MVDoppler-Pose: Multi-Modal Multi-View mmWave Sensing for Long-Distance Self-Occluded Human Walking Pose Estimation

Jaeho Choi[1,2] Soheil Hor[2] Shubo Yang[2] Amin Arbabian[2]
[1]DGIST [2]Stanford University
jhochoi@dgist.ac.kr, {soheilh, shuboy, arbabian}@stanford.edu
https://mvdoppler-pose.github.io/

Abstract

One of the main challenges in reliable camera-based 3D pose estimation for walking subjects is to deal with self-occlusions, especially in the case of using low-resolution cameras or at longer distance scenarios. In recent years, millimeter-wave (mmWave) radar has emerged as a promising alternative, offering inherent resilience to the effect of occlusions and distance variations. However, mmWave-based human walking pose estimation (HWPE) is still in the nascent development stages, primarily due to its unique set of practical challenges including the quality of the observed radar signal dependent on the subject's motion direction. This paper introduces the first comprehensive study comparing mmWave radar to camera systems for HWPE, highlighting its utility for distance-agnostic and occlusion-resilient pose estimation. Building upon mmWave's unique advantages, we address its intrinsic directionality issue through a new approach—the synergetic integration of multi-modal, multi-view mmWave signals, achieving robust HWPE against variations both in distance and walking direction. Extensive experiments on a newly curated dataset not only demonstrate the superior potential of mmWave technology over traditional camera-based HWPE systems, but also validate the effectiveness of our approach in overcoming the core limitations of mmWave HWPE.

1. Introduction

3D human pose estimation (HPE) aims to reconstruct the 3D coordinates of the human body, acting as a key component for diverse applications such as human activity/gait monitoring, clinical rehabilitation, and autonomous driving [2, 5, 17–19, 31, 38, 39, 43]. Typical approaches in 3D HPE primarily rely on vision systems, which involve one-stage [11, 27] or two-stage mapping [15, 42] from input videos into 3D joint coordinates. These systems, however, face critical challenges when applied to walking subjects, *i.e.* human walking pose estimation (HWPE), as the diverse trajectories of walking individuals in a broad area can further complicate reliable data capture of joints. One of the key challenges in camera-based HWPE systems is self-occlusion, where movements of subjects across various directions often cause their joints to be occluded from camera by other body parts. The problem is even more pronounced in scenarios utilizing low-resolution cameras or capturing subjects from longer distances, where reduced effective resolution further exacerbates the impact of occlusion.

Alternatively, millimeter-wave (mmWave) radar systems offer a fundamental solution to the challenges faced by camera-based HWPE. Utilizing transmission and reception of electromagnetic (EM) waves, mmWave radar not only detects the spatial locations of human joints but also captures their velocities *directly* via the Doppler effect [3]. While visible light is easily blocked by occluding objects, mmWave signals can easily bypass and traverse occlusions, identifying whole human figures even when parts of their body are obscured [44]. Moreover, mmWave radar is noted for its resilience against distance-related information loss: such capabilities make the technology particularly valuable in scenarios where visual line-of-sight is compromised, such as in automotive applications where mmWave radars are frequently used to detect distant targets obscured by front vehicles [6, 14, 26].

Despite this promising potential of mmWave technology, its practical efficacy in HWPE still remains underexplored. A significant challenge originates from the 'directional nature' of mmWave signals. Namely, while mmWave's capability to track human joints excels along the radial direction, it shows marked degradation of data quality when detecting movements in lateral directions. Consequently, current mmWave pose estimation models [1, 28, 33, 37, 45] are typically restricted to controlled scenarios (*e.g.* simple limb extension or lunge movements performed in place, or walking only in pre-defined directions within limited spaces), and fail to generalize to more complex, variable motions associated with real-world walking scenarios.

This work introduces the first comprehensive comparison between mmWave radar and camera systems specifically for HWPE, exploring mmWave's unique ability to offer more occlusion-resilient and distance-agnostic pose estimations for general walking subjects. We further enhance the capabilities of mmWave perception for HWPE by addressing its inherent challenges, facilitating robust estimations against both the influence of subject distances and walking directions. Our approach is grounded on two principal insights. First, we demonstrate that the joint use of positional and motional modalities—two distinct signal types derived from a single mmWave sensor—can be mutually beneficial, allowing more reliable HWPE outside controlled settings and restricted region of interests (RoIs). Furthermore, to overcome the challenges posed by the directionality of mmWave signals, we introduce a novel cross-view fusion strategy, enriching our system's ability to accurately capture poses of walking subjects across various orientations and trajectories.

Building upon these two insights, we propose MVDoppler-Pose, a new HWPE model that leverages the fusion of multi-modal and multi-view signals for advanced mmWave HWPE. For synergetic integration across multi-modal and multi-view signals, MVDoppler-Pose strategically decomposes the overall fusion pipeline into multi-layered, dual-stage fusion blocks. Specifically, within the iterative encoding blocks, the positional and motional representations within each sensor are initially combined through cross-modal fusion (CMF), followed by cross-view fusion (CVF) for modality-wise integration between sensors. The overall model is further guided by a cross-domain loss function, promoting an integrated training of positional and motional signatures. Depending on the use of CVF, our MVDoppler-Pose is applicable in both single-view and multi-view setups.

As another contribution of our work, we have expanded the previous MVDoppler dataset[1] [8] for 3D pose estimation task, released as MVDoppler-Pose dataset. This enriched dataset now involves 3D human pose annotations with a comprehensive coverage of diverse locations and walking trajectories of subjects. Relying solely on EM reflections, MVDoppler-Pose with a single-view setup demonstrates relative superiority to camera-based HWPEs in scenarios with occlusion at longer distances, while markedly outperforming the conventional mmWave approaches. More impressively, MVDoppler-Pose with a multi-view setup can fully overcome the challenges of single-view mmWave HWPE, establishing the first distance- and orientation-resilient HWPE across a large area.

