls, which exhibits the potential of privacy invasion. Sengupta *et al.* [19] use millimeter waves (mmWave) radars to collect the point clouds which are used for predicting 3D human skeletons. However, the edge keypoints like wrists are not considered.

2.3. Graph Convolutional Networks (GCN) in HPE

Human skeletons inherently preserve graphical characteristics with its predefined relationship between keypoints. Many GCN-based solutions [2, 29, 25, 17] therefore well leverage this structural information and achieve a great success on HPE. Cai *et al.* [2] exploit spatio-temporal information to construct 3D poses by feeding 2D poses into a GCN. Wang *et al.* [25] refine several sampled human keypoint maps based on guided keypoints to locate the corresponding features. Qiu *et al.* [17], similar to [25], use predicted 2D keypoint coordinates to locate the features in the feature extractor for each keypoint, followed by a GCN to predict 3D keypoints. Our work directly utilizes the keypoint heatmaps as the node features, instead of the features associated with the keypoints, and then refines these heatmaps by a GCN.

3. Dataset

3.1. Sensor Setup

We use 2 identical radar sensors, IWR1843BOOST by Texas Instruments [8]. The radar sensors can provide the 3D information (range, azimuth, and elevation) of the scene; however, the resolution of elevation angle is extremely low (only 2 bins). Therefore, one radar sensor is rotated by 90° to the other radar sensor on the antenna plane. This way we have one radar sensor that focuses on the horizontal plane (range-azimuth-elevation) and another focuses on the vertical plane (range-elevation-azimuth). The physical setup of our radar sensors is shown in Fig. 3, and an illustration of data acquiring process is shown in Fig. 1. Both radar sensors are fixed on a solid plane and kept as near as 3cm from each other. One RGB camera is fixed at the same plane. RGB frames from the camera are used to generate the 2D

Table 1: Settings and properties of our radar sensors

Parameters	Values
Physical Transmitters	4
Physical Receivers	3
Frequency Slope	60.012 MHz
Sample Rate	4400 ksp/s
Chirp Duration	72 μ s
Chirps in Frame	64
ADC Samples	256
Frame Per Second	10

ground-truth keypoints based on a pre-trained image-based 2D human pose estimation network HRNet [22].

for the IWR1843BOOST module the maximum range is set to 11m with the range resolution of 4.8cm, the azimuth angle of 120° with the resolution of 30° , and an elevation angle of 30° with the resolution of 15° . To increase the azimuth resolution, we configure the radar sensor in 1 virtual transmitter antenna and 8 virtual receiver antennas, which in turn increase the azimuth resolution by a factor of 2, resulting in the final azimuth resolution of 15° . Table 1 shows the parameter settings of our radar.

3.2. Dataset Collection

We collected 235 sequences of data in an indoor environment, with each sequence being one-minute long and totally about 4 hour-long video data. Some samples in the dataset are shown in Fig 2. Each sequence has an RGB camera frames, the horizontal radar frames, and the vertical radar frames. The two radars and the camera are synchronously configured to capture 10 frames per second (FPS), and hence each sequence has 600 triplets of camera-radar-radar frames. In total, we have 141,000 triplets in 235 sequences. The resolution of camera frames is 256×256 and the dimensions of raw radar data is $256 \times 64 \times 8 \times 2$, corresponding to analog-to-digital converter (ADC) samples, chirps, azimuth bins, and elevation bins, respectively. Specifically, we select a specific range of the ADC samples to just process signals in a shorter range since human subjects will always perform actions close but not too closed to the sensors and use zero paddings to increase the size of the azimuth and elevation dimension before the pre-processing procedure. Overall, we obtain the raw radar data with the dimensions of $64 \times 64 \times 64 \times 8$. There is always only one person in the scene. The person performs 3 types of movements: static actions, standing and waving hand(s), and walking with waving hand(s). Static actions mean the subject performs a pose without moving for at least 10 seconds. Six subjects are involved in the dataset and the consent of each subject has been obtained by a face-to-face interview. One of them only appears in the test set for fair evaluation.

