



# A Survey of Wireless Sensing Human Posture Recognition Research

Zihan Qi

School of Electrical and Information Engineering, Tianjin University, 300072, Tianjin, China  
qizihan0427@tju.edu.cn

**Abstract.** With the development of human-computer interaction, virtual reality, and other fields, the importance of human posture estimation technology is becoming increasingly prominent. Traditional optical methods are limited by light, occlusion, and privacy issues, whereas wireless sensing techniques have become a core research direction owing to their non-invasiveness and environmental robustness. In this study, nine cutting-edge wireless sensing technologies based on millimeter-wave radar, through-wall radar, and WiFi signals were analyzed in terms of hardware suitability, signal processing frameworks, and deep learning model optimization strategies. Millimeter-wave radar technology forms a technological divide between accuracy, light weight, and multi-person robustness, whereas through-wall radar focuses on the balance between accuracy and generalization ability in walled scenarios, and WiFi signals show unique value in the smart home and security fields. The study points out that current technology faces challenges such as signal sparsity, insufficient cross-modal generalization, and high hardware cost, and explores typical application scenarios such as medical, healthcare, and industrial security. Future research can focus on self-supervised multimodal fusion, privacy-preserving frameworks, and dynamic adaptive algorithms, laying the technical foundation for 6G communication-sensing integration..

**Keywords:** Wireless sensing, Human posture recognition, Millimeter-wave radar, Through-wall radar, WiFi signals;

## 1 Introduction

Human posture estimation, as a core technology of intelligent systems, plays a key role in human-computer interaction, virtual reality, intelligent security, healthcare, and other fields. In recent years, wireless sensing technology has become an important direction to overcome the bottleneck of traditional methods by virtue of its unique advantages, such as non-invasiveness, privacy preservation, and environmental robustness. This technology realizes attitude estimation by integrating deep learning algorithms through the reflection and diffraction features generated by different frequency bands of electromagnetic waves interacting with the human body, such as high-frequency millimeter

waves, low-frequency wall-penetrating radar, and WiFi signals, which show broad application prospects in diverse scenarios, such as medical rehabilitation, smart homes, and industrial security.

Existing wireless sensing technologies can be divided into three categories based on signal frequency bands: high-frequency millimeter-wave radar realizes high-resolution point cloud imaging with the help of short-wavelength characteristics to support joint-level attitude reconstruction; low-frequency wall-penetrating radar relies on signal penetration capability to realize wall tracking in obscured scenes; and WiFi signals are analyzed by multipath reflection to consider both communication and sensing functions. However, these technology paths generally face common challenges, such as signal sparsity, noise interference, and insufficient cross-scene generalization ability, and the algorithm design is fragmented owing to the differences in signal characteristics, which urgently requires systematic sorting and optimization.

This paper systematically reviews nine cutting-edge wireless sensing technologies based on millimeter-wave radar, through-wall radar, and WiFi signals, and analyzes them from the perspective of hardware adaptability, signal processing framework, and deep learning model optimization strategy, aiming to reveal the core frameworks, strengths and weaknesses, and applicable scenarios of the different technologies, analyze the current challenges, and look forward to the direction of future development to provide theoretical references for the landing of cross-scenario technologies.

## 2 Typical technical analysis

### 2.1 Typical technology analysis of millimeter wave radar

**mmPose-FK.** mmPose-FK is a novel millimeter wave (mmWave) radar-based pose estimation method that employs a dynamic forward kinematics (FK) approach to address the challenges posed by low resolution. At its core, the technique achieves high-precision dynamic skeletal pose estimation through dynamic forward kinematic modeling, point cloud feature enhancement, and time sequence stability optimization. It treats the human skeleton as a kinematic chain, directly predicts joint rotation angles, such as Euler angles or quaternions, with the help of deep learning, and generates continuous attitude sequences by combining with bone length constraints, thus ensuring that the attitude outputs conform to the laws of human kinematics. For sparse radar point cloud data, the technique employs volumetric processing to preserve 3D spatial relationships, whereas it extracts local and global features through multi-scale 3D convolution to enhance point cloud feature representation. For timing processing, bidirectional long short-term memory (LSTM) is introduced to model timing dependencies and is combined with FK inverse solving to reduce noise-induced joint jitter and enhance the stability of the attitude sequence [1].