In summary, this work has the following contributions:

- We demonstrate that the use of mmWave technology can address the challenges of camera HWPE, particularly in scenarios involving occlusions and long distances.
- Unlike previous mmWave approaches limited to controlled human movement scenarios in restricted space, our MVDoppler-Pose enables HPE even for random walking subjects across a much larger space. [2]
- We release MVDoppler-Pose dataset by expanding the previous MVDoppler dataset [8], which offers a full characterization and benchmarking of sensor-wise HWPE across comprehensive locations and walking directions.

2. Related Works

2.1. Vision-Based Pose Estimation

Contemporary vision-based 3D HWPE can be broadly categorized into two principal approaches: One-stage and two-stage models. One-stage approaches directly map input images or videos into 3D joint coordinates [11, 27, 36]. In contrast, two-stage approaches initially extract 2D coordinates using pre-trained models, and then elevate the 2D pose into 3D space [7, 20], offering significant resource savings compared to one-stage models. To navigate the intrinsic depth ambiguity in monocular video, another strand of methods deploys temporal correlations from neighboring frames. Pavllo *et al*. [23] introduced a temporal fully-convolutional network that models 3D local joints at a specific time frame using the information of adjacent frames together. Further developments leveraged full transformer architectures for 3D HWPE, exemplified with MHFormer [15] and STCFormer [29], which proposed multiple hypotheses and parallelized spatiotemporal encoding, respectively. Zhao *et al*. [42] further enriched the field by incorporating frequency representations for enhanced efficiency. Despite the efficacy of vision-based approaches at close distances, they face performance degradation with increasing distance and restricted viability even in mildly occluded scenarios, struggling to maintain robust HWPE across diverse positions and walking trajectories.

2.2. mmWave-Based Pose Estimation

Positional Approaches. mmWave sensors, capturing the surroundings using EM waves, offer distinctive advantages for 3D HWPE, especially in scenarios where reliance on visual data may be impractical. The primary approach for mmWave 3D HWPE relies on *positional information* obtained from radar signals. The majority of models treat mmWave reflections in a manner akin to LiDAR data, *i.e*. converting raw signals into point clouds and subsequently mapping these point clouds onto joint coordinates utilizing CNN [1, 37] or PointNet [35] architectures. Other methodologies involve direct processing of multi-dimensional raw

[1] https://mvdoppler.github.io/

[2] The comparison of experimental conditions between previous and our approaches is detailed in the supplementary material.

tensors to further enhance performance with the cost of additional computation [12, 33, 40, 41]. Despite the intuitive nature of these positional approaches in mmWave HWPE, they encounter challenges in capturing the detailed motional dynamics of subjects under uncontrolled settings.

Motional Approaches. A major characteristic that distinguishes mmWave radars from other sensor technologies is their Doppler property. Unlike the conventional positional approaches that concentrate on the spatial features of a target, mmWave sensor's Doppler property allows for the extraction of a subject's *motional dynamics* directly through the macro- and micro-velocity spectra [3]. This unique aspect of the mmWave modality motivated Zhou *et al*. [45] to pioneer a Doppler-based model for motional HWPE, establishing a foundational proof-of-concept. Tang *et al*. [28] augmented this concept by combining initial position information with Doppler-derived motional features, enabling enhanced pose tracking over time. However, compared to positional approaches for HWPE, motional approaches typically exhibit less accurate results due to the absence of essential positional information in the Doppler signatures.

3. Preliminary

3.1. mmWave Signal Model

mmWave radar sensors function by periodically emitting EM waves and capturing their scattering from surroundings. To distinguish each EM scattered reflection, a large body of mmWave approaches employ the frequency-modulated continuous-wave (FMCW) technique. Following a series of fundamental pre-processing steps, the received FMCW signal can be represented as the following complex-valued data [4, 9, 10]:

$$s(R,t) = \sum_i \alpha_i \delta\left(R - R_i(t)\right) \exp\left(j\theta_i(t)\right), \qquad (1)$$

where i corresponds to individual EM scattering centers, reflected from various segments of a human body such as the torso, arms, and legs. R_i signifies the radial distance from the mmWave sensor to each respective i-th scatterer, while t denotes the time instant. Note that the representation of the received mmWave signal is characterized as a linear combination of range-delayed impulse pulses, each modulated in both magnitude α and phase θ by the respective scatterers.

3.2. Positional Measurement from mmWave Signals

As demonstrated in Eq. (1), mmWave radar operates as a radial sensor, primarily capturing the position of reflective targets through the time-of-arrival of pulses. This mechanism implies that the spatial information about each scatterer (encapsulating each human joint) is projected onto the single radial dimension. This single radial component in mmWave data is often insufficient to separate and localize human joints, necessitating additional information sources. The most prevalent strategy to obtain extra dimension involves the use of a multi-input-multi-output setup, which allows an enhanced lateral (*i.e*. cross-range) dimension through its arrangement of spatially distributed antennas [9, 13].

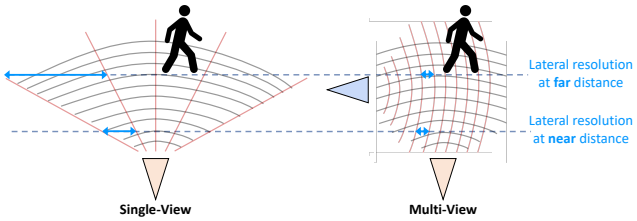

Figure 1. Conceptual figure for radial and lateral sensing in single-view (**left**) or multi-view (**right**) setup. Compared to single-view configuration, multi-view radars can preserve a much finer as well as distance-independent level of lateral resolution.

The principal issue here lies in the large disparity in resolution between the range and cross-range dimensions. This discrepancy in resolution is further pronounced with distance, as visualized on the left side of Fig. 1. For instance, a typical off-the-shelf 3TX-4RX radar system [30] with a 1GHz bandwidth maintains a consistent radial resolution of $\triangle R \approx$ 15cm *regardless of target distance* as long as there is sufficient SNR. However, its cross-range resolution degrades, showcasing $\triangle CR \approx$ 26cm at 1m and an even largely diminished $\triangle CR \approx$ 2.6m at 10m. Given that the standard human stride is around 75cm [21], this acts as a practical bottleneck for HWPE, predominantly restricting the efficacy of current mmWave-based models to scenarios where the individuals perform movements mainly in radial direction (*e.g*. lunge or walking back and forth in front of the sensor) while doing lateral movements (*e.g*. raising arms to the side) only at short distances (<3 m) [1, 37].