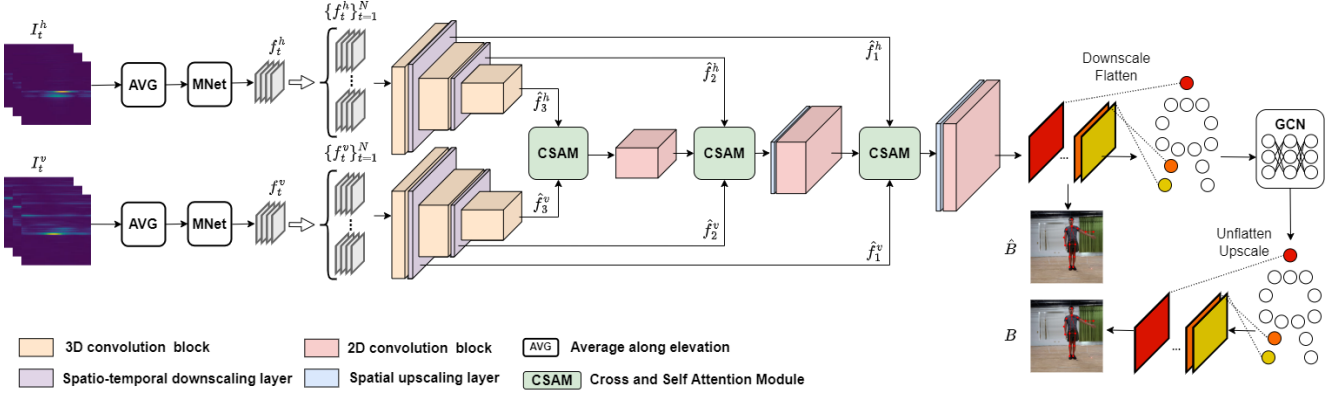


Figure 4: Architecture of our proposed method.

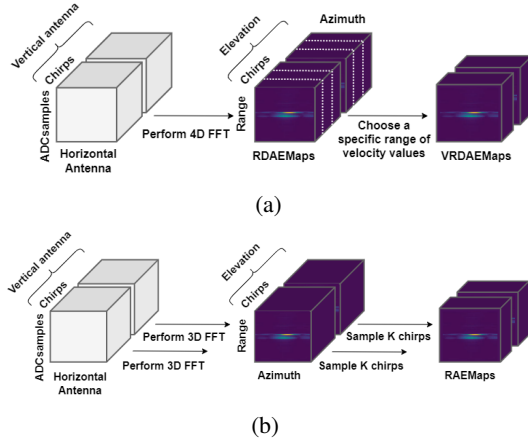


Figure 5: Comparison of radar pre-processing methods: (a) our proposed method and (b) traditional schemes.

4. Proposed Method

This work addresses the task of 2D human pose estimation using mmWave radars. Given two input sequences of length N consisting of the two sets of radar frames $\{I_t^h\}_{t=1}^N$ and $\{I_t^v\}_{t=1}^N$ (to be defined in Section 4.1) from the horizontal and the vertical radar sensors, where $I_t^h, I_t^v \in \mathbb{R}^{2 \times K \times H \times W \times E}$, our task is to predict the 2D keypoint heatmaps $B \in \mathbb{R}^{C \times H \times W}$ for the scene at time instance $t = \frac{N}{2}$. Each keypoint's heatmap $B_c, c \in C$ represents the confidence distribution of the keypoint c . In particular, we take a sliding window approach to include radar frames from not only the past but also the future when the prediction proceeds from one time instance to the next. Fig. 4 shows the overall pipeline of our proposed framework. The two input sequences $\{I_t^h\}_{t=1}^N$ and $\{I_t^v\}_{t=1}^N$ are first processed by a Multi-Scale Spatio-Temporal Feature Fusion module to generate the 2D keypoint heatmaps $\hat{B} \in \mathbb{R}^{C \times H \times W}$. The heatmaps \hat{B} are then refined by Pose Refinement Graph

Convolutional Networks (Section 4.3) for pose confidence refinement to produce the final heatmap B .

4.1. Radar Signal Pre-processing

Fig. 5a depicts our proposed radar pre-processing method for each radar frame, I_t^h or I_t^v . Our method is designed to additionally extract velocity information as an important cue for predicting human pose. With the traditional scheme (Fig. 5b), the range-azimuth-elevation maps (RAEMap) are often generated by performing Fast Fourier transform (FFT) first along the dimension of digitized chirp samples (i.e. ADCsamples), followed by applying the other two FFT's to co-located samples along the horizontal and vertical antenna dimensions of the 3D raw input data tensor to get range, azimuth, and elevation information. The final RAEMaps are formulated by sampling uniformly K chirps along the chirp dimension. Some of the previous methods [26, 9] adopt only range-azimuth maps (RAMap) without elevation information. In comparison to the traditional scheme, our proposed pre-processing method generates the velocity-specific range-doppler-azimuth-elevation map (VRDAEMap) representation by additionally performing FFT along the chirp dimension, to extract the doppler velocity information. Specifically, we first perform 4D FFT on the raw data along all four dimensions of ADCsamples, chirps, horizontal antenna, and vertical antenna to obtain the range-doppler-azimuth-elevation map (RDAEMap). Instead of directly using the RDAEMap, which is sparse and inefficient to process, we choose a specific range of the velocity values (named VRDAEMap) from $-\frac{K}{2}$ to $\frac{K}{2}$ relative to the radar sensors as our input I_t^h, I_t^v with the dimensions of $2 \times K \times H \times W \times E$, where 2 represents the real and imaginary parts of the FFT coefficients. The values of H , W , and E are 64, 64, 8, respectively. We empirically set K to 8 to account for the moderately slow action sequences in our applications.