mmPose-FK has significant advantages, with high-precision characteristics, joint-level error of less than 5 cm in single-target scenarios, suitability for medical rehabili-

tation, and other scenarios with very high requirements for precision; based on FK constraints, the technology avoids posture inversion problems such as knee hyperextension, making the output results physically reasonable; it supports multimodal parameterized expressions such as joint rotation angle and bone length, and can be adapted to gait analysis, and supports multi-modal parameterization such as joint rotation angle and bone length, which can be adapted to diversified application requirements such as gait analysis and virtualized body drive. However, there are some limitations to this technique, such as high computational complexity due to voxelization and 3D convolution operation, and dependence on graphics processing unit (GPU) for real-time inference. Moreover, it needs to pre-calibrate the bone lengths, which is difficult to adapt to dynamic targets, such as children, due to the possibility of changes in the lengths of the bones.

**MobiRFPose.** MobiRFPose is a lightweight 3D pose estimation technique that uses a two-stage detection-estimation framework that combines model compression techniques with cross-view feature fusion to achieve lightweight 3D pose estimation. In the first stage, the technique quickly locates the center of the human body through a human localization network based on an anchor detector and crops local regions to reduce the input dimension and the amount of data for subsequent processing. In the second stage, it utilizes a lightweight 3D Convolutional Neural Networks (CNN) to process the region of interest and regress to the joint coordinates. To support mobile deployment, MobiRFPose adopts model compression techniques, such as channel pruning and quantization, to compress the model volume to 268 KB. Regarding the view angle limitation of a single horizontal radar in the vertical dimension, the technique implicitly captures the vertical dimension information through the signal diffraction property to realize cross-view feature fusion and compensate for the hardware deficiency [2].

The advantages of MobiRFPose lie in its high real-time performance, which can reach 66FPS on the central processing unit (CPU) side and can meet the demand of real-time interaction in a smart home. Its model parameter count is less than 1M, and the memory occupancy is only 3.2MB, which is a good adaptability for edge devices, and by simulating multipath noise through the data enhancement method, the technology effectively improves the environmental robustness in different indoor and outdoor scenarios. However, MobiRFPose has some limitations. Owing to the limitation of the single-view angle, the vertical resolution is insufficient, the Z-axis error is approximately 11 cm, and the accuracy will be decreased in complex action scenes such as raising hands. When regions of interest overlap in a multi-person scene, it is easy to have the feature confusion problem, and the error will be increased by 15% in a multi-person scene.

**RPM2.0.** RPM2.0 is a novel framework that infers accurate multi-person 3D skeletons from commercial millimeter-wave radars. The technical core of RPM2.0 is multi-radar feature alignment, Spatio-Temporal attention network, and occlusion robustness design for coping with multi-person pose estimation scenarios. In terms of feature processing, RPM2.0 encodes horizontal and vertical radar signals into 3D voxels separately, which

are projected to a unified space through a coordinate transformation matrix for weighted fusion to achieve effective integration of multi-view features. The spatiotemporal attention network contains two parts: spatial and temporal attention. Spatial attention models the nonlocal correlation between joints based on graph convolution and is able to recover the pose information of the occluded parts. Temporal attention uses the transformer to encode the long temporal dependency and optimize the action continuity. To improve the anti-obscuring ability, RPM2.0 adopts a dynamic masking training strategy, which randomly masks some joints and forces the network to learn the contextual reasoning ability to accurately estimate the pose in an occluded scene [3].