3.3. Motional Measurement from mmWave Signals

A mmWave sensor is capable of measuring Doppler signatures of targets in addition to their locations, which differentiates it from other sensing modalities. The Doppler effect in the EM waves occurs as a form of frequency shift corresponding to the relative velocity between the source and the target scatterer [9]. This velocity-dependent shift is modulated into the phase $\theta(t)$ of the received signal.

Given that each joint of the human body generally moves at distinct velocities, such motional measurements can introduce another source of information for localizing human joints. This motional data, like its positional counterpart, also exhibits *inherent resilience to the effect of distance*, underscoring its potential for estimating pose information. Nevertheless, the motional data lacks essential positional context, facing the challenge of complex mapping from the motional domain into the positional joints. Another

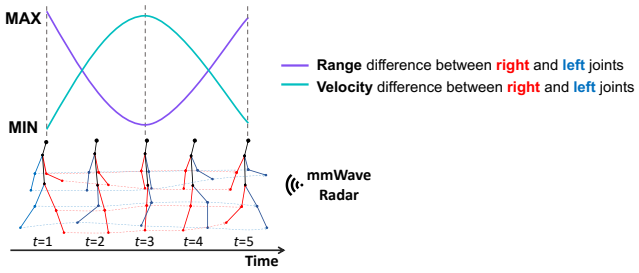

Figure 2. Positional and motional mmWave measurements of a walking person. When the right-left positional difference measured from mmWave radar is minimal ($t = 3$), their motional difference becomes maximal, and vice versa.

complication arises from the directional nature of Doppler measurements: the quality of motional information derived from Doppler shifts is contingent both on the observation angle and the movement direction of the subject. This complexity, in turn, leads to performance degradation when the target's motion is perpendicular to the sensor, forcing the model to function well only with the radial movements [8].

3.4. Motivation

Through examining the dynamics of individual body parts (particularly limb joints) for a moving person, an interesting pattern comes to light: mirroring the cyclical nature of human gaits, both the radial range and velocity of limb joints display a sinusoidal oscillation, marked by a phase difference of approximately $\pi/2$ relative to each other (Fig. 2). This pattern, especially evident in the opposing motions of the right and left joints, unveils a complementary nature between the positional and motional modalities in differentiating and tracking the right and left sides of a human figure. Notably, when the right-left range difference measured from the mmWave sensor is minimal (the hardest case for positional HWPE), their velocity difference becomes maximal (the easiest case for motional HWPE), and vice versa. Motivated by this complementary relationship, instead of relying exclusively on either positional or motional mmWave measurements, we propose a new learning framework based on the fusion of both modalities.

Beyond the cross-modal sensing, we delve into recognizing the potential of multi-view setup to address the domain-specific dependency issues in the '*lateral dimension of radar*'. The essence of this strategy is the introduction of an additional sensor in the lateral direction, whose radial sensing can now cover the cross-range function of the primary sensor. Particularly, considering the *distance-resilient* capability of mmWave signals in the radial direction, this multi-view concept not only overcomes the angle-dependency hurdles associated with positional and motional mmWave measurements, but also allows the full potential of their distance-resilient nature in both radial and lateral dimensions, as depicted in the right side of Fig. 1. This approach consequently allows us to develop the first trajectory-agnostic HWPE over a wide coverage area.

4. Methodology

Building upon our preceding motivation, we leverage dual modalities (*i.e.* positional and motional measurements) captured from multi-view sensors for mmWave HWPE. A notable complexity here is that these four inputs exist independently, each bearing its unique weight and significance in the context of joint tracking. To navigate this complexity, MVDoppler-Pose is designed to trace each mmWave measurement across time and integrate them through a multi-layered, dual-stage fusion strategy. This integration is further supported by a new cross-domain loss function.

Fig. 3 delineates the overall pipeline of our model. It takes time-range (positional) and time-Doppler (motional) data acquired from multi-view mmWave sensors as input, which are then encoded through dedicated blocks. During the iterative encoding pipeline, the CMF module first takes charge of fusing positional and motional representations within each sensor. Following this, CVF modules, specific to each domain, are employed for inter-sensor integration. This encoding and dual-stage fusion sequence is iteratively applied across multiple layers, facilitating a holistic fusion across varying resolution levels.

4.1. Pre-Processing

To form positional inputs spanning the time-range domain, we apply time windowing on the received signals $s(R, t)$, and select only the magnitude component. For motional inputs, we extract Doppler information by applying a short-time Fourier transform along the time axis, representing time-velocity data. These signals are then standardized through uniform resizing, yielding positional and motional mmWave images, as illustrated in Fig. 3.

For effective encoding of each mmWave image, our model adopts the MobileViT backbone [22]. The backbone is characterized by iterative blocks, each structured as a series of convolutional operations and self-attention-based transformer encodings, interspersed with tokenization and de-tokenization phases. This deliberate alternation between convolutional and attention sequences can facilitate the effective use of transformer's capability on global context capturing while securing the local contextual interactions through convolutional processes.