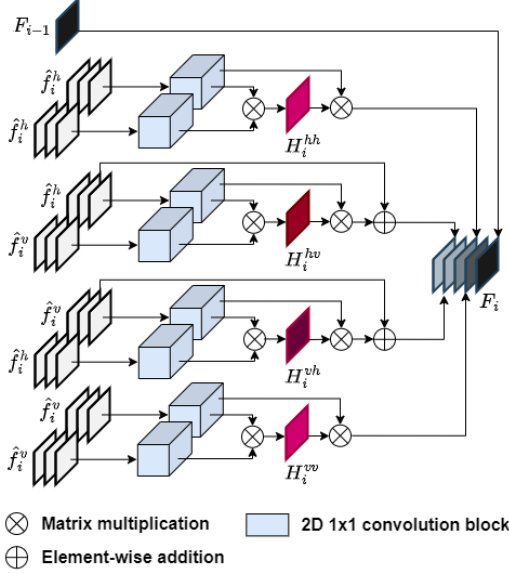


Figure 6: Architecture of Cross and Self Attention Module.

4.2. Multi-Scale Spatio-Temporal Radar Feature Fusion with Cross- and Self-Attention

Our multi-scale spatio-temporal feature fusion is to fuse information captured by the horizontal and vertical radars, in generating the heatmaps of 2D keypoints. To this end, we first process two sequences of the VRDAEMaps $\{I_t^h\}_{t=1}^N$ and $\{I_t^v\}_{t=1}^N$ by two distinct branches yet with the same network structure. As shown in the left part of Fig. 4, each branch begins with an averaging (AVG) function that pools information along the elevation dimension (dimension E) by taking the average of co-located feature samples. The intermediate result of dimension $2 \times K \times H \times W$ is processed further by a fusion network, MNet [26], which fuses the information along the velocity dimension (dimension K) using a convolutional layer with a kernel size $\frac{K}{2} \times 1 \times 1$ and a MaxPooling layer of stride $\frac{K}{2}$ along the velocity dimension. At the output of MNet [26], the number of feature channels is set to D . We therefore get two sets of fused feature maps $\{f_t^h\}_{t=1}^N$ and $\{f_t^v\}_{t=1}^N$, where $f_t^v, f_t^h \in \mathbb{R}^{D \times H \times W}$.

Multi-Scale Spatio-temporal Feature Fusion: The middle part of Fig. 4 illustrates our multi-scale spatio-temporal feature fusion. We adopt multiple 3D convolutional layers with strides to aggregate information with respect to both the spatial and the temporal dimensions. The output of each spatial scale is followed by a 3D convolutional layer with a kernel size $\frac{N}{2^{i-1}} \times 1 \times 1$ to aggregate the remaining temporal information, resulting in the encoding features \hat{f}_i^v, \hat{f}_i^h , where i is the scale index and $\hat{f}_i^v, \hat{f}_i^h \in \mathbb{R}^{2^i \cdot D \times \frac{H}{2^{i-1}} \times \frac{W}{2^{i-1}}}$. The encoding features of each scale are processed by a Cross- and Self-Attention module to generate the decoding features. We then use mul-

Table 2: Comparison of pre-processing methods in AP . Total denotes the average precision over 14 keypoints.

Pre-processing	Model	Elbow	Wrist	Total
RAEMap	RF-Pose [30]	13.6	6.0	40.6
RAEMap	Ours	40.9	17.4	61.6
VRDAEMap	Ours	45.6	23.5	64.3

tipple 2D convolutional layers, followed by a sigmoid function, to generate 2D keypoint heatmaps \hat{B} .