RPM2.0 achieves an average accuracy of 125.3 mm in multiplayer scenarios through the spatiotemporal attention mechanism, supporting gesture estimation for intensive interactions, such as handshaking and hugging. The architecture has outstanding occlusion resistance, and the spatiotemporal attention mechanism only increases the error by 8% under 50% occlusion. It also has cross-device generalization capability and can be adapted to different radar models, but there is a bottleneck of high computational complexity: the multi-branch network architecture leads to a 200ms inference delay, making it difficult to meet real-time tracking requirements. When the error of multi-radar time synchronization is more than 2ms, the fusion performance decreases by 20%. This contradiction between performance and cost highlights the core challenge of high-precision multi-person sensing systems in engineering applications.

## 2.2 Typical technology analysis of through-wall radar

**ST<sup>2</sup>W-AP.** ST<sup>2</sup>W-AP is a Spatio-Temporally separated 4D imaging radar framework whose technical core lies in efficient bulkhead attitude estimation and behavior recognition through a Spatio-Temporal step-by-step strategy, deep echo domain compensation, and multi-task joint optimization. The technology system decomposes the 4D radar data into two dimensions of 3D spatial voxels composed of distance-orientation-height and temporal features and extracts the spatial voxel features and temporal dynamic features through a cascaded 3D convolutional network and 1D temporal network, respectively, to avoid direct processing of high-complexity 4D data and to improve computational efficiency. In terms of signal compensation, technology utilizes deep learning to fit the signal attenuation model after wall penetration, replacing the traditional iterative algorithm, reducing the computational overhead by 90%, and effectively suppressing the wall reflection interference. In addition, the technology parallel outputs 3D joint point pose estimation and action category behavior recognition through a shared spatial feature extraction layer and balances the multi-task learning weights with the help of a weighted loss function to achieve synergistic optimization of the two [4].

ST<sup>2</sup>W-AP technology has significant advantages, the single-target pose estimation error is only 73 mm, and the behavior recognition F1 score reaches 92.4% in the wall-separated scene, which is a higher accuracy; the signal-to-noise ratio of the signal is increased by 8 dB after depth compensation, and the wall interference is greatly reduced; under GPU acceleration, the 4D data processing speed reaches 20FPS, which supports real-time monitoring of dynamic scenes. However, its limitation is the weak

multi-target processing ability: the pose error increases to 125 mm in the 3-person scene, and the behavior recognition accuracy decreases by 12%, which relies on the pre-calibrated data of wall thickness and material, and the compensator needs to be re-trained for cross-scene migration, so the flexibility is limited.

**Dual-Task Network Framework.** The dual-task joint learning framework achieves synergistic optimization of pose reconstruction and behavior recognition through a feature sharing mechanism, dynamic loss balancing, and temporal modeling. The framework adopts ResNet 3D as the backbone network to extract radar voxel features, and the branch network regresses 3D joint coordinates for pose estimation and outputs action category probabilities for behavior recognition to reduce redundant computation through feature sharing. In loss function design, homoscedastic uncertainty is introduced to automatically adjust the weights of attitude and behavior recognition to avoid an imbalance between tasks. For the temporal characteristics of behavior recognition, a gated recurrent unit (GRU) module is added to the branch network to capture the temporal correlation of successive actions, such as "fall-down-get-up," to improve the recognition ability of dynamic actions [5].

Its advantages are obvious task synergy effect, gesture reconstruction can assist behavior recognition, in the "hand up alarm" and other composite action recognition accuracy increased by 15%; shared features to reduce the model's dependence on the amount of data, in small sample scenarios better than the performance of the single-task model; model parameter count is only 5.2M, lightweight design to adapt to the edge of the computing device. The model parameter count is only 5.2M, and the lightweight design is suitable for edge-computing devices. However, the framework has the problem of task conflict, and the pose accuracy and behavior recognition target may constrain each other in extreme scenes such as fast motion, resulting in error amplification; and the occlusion scene is not explicitly modeled, and when the proportion of occlusion reaches 50%, the accuracy of behavior recognition decreases by 25%, and the robustness needs to be improved.