4.2. Multi-Modal, Multi-View Fusion

We note that a sequence of representation tokens is generated during the tokenization and de-tokenization cycles within the MobileViT block. This recognition leads us to strategically apply fusion at the token level within this cycle,

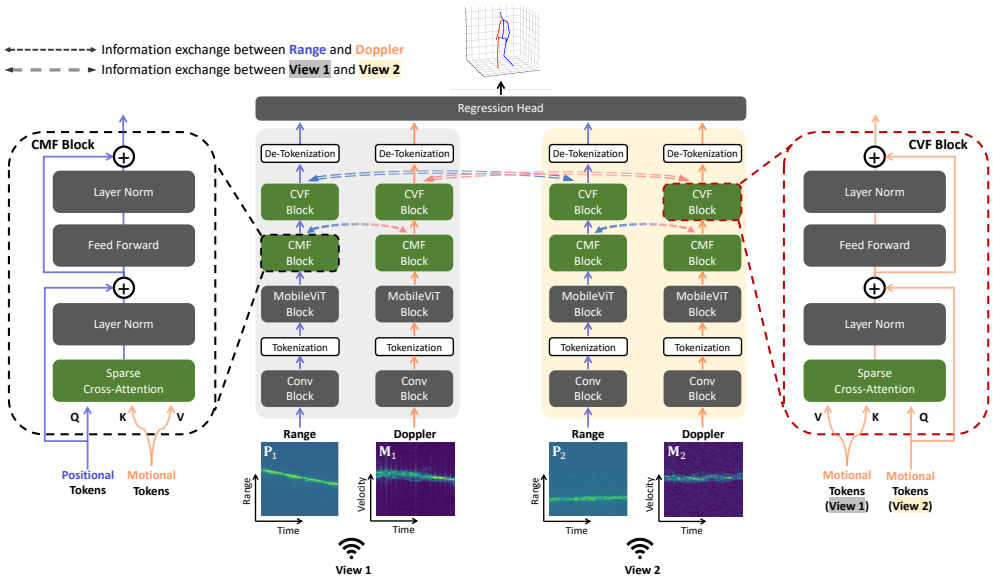

Figure 3. Overall pipeline of the proposed MVDoppler-Pose model.

which not only ensures an efficient deployment of attention-based fusion modules but also fosters a multi-layered fusion across the iterative MobileViT blocks. We further decompose the fusion process into the dual-stage cross-attention mechanism, aiming for a cohesive integration of the four distinct representations. As illustrated in Fig. 3, the initial stage involves the CMF module, where the positional and motional tokens derived from the same sensor are selected as the base query and target, respectively, for cross-modal integration. Subsequently, in the CVF stage, tokens from each modality of cross-view sensors are merged for inter-sensor incorporation. This two-stage fusion pipeline continues throughout iterative encoding blocks to maximize the benefits of multi-layered integration.

Meanwhile, it's noteworthy to state that salient information is often densely concentrated in specific regions in the mmWave images, as evident in the input images of Fig. 3. Accordingly, applying attention indiscriminately across entire patch sequences within such mmWave images can lead to a drastic increase in computational burden as well as an inadvertent intervention of noise elements. To circumvent these challenges, we use sparse attention [25], which selectively correlates only the top-k tokenized mmWave patches from an n-length input sequence ($k \ll n$), ensuring efficient fusion as well as minimizing noise interference.

4.3. Cross-Domain Loss Function

The final 3D pose coordinates $\boldsymbol{p} \in \mathbb{R}^{3\times T_p \times J}$ are obtained by averaging the multi-modal, multi-view representations at the final layer and applying a linear projection, where T_p denotes the temporal frame of the output and J signifies the defined number of joints. To guide the model towards more integrative learning across both positional and motional domains, in harmony with our multi-modal inputs, we introduce the additional employment of cross-domain loss. Extending beyond the conventional mean per joint position error (MPJPE) loss that centers primarily on the positional aspects, inspired by [34], our loss function additionally incorporates the following distance metric newly defined in the motion space:

$$\boldsymbol{m}(t,j) = \boldsymbol{p}(t+\tau, j) \star \boldsymbol{p}(t,j), \tag{2}$$

where j and τ indicate the joint index and the time interval, respectively, and $\star$ represents a cross-product binary operation. Integrating the new motional distance with MPJPE, the final cross-domain loss is designed as

$$\mathcal{L} = \frac{\lambda_p}{T_p J} \sum_{t=1}^{T_p} \sum_{j=1}^{J} \left\| \boldsymbol{p}(t,j) - \boldsymbol{p}^{GT}(t,j) \right\|_2 + \frac{\lambda_m}{T_p \left|\mathbb{T}\right| \left|\mathbb{J}_L\right|} \sum_{\tau \in \mathbb{T}} \sum_{t=1}^{T_p - \tau} \sum_{j \in \mathbb{J}_L} \left(1 - \varphi\left(\boldsymbol{m}(t,j), \boldsymbol{m}^{GT}(t,j)\right)\right), \tag{3}$$

where λ_p and λ_m serve as balancing parameters for positional and motional losses, respectively. $\mathbb{T}$ is the interval set to encompass multiple scales of τ. In particular, leveraging the insight that the motional signatures within a human body are primarily evident in the limb joints and manifest as sinusoidal patterns, the motional component in our loss function is specifically tailored to limb joints $\mathbb{J}_L$, aiming at enhancing Pearson's correlation coefficient (PCC) $\varphi(\cdot)$ between the predicted and ground-truth motional patterns.

Table 1. Quantitative comparison of MVDoppler-Pose-variants against the vision- and mmWave-based models according to different levels of target distance, specifically under self-occluded walking scenarios. Our multi-modal model surpasses the SOTA vision baselines beyond 8.5m in a single-view setup, and demonstrates significantly superior performance across all protocols in a multi-view setup.

Method	**Modality**	y <**8.5m**		**8.5m**$\leq y \leq$**11.5m**		**11.5m**< y	
		MPJPE↓	ρ↑	MPJPE↓	ρ↑	MPJPE↓	ρ↑
Sun *et al*. [27]	RGB	91.03	0.28	114.57	0.18	137.16	0.17
Wei *et al*. [36]	RGB	88.16	0.28	113.97	0.19	133.80	0.17
Li *et al*. [15]	RGB	95.52	0.34	119.75	0.24	141.01	0.21
Zhao *et al*. [42]	RGB	93.34	0.35	116.16	0.24	138.35	0.22
Wang *et al*. [33]	mmW-Position	111.69	0.27	107.86	0.27	109.02	0.26
Zhou *et al*. [45]	mmW-Motion	116.66	0.28	112.95	0.29	112.75	0.29
Tang *et al*. [28]	mmW-Motion	114.24	0.30	109.52	0.31	110.13	0.30
Ours (Single-View)	mmW-Position	108.09	0.27	102.95	0.28	105.93	0.26
Ours (Single-View)	mmW-Motion	112.08	0.30	109.77	0.31	108.62	0.31
Ours (Single-View)	mmW-Multi	96.59	0.33	89.35	0.32	88.32	0.30
Ours (Multi-View)	mmW-Position	70.10	0.47	67.22	0.49	64.83	0.48
Ours (Multi-View)	mmW-Motion	71.57	0.53	67.88	0.52	68.25	0.52
Ours (Multi-View)	mmW-Multi	**63.27**	**0.53**	**59.11**	**0.53**	**58.24**	**0.54**