Cross- and Self-Attention Module: Our cross- and self-attention module (CSAM) is to discover useful context information for human pose detection across and within the radar frames captured by the horizontal and vertical radars. It performs cross-attention and self-attention at different scales of the horizontal and vertical radar features. We first describe the cross-attention mechanism (the central part of Fig. 6). It begins by applying 2D 1x1 convolution to \hat{f}_i^h and \hat{f}_i^v independently to generate the projected features. The resulting features are then flattened along the spatial dimensions to obtain the attention maps H_i^{hv}, H_i^{vh} , where $H_i^{hv} \in \mathbb{R}^{\frac{HW}{2^{i-1}} \times \frac{HW}{2^{i-1}}} \times \frac{HW}{2^{i-1} \times 2^{i-1}}$ (respectively, H_i^{vh}) indicates the use of \hat{f}_i^v (respectively, \hat{f}_i^h) as the guiding signal to formulate the attention map with \hat{f}_i^h (respectively, \hat{f}_i^v) serving as the reference signal. Finally, \hat{f}_i^h is updated (respectively, \hat{f}_i^v) by performing cross-attention with H_i^{hv} (respectively, H_i^{vh}) to generate the cross-attended residual features. Note that the width (dimension W) of horizontal and vertical features represent different meanings (azimuth and elevation information, respectively). Directly relating two features may make the training process unstable. We therefore add a residual connection at the end of the layer to stabilize the whole training process. The self-attention mechanism (the top and bottom branches of Fig. 6) acts in a way similar to the cross-attention one. Two major distinctions are that both the guiding and reference signals come from the same \hat{f}_i^h or \hat{f}_i^v , and that there is no skip connection to generate the self-attended residual features. Finally, we concatenate all the output features, generating F_i as the input to the convolution blocks and CSAM of the next scale.

4.3. Pose Refinement Graph Convolutional Networks (PRGCN)

The locations of different body keypoints are strongly related to each other. That is, knowing the locations of some keypoints is able to help predict the locations of the others. This motivates us to develop a pose refinement module based on GCN to refine the initial keypoint heatmaps \hat{B} . We first define the graph structure. Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be a graph, where $\mathcal{V} = \{v_1, \dots, v_C\}$ is a collection of nodes, each corresponding to a keypoint to be estimated. $\mathcal{E} = \{A_{i,j} | v_i, v_j \in \mathcal{V}\}$ specifies edges connecting nodes v_i

Table 3: Comparison of baseline and the variants of our proposed method. CSAM denotes that we train the model with cross-attention (CA) + and self-attention (SA) module. † indicates the results are reproduced by us.

Model	AP								AP	AP^{50}	AP^{75}
	Head	Neck	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	Total		
RF-Pose† [30]	61.0	65.3	52.5	16.1	6.3	73.5	65.7	62.0	41.4	82.9	37.0
Ours	71.3	76.7	64.4	36.9	18.8	84.8	78.1	70.4	57.6	96.5	64.1
Ours (CA)	79.6	83.8	73.3	39.3	16.7	84.3	80.6	73.1	60.0	96.8	69.7
Ours (SA)	76.4	81.6	71.4	42.1	18.7	87.2	81.7	72.4	61.3	96.9	71.2
Ours (CSAM)	74.4	80.8	69.0	42.1	17.4	87.2	82.4	75.6	61.9	97.3	73.1
Ours (CSAM + PRGCN)	77.5	81.9	70.3	45.5	22.3	88.1	82.2	73.1	63.4	97.0	74.0

Table 4: Comparison of 3D keypoint performance based on MPJPE in millimeters. Ours + VideoPose3D means that we adopt our proposed method to generate 2D keypoints, which are lifted to 3D by VideoPose3D.

Model	Elbow	Wrist	Total
mmMesh [28]	112.9	218.2	71.3
Ours + VideoPose3D [16]	85.3	156.4	68.2

and v_j with the weight $A_{i,j} = 1$ if the two body keypoints i, j are physically connected or 0 otherwise. In our task, $v_i \in \mathbb{R}^{\frac{H \times W}{4}}$ represents the downsampled heatmap of the i -th keypoint. It is obtained by downscaling and flattening the final features $\hat{f} \in \mathbb{R}^{C \times H \times W}$ (see Fig. 4) along the spatial dimensions into a total of C 1-dimensional vectors, with each being a node feature $v_i, i \in C$. To generate the refined keypoint heatmaps B , we perform the feature propagation and the inference on graph \mathcal{G} through a 3-layer GCN. The forward pass proceeds as follows:

$$\begin{aligned} V' &= \sigma_r(\hat{A}VW_1), \\ V'' &= \sigma_r(\hat{A}V'W_2), \\ H^g &= \sigma_s(\hat{A}V''W_3) \end{aligned} \quad (1)$$