**MIMDSN.** MIMDSN is a mutual information maximizing deeply supervised network that improves the anti-obscuration capability and cross-domain generalization of bulkhead attitude estimation through cross-modal mutual information maximization, a deep supervisory mechanism, and non-local attention. The core of this technique is the use of contrast learning to align the mutual information of radar features and optical pseudo-labels generated by multiview cameras to enhance the discriminative nature of the features so that the model can extract effective attitude information from noise. Simultaneously, supervisory signals are injected at the intermediate network layer to constrain the geometric consistency of shallow features with the coordinates of the joints, ensuring that feature learning conforms to the human structure a priori. A self-attention module is introduced into the decoder to construct a non-local attention mechanism to model the long-range dependencies between joints, such as "left shoulder-right hip," and recover the pose information of the occluded parts [6].

The significant advantages of this technique are outstanding occlusion resistance, which can still recover a reasonable attitude in 80% occlusion scenarios; the error is controlled within 150 mm; the technique does not need to be retrained when migrating across domains, and the error only increases by 8% from concrete walls to new environments, such as brick walls, which is a strong generalization; the feature visualization of this technique shows that the joints' heat maps are highly aligned with the optical data, and the model interpretability is better. However, its limitation is the strong dependence on optical pseudo-labeling: if the optical calibration error is more than 20 mm, it will be directly transferred to the radar model to affect the accuracy. In addition, mutual information comparison learning substantially increases the computational cost, and the training time is extended by three times compared with the traditional method, which requires high hardware resources.

### 2.3 Typical technology analysis of WiFi signal

**Person-in-WiFi3D.** Person-in-WiFi3D is an end-to-end multi-person 3D pose estimation technique based on a transformer, and its technical core lies in the realization of end-to-end multi-person 3D pose estimation through multi-receiver signal fusion, transformer architecture, and dynamic link selection. This technology deploys three WiFi receivers with orthogonal layouts to capture 3D spatial reflections by utilizing the diffraction characteristics of the multilink signals, which effectively enhances the vertical resolution and compensates for the natural limitation of WiFi signals in the Z-axis direction. In signal processing, the transformer architecture is adopted: the encoder chunks channel state information (CSI) signals into Spatio-Temporal tokens, extracts global contextual features through a multi-head self-attention mechanism, and captures complex signal correlations in multi-person interactions; the decoder interacts with the encoded features through a learnable query and directly regresses to the 3D joint coordinates of multi-person interactions, avoiding the error of the traditional method of converting a heatmap to coordinates and avoiding the accumulation problem of heatmap-to-coordinate conversion in traditional methods. In addition, a dynamic link selection strategy is designed based on the signal-to-noise ratio of the signal, which adaptively weights the contributions of different links, suppresses environmental interference such as the multipath effect, and improves the reliability of the features [7].

Its advantages are reflected in the good support for multi-person scenes, which can track 3 people at the same time, and the average joint error is 125.3 mm in 3-people scenes to meet the needs of multi-people interaction scenes; the model volume is only 3.2MB, and the GPU inference speed reaches 22FPS, which is both highly efficient and lightweight; the technology relies on the amplitude information of the CSI signals only, and there is no need for visual data acquisition, which guarantees user privacy from the bottom of the technology. However, owing to the limitation of the WiFi signal wavelength, the vertical resolution is insufficient, and the error of the Z-axis is approximately 150 mm. In the dense overlapping scene of multiple people, the query mechanism is prone to feature confusion, which leads to an increase of 30% in the error, and the performance of complex occlusion scenarios needs to be improved.