5. Experiments

5.1. Experiment Design and Evaluation Metrics

To train and evaluate the proposed MVDoppler-Pose model, we have expanded the scope of the previous MVDoppler dataset [8], newly released as MVDoppler-Pose dataset with enhanced annotations for 3D HWPE tasks. The dataset originates from a configuration involving a stereo camera and two cross-view FMCW radars. The capturing setup guarantees a symmetrical coverage of 10m$\times$10m RoI, with each sensor placed 5m away from this area. Note that our RoI is significantly larger compared to those used in previous studies, which introduces greater complexity for HWPE within the expanded space.

Within this extensive coverage space, a total of 13 subjects were instructed to act different types of hand movements, including normal walking, hands in pockets, and texting. Such activities were performed in seven distinct walking patterns (including random walking) across the entire RoI, encompassing *a rich spectrum of* possible walking directions and trajectories. The final paired data spans 6.3 hours. See supplementary for further details about dataset.

Following the protocols in [1, 37], our model is evaluated based on the subject-independent split, where the models are trained on data from 10 subjects and evaluated against the measurements from the remaining 3 subjects. We utilize the MPJPE (*mm* scale) and correlation (ρ) as evaluation metrics, which assess the spatial and temporal quality of the predicted 3D poses, respectively (see supplementary for details). For ablation studies, we also report the percentage of correct keypoints (PCK) with a threshold of 150mm.

5.2. Experimental Results

Comparison with State-of-the-Art. We first compare the proposed MVDoppler-Pose against state-of-the-art (SOTA) vision- and mmWave-based models, specifically under self-occluded scenarios across varying distances. Selecting data instances where subjects move away from the sensor (*i.e*. self-occluded trajectories), the assessments are categorized according to three different distance levels (y <8.5m; 8.5m$\leq$ y $\leq$11.5m; 11.5m< y), which are summarized in Table 1. For the vision baselines, we present the results of end-to-end models leveraging one-stage video encoding [27, 36] and two-stage models [15, 42], each fine-tuned using the synchronized RGB clips from the MVDoppler-Pose dataset. For mmWave baselines, we have reimplemented and retrained existing position- [33] and motion-only [28, 45] models to align with our hardware setup. We also examine the variations of our model utilizing single-view, single-modal inputs.

The comparison illuminates the distinct attributes of each sensor modality. Vision-based models, while effective at closer ranges, start to significantly falter as distance level increases, primarily due to the distance-induced information loss per pixel. Conversely, mmWave models exhibit great resilience to distance variations, maintaining relatively consistent performance across the three distance levels. Meanwhile, the examination comparing different input domains within the mmWave model variants reveals compelling findings as well. The position-only mmWave models generally excel in the MPJPE metric compared to their motion-only counterparts, while motion-only mmWave models display a marked superiority in capturing the temporal coherence of joints compared with their position-only counterparts. This numerically illuminates

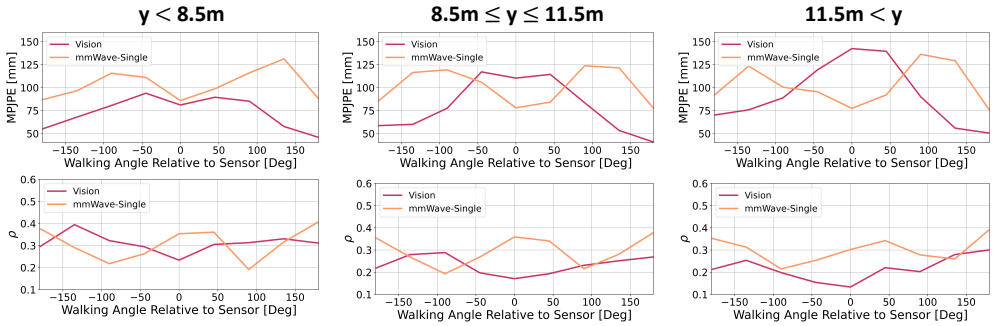

Figure 4. MPJPE (**Top**) and temporal correlation (**Bottom**) of vision and mmWave-single baselines input with respect to walking directions. 0° or 180° represents a person walking away or towards the sensor, respectively. This illustrates that the vision baseline degrades with range, showing lower performance than mmWave-Single beyond 8.5m as subjects walk away from the sensor.

the intrinsic complementarity between positional and motional mmWave inputs, supporting our insight for our multi-modal model and highlighting its capacity to adeptly navigate both spatial precision and temporal coherence. Furthermore, compared to single-view, multi-view setup shows significant superiority across all protocols.

Single-View Camera vs Single-View mmWave. Note that the results in Table 1 are focused on the performance in self-occluded scenarios. We extend our analysis for the vision [36] and single-view mmWave baselines to all walking scenarios, as showcased in Fig. 4. Overall, the vision baseline exhibits a decline in performance as the distance level increases. Additionally, examining the effect of walking angles reveals significant difficulties for the vision baseline when subjects walk away from the camera (*i.e.* self-occluded joints), particularly between -60° and 60°, where self-occlusions are pronounced. In contrast, the mmWave-single baseline maintains consistent robustness both against distance variations and occlusions, outperforming the vision model in challenging conditions like distant occlusions, particularly at distances beyond 8.5m.

Single-View mmWave vs Multi-View mmWave. Despite the potential of mmWave signals in dealing with occlusions and long distances, the single-view mmWave baseline encounters notable performance declines when subjects move tangentially to the radar (*i.e.* around ±90° in Fig. 5), posing limited capabilities over specific trajectories of walking subjects. Our multi-view setup effectively resolves this challenge, unleashing the full potential of mmWave technology in HWPE: it consistently shows significant resilience across nearly all scenarios, uniquely ensuring distance- and orientation-invariant HWPE within the entire coverage.