where $V = [v_1, \dots, v_C]^T \in \mathbb{R}^{C \times \frac{H \times W}{4}}$ is a matrix of node features; $W_1, W_2, W_3 \in \mathbb{R}^{\frac{H \times W}{4} \times \frac{H \times W}{4}}$ are learnable parameters; $\sigma_r(\cdot), \sigma_s(\cdot)$ are ReLU and sigmoid functions, respectively; and $\hat{A} = A + I$ is the sum of the adjacency matrix $A \in \mathbb{R}^{C \times C}$ and an identity matrix I . The refined output B is un-flattened back to heatmaps with dimensions $\frac{H}{2} \times \frac{W}{2}$, followed by upscaling. Note that the previous GCN-based approach [29] produces the node features by selecting the co-located feature samples of intermediate features at the coordinates of keypoints. Our work, however, does not follow the same feature selection process. The reason is that the radar signals and the keypoint heatmaps have different coordinate systems. We are unable to map radar features using keypoint coordinates. To train the model end-to-end, we impose pixel-wise binary cross-entropy on both

the initial keypoint heatmaps \hat{B} and the GCN-refined keypoint heatmaps B . Regularizing \hat{B} makes the training of B more stable since its inputs depend completely on \hat{B} . To be specific, the objective function is given by

$$\mathcal{L} = \mathcal{L}_{bce}(\hat{B}, T) + \alpha \cdot \mathcal{L}_{bce}(B, T) \quad (2)$$

where α is a hyper-parameter and is set to 1 in our implementation. Taking $\mathcal{L}_{bce}(B, T)$ as an example, the binary cross entropy loss is defined as:

$$\begin{aligned} \mathcal{L}_{bce}(B, T) &= \\ &= - \sum_{c,i,j} T_{c,i,j} \log(B_{c,i,j}) + (1 - T_{c,i,j}) \log(1 - B_{c,i,j}) \end{aligned} \quad (3)$$

where $T \in \mathbb{R}^{C \times H \times W}$ is the heatmaps of ground-truth keypoints generated with HRNet [22]. Note that the heatmap of each keypoint is a Gaussian distribution with mean at the ground-truth keypoint coordinates and a pre-determined variance.

5. Experimental Results

5.1. Setup and Evaluation Metrics

We train the proposed system on our HuPR dataset, which contains total 235 video sequences, each being 1-minute long. We take 193 video sequences for training, 21 for validation, and 21 for test. For fair comparison, we only choose RF-Pose [30] with RF-based input data as our heatmap-based 2D HPE baseline since there are only a handful of prior works, most of which do not even release nor provide hardware settings and training data. It is challenging to reproduce their results. We additionally include mmMesh [28], which is a pointcloud-based 3D mesh prediction network. We modified it to 3D HPE network, to see how our RF-based scheme performs as compared to the pointcloud-based scheme. Following common practice [30, 12], we adopt average precision over different object keypoint similarity (OKS) as the performance metric of 2D keypoints. Specifically, we evaluate the model over 14 keypoints, including head, neck, shoulders, elbows, wrists,

hips, knees, and ankles. We present three variants of average precision AP^{50} , AP^{75} , and AP . The value of 50 and 75 indicate the loose and strict constraints of OKS. AP denotes the average precision over 10 OKS thresholds, namely, 0.5, 0.55, ..., 0.9, 0.95. Following [16], we report mean per-joint position error (MPJPE) in millimeters as the evaluation metric of 3D keypoints.

5.2. Implementation Details

We implement our pre-processing method and the proposed framework with Python and Pytorch. The 2D ground-truth keypoints for training video sequences are generated from the RGB frames by the state-of-the-art 2D human pose estimation network, HRNet [22], which is pre-trained on MPII dataset [1]. We also choose the keypoints generated by HRNet as the 2D pseudo ground-truths for our test set. We remark that our dataset contains mostly simple actions under a single scene. A side experiment shows that HRNet achieves an AP of 99%. In other words, HRNet provides satisfactorily good ground-truths under our application scenarios. Likewise, to generate the 3D ground-truths of the training and test sets, we choose VideoPose3D [16]. Since our method generates only 2D keypoints, we adopt the pre-trained network of VideoPose3D [16] to convert them into 3D keypoints. We use the Adam optimizer with a learning rate 0.0001, which is decreased by a factor of 0.999 for every 2000 iterations. We set the batch size to 20 and the weight decay to 0.0001. The number (N) of input frames is chosen to be 8 frames (0.8 seconds) for our method, and 24 frames (2.4 seconds) for RF-Pose [30] which shows the best results on our dataset. Because the code of RF-pose [30] is unavailable, we implement it by following the network structure reported in the paper. In addition, we train the model from mmMesh [28] from scratch using their official code. In particular, we replace the mesh reconstruction module with MLPs to serve our purpose of predicting 3D keypoints. Each convolutional block in Fig. 4 contains two basic blocks from ResNet [7]. The activation functions used in the 3D and 2D blocks are ReLU and PReLU, respectively. Batch normalization is not included in the 2D blocks. In these basic blocks, the 2D convolutional layer has a kernel size 3×3 , and the 3D one has a kernel size $3 \times 3 \times 3$. We use a single Tesla V100 to train the networks.