**Widar 3.0.** Widar 3.0 is a zero-sample cross-domain gesture recognition technique which realizes zero-sample cross-domain gesture recognition by means of body-coordinate velocity profile (BVP), contrast learning alignment and dynamic noise suppression. The core of the technique is to extract the domain-independent velocity distribution of the target motion based on the Fresnel diffraction model, which eliminates the effect of the environment and positional differences and enables the model to adapt to different scenarios from laboratory to home. Comparative learning with the EI and Cross-Sense framework maximizes the mutual information between BVP features and gesture categories, enabling cross-environmental migration without re-training and overcoming the limitations of the traditional method that relies on scene-specific data. Simultaneously, a dynamic noise suppression mechanism is designed based on the BVP energy distribution, which automatically detects irrelevant human movements such as walking and shields interfering signals to improve the purity of gesture recognition [8].

The significant advantages of this technology are strong cross-domain generalization ability, cross-environment recognition accuracy can reach 82.6%-92.4%, truly realizing "one-time training, all-domain application"; Widar3.0 adopts a lightweight CNN architecture, inference latency <10ms, adapted to low-level arithmetic edge equipment; this technology through BVP Personalized encoding distinguishes the gesture styles of different users, with an accuracy rate of 88.9% in multi-user scenarios. However, the limitation is that the action granularity is coarse, only supports 15 basic gestures such as waving and clapping, and has insufficient recognition ability for complex continuous actions such as writing and typing. Relying on the amplitude of gesture movement, small actions such as finger micro-movement are prone to false detection due to weak signal characteristics.

**WiFense.** WiFense focuses on boundary crossing detection, and its core technologies include Fresnel diffraction signature analysis, dynamic link selection algorithms, and automatic gain control (AGC) feature enhancement. The technology utilizes the diffraction signal changes generated when a target traverses a WiFi transmit-receive link to extract rayleigh distribution in fresnel diffraction model (RFD) features to construct a boundary-crossing event detection model without relying on visual or contact sensors. Through the dynamic link selection algorithm, the most sensitive link is automatically screened according to the link state to enhance the robustness of detection in complex environments. Combined with the sudden change characteristics of the automatic gain control signal during occlusion, it enhances the ability to differentiate between the real boundary-crossing behaviors and the environmental noise, and reduces the false alarm rate[9].

Its advantages include high detection accuracy, >89% detection accuracy in home scenarios, and false alarm rate <5%. To meet the reliability requirements of security monitoring, the technology's processing delay is <50ms, supporting real-time alarm response, suitable for emergency scenarios. WiFense adopts an unsupervised learning approach to modeling the RFD feature distribution, without the need to manually label the data, significantly reducing the cost and time of deployment. WiFense adopts unsupervised learning to model the RFD feature distribution without manual labeling of the data, which greatly reduces the deployment cost and time. However, this technology

has a single function, which can only determine whether the target is out of bounds or not and cannot provide detailed information such as posture and movement. The detection range of a single link is limited, and it is necessary to deploy multiple devices to extend the coverage area, which increases the complexity of hardware deployment.

### **3 Challenges**

#### **3.1 Technical bottlenecks and physical limitations**

Wireless sensing technology faces multiple challenges in terms of signal characterization and cross-modal fusion. First, signal sparsity and noise interference are significant problems, making it difficult to capture details of complex movements, and dynamic noises are prone to false detections. WiFi signals are affected by the multipath effect, and signal superposition in multi-person scenarios can lead to feature confusion. In addition, the signal attenuation of through-wall radar creates a significant constraint on the detection capability of weakly reflective targets.

Second, the problem of insufficient cross-modal generalization capability must be addressed. Existing models perform erratically when migrating across devices or scenarios, which usually needs to be mitigated. Meanwhile, the intrinsic differences in feature dimensions between signals result in the absence of a unified characterization framework, contributing to the limitation of early multimodal fusion schemes. In terms of privacy and ethical controversies, wireless sensing technologies may leak user identities through biometric features, highlighting the need for privacy protection mechanisms.