Qualitative Results. Fig. 6 showcases the 3D poses produced by our model alongside several baselines under various scenarios, accompanied by RGB visualizations. It is evident that mmWave model successfully predicts the actual pose even from challenging conditions with considerable distances, occlusion, and dark lighting—scenarios where other models often falter. Furthermore, the multi-view mmWave model maintains its accuracy in the perpendicular movement of the subject, providing an effective solution for the challenges in the single-view mmWave model.

Table 2. Performance comparison by applying other multi-modal fusion strategies or the variants of our fusion module.

Fusion Module	MPJPE↓	PCK↑	ρ↑
Wang *et al.* [32]	67.95	87.17	0.48
Prakash *et al.* [24]	65.26	89.31	0.50
Li *et al.* [16]	63.56	90.85	0.52
Ours w Single-Layer Fusion	65.09	90.65	0.52
Ours w/o Sparse Attention	62.77	91.17	0.52
Ours	**60.96**	**93.24**	**0.53**

5.3. Ablation Studies

Effectiveness of Fusion Strategy. To investigate the efficacy of our model's fusion strategy employed for harmonizing multi-modal and multi-view inputs, we conduct a comparative study with a range of alternative fusion strategies [16, 24, 32], encompassing both traditional multi-modal fusion modules and variants of our fusion approach. The results detailed in Table 2 demonstrate that the proposed approach—characterized by dual-stage multi-layered sparse cross-attention—consistently surpasses other alternatives across all assessed metrics. Moreover, the comparison with our module's variants numerically validates that applying the multi-layered fusion across the encoding pipeline, coupled with sparse attention, can further enhance the overall efficacy of the fusion process.

Effectiveness of Cross-Domain Loss Function. We delve into the impact of our cross-domain loss by comparing it with mmWave variants trained with the standard po-

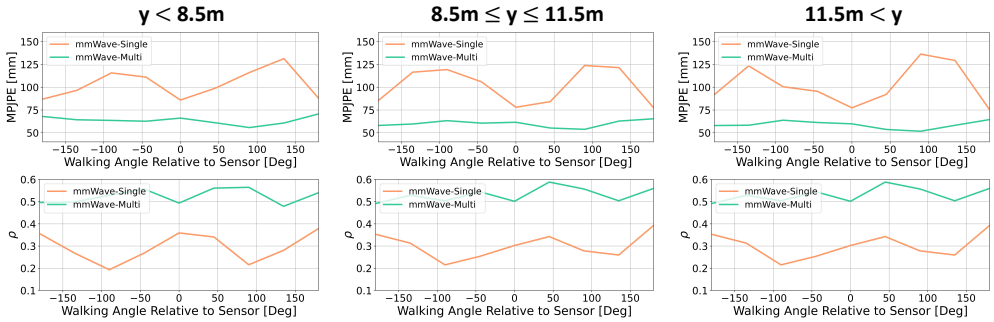

Figure 5. MPJPE (**Top**) and temporal correlation (**Bottom**) of mmWave-single and mmWave-multi baselines with respect to walking directions. 0° or 180° represents a person walking away or towards the sensor, respectively. While mmWave-Single experiences significant performance drops in perpendicular walking scenarios, mmWave-Multi consistently maintains reliable HWPE across nearly all scenarios.

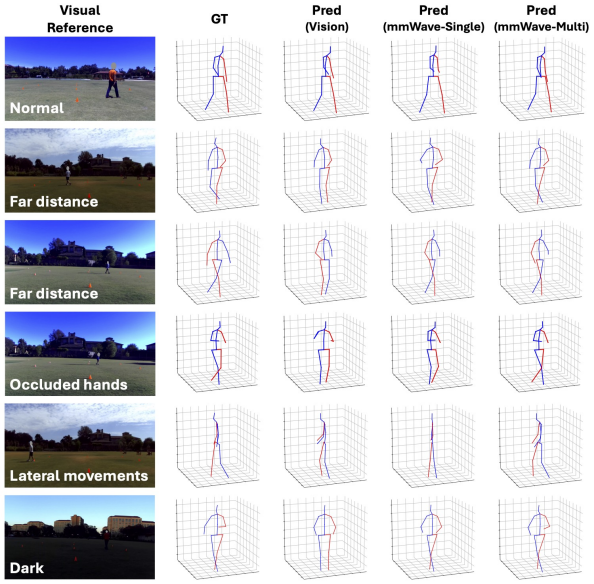

Figure 6. Qualitative comparison of each baseline under various walking scenarios.

Table 3. Performance comparison for different loss functions.

Modality	Loss–Positional			Loss–Cross-Domain		
	MPJPE↓	PCK↑	ρ↑	MPJPE↓	PCK↑	ρ↑
mmW-Position	71.83	83.45	0.40	66.92	89.02	0.48
mmW-Motion	72.82	81.22	**0.45**	68.41	86.46	0.52
mmW-Multi	**67.53**	**89.37**	**0.45**	**60.96**	**93.24**	**0.53**

sitional loss (Table 3). Overall, the inclusion of an additional loss component from the motional domain enhances not only the temporal correlation of the models but also their spatial accuracy. Particularly noteworthy is the remarkable synergy this cross-domain loss yields with our model's multi-modal inputs, compared to single-modal cases.

5.4. Limitations

Despite the achievements of our MVDoppler-Pose, it still has several limitations. Although we have substantially mitigated constraints in traditional mmWave HWPE systems, our testing conditions are still somewhat limited— scenarios involving a single subject and environments with minimal clutter. Additionally, the scope of human walking activities is relatively narrow as well. The issues related with single subject and limited clutter scenarios could potentially be mitigated by integrating additional detection/tracking and decluttering algorithms into our existing framework. Additionally, considering that these overall limitations in validation stem from the constraints of the MVDoppler dataset [8], we aim to collect and validate data from in-the-wild scenarios to enhance the robustness and applicability.