5.3. Quantitative Results

Comparison of radar pre-processing methods: Table 2 shows the results for two pre-processing methods. These methods are trained on a simplified dataset, containing 80 video sequences for training, 10 for validation, and 10 for test. Comparing with the RAEMap, the traditional method, our proposed VRDAEMap achieves higher AP , especially on difficult keypoints like wrists and elbows. We find that RF-Pose [30] shows relatively poor overall

Table 5: Comparison of different range of velocity values.

K	Elbow	AP	
		Wrist	Total
2	33.9	13.7	53.1
4	36.7	14.6	58.9
8	45.5	22.3	63.4
16	39.2	20.4	60.6

accuracy when compared with RAEMap (Ours). As RF-Pose [30] does not consider the chirp or velocity information in the Wi-Fi data, which plays an important role in the radar data, it is unable to take full advantage of the temporal information despite using a longer sequence of frames.

Comparison of methods: Table 3 compares the values of average precision of different methods. Our method (CSAM + PRGCN) outperforms RF-Pose [30] in all AP values. It also attains 5.8% and 9.9% gain in terms of AP and AP^{75} , as compared to our baseline method (Ours). The model "Ours" denotes that we only perform concatenation of features of different scales from two radars without CSAM and PRGCN for refinement. The increase in AP^{75} , which is a stricter metric, indicates that our proposed method (CSAM + PRGCN) is able to predict the keypoints more precisely. We further analyze the results of the variants of our method in Section 5.4. Table 4 further compares the 3D keypoint detection of our method to mmMesh [28], i.e. the pointcloud-based state-of-the-art method. It is seen that our method outperforms mmMesh [28], achieving 27.6, 61.8, and 3.1 less MPJPE of elbows, wrists, and total, respectively. Note that the results for mmMesh [28] are produced by training their model end-to-end to predict 3D keypoints. In comparison, our scheme predicts 3D keypoints in 2 sequential steps (i.e. in a non-end-to-end optimized manner): the generation of 2D keypoints followed by converting them into 3D keypoints using VideoPose3d [16]. We clarify that as the network of VideoPose3D [16] is pre-trained on another dataset, it possesses the ability to improve the performance of 3D HPE by lifting the 2D predictions. In other words, it contributes partly to our superior performance which makes the comparison unfair to some extent.

5.4. Ablation Experiments

Component breakdown: We investigate the effect of each component in our proposed method. It is seen in Table 3 that with cross-attention only (CA), self-attention only (SA), or cross- and self-attention (CSAM), the predictions related to the torso, e.g. head, neck, and shoulders, become more stable. Moreover, adopting PRGCN further improves the detection performance of wrists and elbows (i.e. fast-moving keypoints) up to 4.9% and 3.4% respectively. The reason may be attributed to that the pre-defined adjacency matrix propagates only the neighboring features to refine

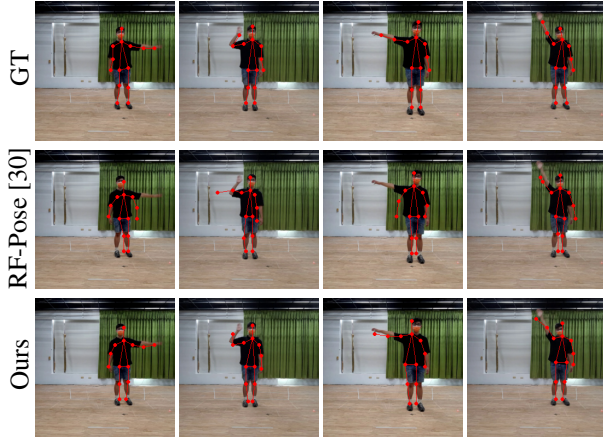


Figure 7: Comparison of 2D qualitative results. GT: Ground-truth keypoints generated by HRNet [22]. Ours: The predicted results generated by our method (CSAM + PRGCN).