#### **3.2 Hardware and deployment costs**

The high cost and complex deployment at the hardware level constrain the popularization of the technology. The problem of high-precision equipment dependence is particularly prominent, which seriously hampers promotion in home scenarios. The deployment of multi-antenna WiFi arrays is also challenging, significantly increasing the complexity of technology landing. In terms of the contradiction between energy consumption and real-time performance, high-precision models require significant power, whereas edge computing devices have limited processing speed, making it difficult to satisfy real-time security requirements. Although low-power models achieve certain capabilities, they have performance defects, indicating that the current technology still needs improvement in balancing performance and energy consumption.

#### **3.3 Algorithm Robustness**

The robustness of algorithms in complex scenes must be improved urgently. In dynamic occlusion and multi-person interaction scenarios, the dense overlapping of targets causes feature overlapping effects. Existing methods still have limitations. In wall penetration scenarios, phenomena generated by multiple reflections can interfere with real

target localization. There are obvious shortcomings in cross-body type and action generalization ability. Physiological characteristic differences significantly impact model adaptability and increase recognition error. For unconventional action patterns, the algorithm is prone to misjudgment due to lack of relevant samples, restricting reliable application in diverse scenarios.

## 4 Application scenarios

### 4.1 Medical health and rehabilitation monitoring

Wireless sensing technology provides a new path for precision medicine through non-contact monitoring in the field of healthcare. In postoperative rehabilitation assessment, mmPose-FK, a millimeter-wave radar technology, can monitor the flexion and extension angles of knee replacement patients in real time with a precision of  $\pm 1.5^\circ$ , which is significantly better than the  $\pm 5^\circ$  error of manual measurement owing to its high precision. It generates continuous posture sequences through dynamic forward kinematics modeling and automatically generates a rehabilitation report after the data are uploaded to the cloud in real time, assisting the doctor in formulating a personalized rehabilitation plan.

In respiration and sleep monitoring scenarios, WiFense, a WiFi signal, and millimeter-wave radar form a complementary technology. WiFense detects respiration rate by capturing micromotion signals from the chest cavity, with an error of  $< 0.5$  breaths/min, and combined with the millimeter-wave radar's ability to monitor sleep postures, it can effectively identify sleep apnea syndromes. This non-contact solution avoids the discomfort of electrode patches on the human body and is especially suitable for sensitive scenarios, such as neonatal monitoring, which improves monitoring comfort while safeguarding data accuracy.

### 4.2 Smart home and human-computer interaction

Widar3.0 is based on the domain-independent speed characteristics of WiFi signals to realize cross-room gesture control, such as "sliding in the air" to adjust the temperature of air conditioning with an accuracy of 92%, and supports user-defined gestures, such as "clench fist" to trigger the command to turn off the lights, which greatly improves the naturalness of interaction. This technology avoids uploading raw signals by locally processing the hash value of gesture features, which protects the user's privacy in terms of mechanism.

To meet the needs of elderly monitoring, the wall-penetrating radar ST<sup>2</sup>W-AP recognizes fall events by detecting sudden changes in acceleration and falling to the ground in a stationary state and triggers an alarm within 5 s, with a measured false alarm rate of  $< 3\%$  in nursing home scenarios. Compared with the traditional camera solution, it does not require visual data collection, avoids privacy invasion, and can work stably at night in a lightless environment, providing a safe and reliable non-contact monitoring solution for senior living scenarios.

### 4.3 Industrial security and emergency response

In the industry and security fields, wireless sensing technology exhibits high robustness and real-time advantages. WiFense is based on Fresnel diffraction characteristics to delineate the electronic boundaries of hazardous areas, such as robotic arm operations, and triggers an emergency stop response when the personnel cross the boundaries, with a delay of <50ms, and a false alarm rate of <2%, as measured by the Siemens factory in Germany. Its extended application can penetrate chemical plant protective clothing to monitor vital signs and provide technical support for toxic gas intrusion detection, considering both safety and practicality.

In disaster rescue scenarios, the wall-penetrating radar MIMDSN with its cross-modal mutual information maximization technology can penetrate 2-meter-thick concrete to detect the breathing signals of trapped people, and the joint UAV positioning efficiency is 50% higher than that of infrared thermal imaging. With its anti-obscuration design and cross-domain generalization capability, the technology can quickly lock the signs of life in complex environments such as ruins and buying golden time for emergency rescue.