6. Conclusion

This study introduces MVDoppler-Pose, a new mmWave-based HWPE model designed to deal with the challenges of self-occlusions and long-range estimation in traditional HWPE systems. Our system is grounded in two foundational insights: the synergistic relationship between positional and motional signals in mmWave sensor, and the potential of multi-view fusion to overcome the challenges in capturing details of lateral movements. The experimental results validate that our model not only mitigates the inherent shortcomings of vision-based HWPE—such as self-occlusions and distance-related degradation—but also overcomes the primary issues of current mmWave-based HWPE approaches. Ultimately, MVDoppler-Pose first establishes a robust framework for consistent and accurate HWPE across a variety of subject distances and trajectories.

Acknowledgments

This work has been supported by Samsung Electronics (Samsung Semiconductor USA) and Texas Instruments.

References

[1] Sizhe An, Yin Li, and Umit Ogras. mRI: Multi-modal 3D human pose estimation dataset using mmWave, RGB-D, and inertial sensors. In *NeurIPS*, pages 1–13, 2022. 1, 2, 3, 6

[2] Andrea Avogaro, Federico Cunico, Bodo Rosenhahn, and Francesco Setti. Markerless human pose estimation for biomedical applications: a survey. *Frontiers Comput. Sci.*, 5, 2023. 1

[3] Victor C Chen. *The micro-Doppler effect in radar*. Artech house, 2019. 1, 3

[4] Jae-Ho Choi, Ki-Bong Kang, and Kyung-Tae Kim. Remote respiration monitoring of moving person using radio signals. In *ECCV*, pages 253–270, 2022. 3

[5] Lijie Fan, Tianhong Li, Yuan Yuan, and Dina Katabi. In-home daily-life captioning using radio signals. In *ECCV*, pages 105–123, 2020. 1

[6] Lili Fan, Junhao Wang, Yuanmeng Chang, Yuke Li, Yutong Wang, and Dongpu Cao. 4D mmWave radar for autonomous driving perception: A comprehensive survey. *IEEE Trans. Intell. Veh.*, 9(4):4606–4620, 2024. 1

[7] Hao-Shu Fang, Yuanlu Xu, Wenguan Wang, Xiaobai Liu, and Song-Chun Zhu. Learning pose grammar to encode human body configuration for 3D pose estimation. In *AAAI*, pages 1–8, 2018. 2

[8] Soheil Hor, Shubo Yang, Jae-Ho Choi, and Amin Arbabian. MVDoppler: Unleashing the power of multi-view doppler for micromotion-based gait classification. In *NeurIPS*, pages 1–11, 2023. 2, 4, 6, 8

[9] Cesar Iovescu and Sandeep Rao. The fundamentals of millimeter wave sensors. *Texas Instrum.*, pages 1–8, 2017. 3

[10] Chengkun Jiang, Junchen Guo, Yuan He, Meng Jin, Shuai Li, and Yunhao Liu. mmVib: Micrometer-level vibration measurement with mmWave radar. In *ACM Annu. Int. Conf. Mobile Comput. Netw. (MobiCom)*, pages 1–13, 2020. 3

[11] Angjoo Kanazawa, Michael J. Black, David W. Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *CVPR*, pages 7123–7131, 2018. 1, 2

[12] Shih-Po Lee, Niraj Prakash Kini, Wen-Hsiao Peng, Ching-Wen Ma, and Jenq-Neng Hwang. HuPR: A benchmark for human pose estimation using millimeter wave radar. In *WACV*, pages 5715–5724, 2023. 3

[13] Jian Li and Petre Stoica. *MIMO radar signal processing*. John Wiley & Sons, Hoboken, New Jersey, USA, 2008. 3

[14] Jingzhong Li, Yuyi Wang, Lin Yang, Jun Lin, Gaoqiang Kang, Zhen Shi, Yuxuan Chen, Yue Jin, and Kanta Akiyama. L-RadSet: A long-range multimodal dataset with 4D radar for autonomous driving and its application. *IEEE Trans. Intell. Veh.*, pages 1–16, 2024. 1

[15] Wenhao Li, Hong Liu, Hao Tang, Pichao Wang, and Luc Van Gool. Mhformer: Multi-hypothesis transformer for 3d human pose estimation. In *CVPR*, pages 13147–13156, 2022. 1, 2, 6

[16] Yingwei Li, Adams Wei Yu, Tianjian Meng, Ben Caine, Jiquan Ngiam, Daiyi Peng, Junyang Shen, Yifeng Lu, Denny Zhou, Quoc V. Le, Alan Yuille, and Mingxing Tan. DeepFusion: Lidar-camera deep fusion for multi-modal 3D object detection. In *CVPR*, pages 17182–17191, 2022. 7

[17] Mengyuan Liu and Junsong Yuan. Recognizing human actions as the evolution of pose estimation maps. In *CVPR*, 2018. 1

[18] Diogo C. Luvizon, David Picard, and Hedi Tabia. Multi-task deep learning for real-time 3D human pose estimation and action recognition. *IEEE TPAMI*, 43(8):2752–2764, 2021.

[19] Diogo C. Luvizon, David Picard, and Hedi Tabia. Multi-task deep learning for real-time 3D human pose estimation and action recognition. *IEEE TPAMI*, 43(8):2752–2764, 2021. 1

[20] Julieta Martinez, Rayat Hossain, Javier Romero, and James J. Little. A simple yet effective baseline for 3D human pose estimation. In *ICCV*, pages 2659–2668, 2017. 2

[21] Margaret A McDowell, Cheryl D Fryar, and Cynthia L Ogden. Anthropometric reference data for children and adults: United States, 1988-1994. *Vital Health Stat.*, pages 1–68, 2009. 3

[22] Sachin Mehta and Mohammad Rastegari. MobileViT: Lightweight, general-purpose, and mobile-friendly vision transformer. In *ICLR*, pages 1–26, 2022. 4

[23] Dario Pavllo, Christoph Feichtenhofer, David Grangier, and Michael Auli. 3D human pose estimation in video with temporal convolutions and semi-supervised training. In *CVPR*, pages 7745–7754, 2019. 2