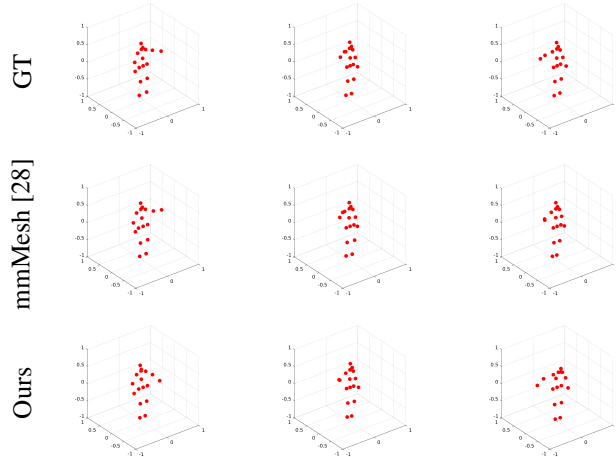


Figure 8: Comparison of 3D qualitative results. GT: Ground-truth keypoints generated by VideoPose3D [16]. Ours: The results generated by VideoPose3D based on our 2D predicted results.

the final confidence output.

Velocity range: We study how the velocity information derived from radar signals affect the performance of our pre-processing scheme. Table 5 shows that increasing the velocity range, as specified by K in Section 4.1, is able to capture the movement of the edge keypoints more precisely. However, the larger velocity range makes the VR-DAMaps become larger and sparser, leading to increased network complexity and memory footprint. In particular, when K goes beyond 8, the performance declines. This may be because an extremely large K value involves more noise rather than useful velocity information.

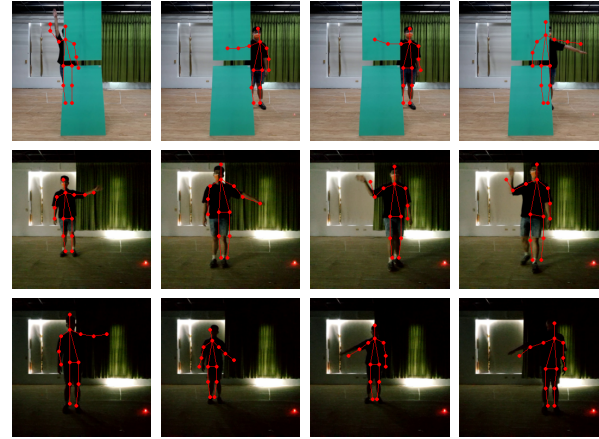


Figure 9: Our results under different conditions. Top row: obstacles in front of the person. Middle row: low-light condition. Bottom row: dark condition.

5.5. Qualitative Results

Fig. 7 and Fig. 8 demonstrate the 2D and 3D qualitative results of different types of actions. Our predicted keypoints mostly follow the ground-truths. It is worth noting that the ground-truth keypoints from HRNet [22] precisely follow the actions in the RGB frames, demonstrating a quality level similar to manual annotations. We also test our method on an extra test set under conditions of obstacle, low light, and no light (Fig. 9). Our proposed method produces promising results under these conditions, showing that radar signals and our method are robust to adverse lighting and visual blocking conditions.

6. Conclusion

The paper introduces a novel human pose estimation scheme for mmWave radar sensing inputs, develops a radar pre-processing method, and proposes CSAM for multi-scale radar feature fusion along with PRGCN for pose confidence refinement. Our major findings and limitations include (1) the additionally extracted velocity information is critical to HPE; (2) our proposed CSAM and PRGCN improve the performance of 2D and 3D HPE as compared to the state-of-the-art RF-based and pointcloud-based methods; (3) the predicted results by taking radar signals as the input are qualitatively robust to dark, low-light, and visually blocked conditions; (4) the low azimuth and elevation angle resolution due to the limited number of antennas on the radar module [8] result in the ambiguity between some actions, e.g., waving hands and raising hands. Lastly, the potential of this line of research is very promising. Further investigation, e.g., extending our system to multi-person 2D, 3D HPE with more complex poses, is highly encouraged.