### 4.4 Virtual reality and the metaverse

Wireless sensing technology provides innovative solutions for low-cost motion capture and cross-space interaction. Person-in-WiFi 3D utilizes CSI signals from home routers to replace \$10,000 optical motion capture systems to drive virtualized bodies, with a joint positioning error of <5 cm, and supports the plug-in deployment of the Steam VR platform. Independent game studio practice shows that this solution reduces full remote motion capture development costs by 80%, promoting lightweight and popularization of virtual content creation.

Millimeter-wave radar RPM2.0 combines multi-view fusion and Spatio-Temporal attention mechanism to capture real actions and render virtual scenes in real time through generative modeling, realizing cross-space interaction experiences such as "high-five," supplemented with ultrasonic arrays to simulate haptic feedback, and enhancing the immersion of meta-universe scenes. Its high-precision performance in intensive multi-person interactions lays the technical foundation for virtual socialization, remote collaboration, and other applications.

## 5 Future directions

The future of wireless sensing technology will focus on multimodal fusion, privacy protection, and intelligent adaptation, breaking through signal characteristics and scene landing bottlenecks. Multi-modal unified characterization integrates millimeter wave, WiFi, and through-wall radar data through self-supervised learning to construct a cross-modal universal model, break the fragmentation of the technology, and realize the generalization ability of "one training, multi-source compatibility"[10].

The privacy protection mechanism will be deepened to the physical layer through the edge of the signal desensitization, such as joint code hashing and RF noise fingerprinting injection, to block the identity back-propagation path from the source and ensure that the closed-loop experience of "perception is privacy and security."

Cross-scene adaptive technology relies on federated learning and physical a priori constraints to achieve rapid model "zero sample" migration, combined with dynamic gradient optimization to reduce the cost of cross-domain deployment and to promote plug-and-play technology in factories, hospitals, and other scenarios[11].

Intelligent noise robustness dynamically adjusts the data enhancement strategy through real-time environment sensing to adaptively cope with complex disturbances, such as industrial noise and home multipaths, to improve the stability and reliability of the model in real scenes[12].

Along with the evolution of 6G communication and perception integration, wireless perception technology will integrate multidisciplinary innovations and become the infrastructure of an intelligent society, providing more ubiquitous and secure underlying support for scenarios such as digital twins and meta-universe.

## 6 Conclusion

This study focuses on wireless sensing human gesture recognition technology and reviews the research background, typical technology, challenge analysis, and application scenarios.

Millimeter-wave radar technology presents differentiated development paths in the triad of "high accuracy - light weight - multi-person robustness"; specifically, mmPose-FK achieves medical-grade accuracy through complex algorithm design, MobiRFPose promotes the popularization and application of smart home scenarios with the help of model compression technology, and RPM2.0 adopts a multi-view fusion strategy to effectively solve key challenges in multi-person interaction scenarios. and RPM2.0, adopts a multi-view fusion strategy to effectively solve the key problems in multi-person interaction scenarios.

Through-the-wall radar technology expands application boundaries by balancing accuracy and generalizability, in which ST<sup>2</sup>W-AP innovatively adopts a Spatio-Temporal separation architecture to achieve real-time multitasking capability, and MIMDSN significantly improves anti-obscuration performance through cross-modal comparison learning mechanisms, which makes through-the-wall sensing reliable in complex environments, such as security rescue.

WiFi signaling technologies form characteristic solutions for differentiated scenario requirements. Person-in-WiFi3D breaks through and solves the technical bottleneck of multi-person gesture estimation, Widar3.0 innovates the zero-sample cross-domain gesture recognition paradigm. WiFense realizes the feasibility of border monitoring through cost-optimized solutions.