[24] Aditya Prakash, Kashyap Chitta, and Andreas Geiger. Multi-modal fusion transformer for end-to-end autonomous driving. In *CVPR*, pages 7077–7087, 2021. 7

[25] Byungseok Roh, JaeWoong Shin, Wuhyun Shin, and Saehoon Kim. Sparse DETR: Efficient end-to-end object detection with learnable sparsity. In *ICLR*, pages 1–23, 2022. 5

[26] Shunqiao Sun, Athina P. Petropulu, and H. Vincent Poor. MIMO radar for advanced driver-assistance systems and autonomous driving: Advantages and challenges. *IEEE Signal Process. Mag.*, 37(4):98–117, 2020. 1

[27] Xiao Sun, Bin Xiao, Fangyin Wei, Shuang Liang, and Yichen Wei. Integral human pose regression. In *ECCV*, pages 536–553, 2018. 1, 2, 6

[28] Chong Tang, Wenda Li, Shelly Vishwakarma, Fangzhan Shi, Simon Julier, and Kevin Chetty. MDPose: Human skeletal motion reconstruction using WiFi micro-Doppler signatures. *IEEE Trans. Aerosp. Electron. Syst.*, pages 1–12, 2023. 1, 3, 6

[29] Zhenhua Tang, Zhaofan Qiu, Yanbin Hao, Richang Hong, and Ting Yao. 3D human pose estimation with spatio-temporal criss-cross attention. In *CVPR*, pages 4790–4799, 2023. 2

[30] Texas Instruments. AWR1843 Single-Chip 77- to 79-GHz FMCW Radar Sensor datasheet. https://www.ti.com/lit/ds/symlink/awr1843.pdf?ts=1705831850034/, 2022. Accessed: 2024-01-21. 3

[31] Yating Tian, Hongwen Zhang, Yebin Liu, and Limin Wang. Recovering 3d human mesh from monocular images: A survey. *IEEE TPAMI*, 45(12):15406–15425, 2023. 1

[32] Chunwei Wang, Chao Ma, Ming Zhu, and Xiaokang Yang. PointAugmenting: Cross-modal augmentation for 3D object detection. In *CVPR*, pages 11794–11803, 2021. 7

[33] Fei Wang, Sanping Zhou, Stanislav Panev, Jinsong Han, and Dong Huang. Person-in-WiFi: Fine-grained person perception using WiFi. In *ICCV*, 2019. 1, 3, 6

[34] Jingbo Wang, Sijie Yan, Yuanjun Xiong, and Dahua Lin. Motion guided 3D pose estimation from videos. In *ECCV*, pages 764–780, 2020. 5

[35] Shuai Wang, Dongjiang Cao, Ruofeng Liu, Wenchao Jiang, Tianshun Yao, and Chris Xiaoxuan Lu. Human parsing with joint learning for dynamic mmwave radar point cloud. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol. (IMWUT)*, 7(1):1–22, 2023. 2

[36] Wen-Li Wei, Jen-Chun Lin, Tyng-Luh Liu, and Hong-Yuan Mark Liao. Capturing humans in motion: Temporal-attentive 3D human pose and shape estimation from monocular video. In *CVPR*, 2022. 2, 6, 7

[37] Jianfei Yang, He Huang, Yunjiao Zhou, Xinyan Chen, Yuecong Xu, Shenghai Yuan, Han Zou, Chris Xiaoxuan Lu, and Lihua Xie. MM-Fi: Multi-modal non-intrusive 4D human dataset for versatile wireless sensing. In *NeurIPS*, pages 1–13, 2023. 1, 2, 3, 6

[38] Jae Shin Yoon, Lingjie Liu, Vladislav Golyanik, Kripasindhu Sarkar, Hyun Soo Park, and Christian Theobalt. Pose-guided human animation from a single image in the wild. In *CVPR*, pages 15034–15043, 2021. 1

[39] Andrei Zanfir, Mihai Zanfir, Alex Gorban, Jingwei Ji, Yin Zhou, Dragomir Anguelov, and Cristian Sminchisescu. HUM3DIL: Semi-supervised multi-modal 3D humanpose estimation for autonomous driving. In *CoRL*, pages 1114–1124, 2023. 1

[40] Mingmin Zhao, Yonglong Tian, Hang Zhao, Mohammad Abu Alsheikh, Tianhong Li, Rumen Hristov, Zachary Kabelac, Dina Katabi, and Antonio Torralba. RF-based 3D skeletons. In *Conf. ACM Special Interest Group Data Commun. (SIGCOMM)*, page 267–281, 2018. 3

[41] Mingmin Zhao, Yingcheng Liu, Aniruddh Raghu, Hang Zhao, Tianhong Li, Antonio Torralba, and Dina Katabi. Through-wall human mesh recovery using radio signals. In *ICCV*, pages 10112–10121, 2019. 3

[42] Qitao Zhao, Ce Zheng, Mengyuan Liu, Pichao Wang, and Chen Chen. PoseFormerV2: Exploring frequency domain for efficient and robust 3D human pose estimation. In *CVPR*, pages 8877–8886, 2023. 1, 2, 6

[43] Ce Zheng, Wenhan Wu, Chen Chen, Taojiannan Yang, Sijie Zhu, Ju Shen, Nasser Kehtarnavaz, and Mubarak Shah. Deep learning-based human pose estimation: A survey. *ACM Comput. Surv.*, 56(1), 2023. 1

[44] Zhijie Zheng, Jun Pan, Zhikang Ni, Cheng Shi, Diankun Zhang, Xiaojun Liu, and Guangyou Fang. Recovering human pose and shape from through-the-wall radar images. *IEEE Trans. Geosci. Remote Sens.*, 60:1–15, 2022. 1

[45] Xiaolong Zhou, Tian Jin, Yongpeng Dai, Yongkun Song, and Zhifeng Qiu. MD-pose: Human pose estimation for single-channel UWB radar. *IEEE J. Biom. Behavior Id. Sci.*, 5(4): 449–463, 2023. 1, 3, 6