References

- [1] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014.
- [2] Yujun Cai, Liuhao Ge, Jun Liu, Jianfei Cai, Tat-Jen Cham, Junsong Yuan, and Nadia Magnenat Thalmann. Exploiting spatial-temporal relationships for 3d pose estimation via graph convolutional networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.
- [3] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.
- [4] Han Cui and Naim Dahoun. Real-time short-range human posture estimation using mmwave radars and neural networks. *IEEE Sensors Journal*, 2022.
- [5] Lijie Fan, Tianhong Li, Rongyao Fang, Rumen Hristov, Yuan Yuan, and Dina Katabi. Learning longterm representations for person re-identification using radio signals. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.
- [6] Kaiming He, Georgia Gkioxari, Piotr Dollar, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2017.
- [7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.
- [8] Texas instruments.
- [9] Guangzheng Li, Ze Zhang, Hanmei Yang, Jin Pan, Dayin Chen, and Jin Zhang. Capturing human pose using mmwave radar. In *2020 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops)*, March 2020.
- [10] Jiefeng Li, Siyuan Bian, Ailing Zeng, Can Wang, Bo Pang, Wentao Liu, and Cewu Lu. Human pose regression with residual log-likelihood estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2021.
- [11] Jia Li, Wen Su, and Zengfu Wang. Simple pose: Rethinking and improving a bottom-up approach for multi-person pose estimation. In *Proceedings of the Thirty-Fourth Conference on Association for the Advancement of Artificial Intelligence (AAAI)*, February 2020.
- [12] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *European Conference on Computer Vision (ECCV)*, September 2014.
- [13] Xiaorui Ma, Yunong Zhao, Liang Zhang, Qinghua Gao, Miao Pan, and Jie Wang. Practical device-free gesture recognition using wifi signals based on metalearning. In *IEEE International Conference on Machine Learning and Applications (ICMLA)*, December 2018.
- [14] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *European Conference on Computer Vision (ECCV)*, October 2016.
- [15] Xuecheng Nie, Jianfeng Zhang, Shuicheng Yan, and Jiashi Feng. Single-stage multi-person pose machines. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.
- [16] Dario Pavlo, Christoph Feichtenhofer, David Grangier, and Michael Auli. 3d human pose estimation in video with temporal convolutions and semi-supervised training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.
- [17] Zhongwei Qiu, Kai Qiu, Jianlong Fu, and Dongmei Fu. Dgcn: Dynamic graph convolutional network for efficient multi-person pose estimation. In *Proceedings of the Thirty-Fourth Conference on Association for the Advancement of Artificial Intelligence (AAAI)*, February 2020.
- [18] Arindam Sengupta and Siyang Cao. mmpose-nlp: A natural language processing approach to precise skeletal pose estimation using mmwave radars. *IEEE Transactions on Neural Networks and Learning Systems*, 2022.
- [19] Arindam Sengupta, Feng Jin, Renyuan Zhang, and Siyang Cao. mm-pose: Real-time human skeletal posture estimation using mmwave radars and cnns. *IEEE Sensors Journal*, May 2020.
- [20] Akash Deep Singh, Sandeep Singh Sandha, Luis Garcia, and Mani Srivastava. Radhar: Human activity recognition from point clouds generated through a millimeter-wave radar. In *Proceedings of the 3rd ACM Workshop on Millimeter-Wave Networks and Sensing Systems*, October 2019.
- [21] Akash Deep Singh, Sandeep Singh Sandha, Luis Garcia, and Mani Srivastava. Radhar: Human activity recognition from point clouds generated through a millimeter-wave radar. In *Proceedings of the 3rd ACM Workshop on Millimeter-Wave Networks and Sensing Systems*, October 2019.
- [22] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.
- [23] Fei Wang, Sanping Zhou, Stanislav Panev, Jinsong Han, and Dong Huang. Person-in-wifi: Fine-grained person perception using wifi. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.
- [24] Jie Wang, Qinghua Gao, Xiaorui Ma, Yunong Zhao, and Yuguang Fang. Learning to sense: Deep learning for wireless sensing with less training efforts. In *IEEE Wireless Communications*, August 2020.
- [25] Jian Wang, Xiang Long, Yuan Gao, Errui Ding, and Shilei Wen. Graph-pcnn: Two stage human pose estimation with graph pose refinement. In *European Conference on Computer Vision (ECCV)*, August 2020.
- [26] Yizhou Wang, Zhongyu Jiang, Xiangyu Gao, Jenq-Neng Hwang, Guanbin Xing, and Hui Liu. Rodnet: Radar object

- detection using cross-modal supervision. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, January 2021.
- [27] Fangyun Wei, Xiao Sun, Hongyang Li, Jingdong Wang, and Stephen Lin. Point-set anchors for object detection, instance segmentation and pose estimation. In *European Conference on Computer Vision (ECCV)*, August 2020.
 - [28] Hongfei Xue, Yan Ju, Chenglin Miao, Yijiang Wang, Shiyang Wang, Aidong Zhang, and Lu Su. mmmesh: Towards 3d real-time dynamic human mesh construction using millimeter-wave. In *Proceedings of the 19th Annual International Conference on Mobile Systems, Applications, and Services (MobiSys)*, June 2021.
 - [29] Long Zhao, Xi Peng, Yu Tian, Mubbasir Kapadia, and Dimitris N. Metaxas. Semantic graph convolutional networks for 3d human pose regression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.
 - [30] Mingmin Zhao, Tianhong Li, Mohammad Abu Alsheikh, Yonglong Tian, Hang Zhao, Antonio Torralba, and Dina Katabi. Through-wall human pose estimation using radio signals. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.