In terms of signal characterization, millimeter-wave point cloud sparsity and WiFi multipath interference constitute physical constraints on sensing accuracy; the lack of cross-modal generalization capability leads to model performance degradation; and

there is a privacy risk from biometric leakage. Future research should rely on self-supervised learning to build a cross-modal unified characterization framework and combine it with physical a priori knowledge to improve the adaptability of the algorithm. Hardware deployment needs to solve the contradiction between cost and energy consumption, and the co-optimization of lightweight models and edge computing is key. Feature extraction needs to deepen the desensitization and counter RF fingerprinting techniques. Multi-modal fusion, a dynamic adaptive computing enhancement strategy, and an incremental federated learning framework are also needed to promote wireless sensing technology to realize scale landing.

## References

1. S.T. Hu, S.Y. Cao, N. Toosizadeh, J.Barton, M.G. Hector, M.J. Fain, mmPose-FK: A forward kinematics approach to dynamic skeletal pose estimation using mmWave radars. *IEEE Sensors J.* 24, 6469-6481 (2024)
2. C. Yu, D.H. Zhang, Z. Wu, C.Y. Xie, Z. Lu, Y. Hu, Y. Chen, MobiRFPose: portable RF-based 3D human pose camera. *IEEE Trans. Multimedia.* 26, 3715-3727 (2023)
3. C.Y. Xie, D.H. Zhang, Z. Wu, C. Yu, Y. Hu, Y. Chen, RPM 2.0: RF-based pose machines for multi-person 3D pose estimation. *IEEE Trans. Circuits Syst. Video Technol.* 34, 490-503 (2023)
4. R. Zhang, H.Q. Gong, R.Y. Song, Y.D. Li, Z. Lu, D.H. Zhang, Y.Hu, Y. Chen, Through-wall human pose reconstruction and action recognition using four-dimensional imaging radar. *Journal of Radars.* 14, 44-61 (2025)
5. Y. Song, Y. Dai, T. Jin, Y. Song, Dual-task human activity sensing for pose reconstruction and action recognition using 4-D imaging radar. *IEEE Sensors J.* 23, 23927-23940 (2023)
6. Z.J. Zheng, J. Pan, D.K. Zhang, X. Liang, X.J. Liu, G.Y. Fang, Through-wall human pose estimation by mutual information maximizing deeply supervised nets. *IEEE Internet Things J.* 11, 3190-3205 (2023)
7. K.W. Yan, F. Wang, B. Qian, H. Ding, J.S. Han, X. Wei, Person-in-wifi 3d: End-to-end multi-person 3d pose estimation with wi-fi, In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Seattle, WA, USA, June 16 (2024), 969-978.
8. Y. Zhang, Y. Zheng, K. Qian, G.D. Zhang, Y.H. Liu, C.S. Wu, Z. Yang, Widar3. 0: Zero-effort cross-domain gesture recognition with Wi-Fi. *IEEE Trans. Pattern Anal. Mach. Intell.* 44, 8671-8688 (2021)
9. Z.P. Liu, S.J. Li, Y. Zhang, Y.W. Zeng, D.Q. Zhang, WiFense: from diffraction modeling to boundary monitoring. *Journal of Software.* 35, 1515-1533 (2023)
10. K. Teramoto, T. Haruyama, T. Shimoyama, F. Kato, H. Mineno, Human Activity Recognition Using FixMatch-based Semi-supervised Learning with CSI. *J. Inf. Process.* 32, 596-604 (2024)
11. Z. He, M. Bouazizi, G. Gui, T. Ohtsuki, A Cross-subject Transfer Learning Method for CSI-based Wireless Sensing. *IEEE Internet Things J.*
12. H. El Zein, F. Mourad-Chehade, H. Amoud, CSI-based Human Activity Recognition via Lightweight CNN Model and Data Augmentation. *IEEE Sensors J.* 24, 25060-25069 (2024)

**Open Access** This chapter is licensed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License (<http://creativecommons.org/licenses/by-nc/4.0/>), which permits any noncommercial use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

