Towards Robust mmWave-based Human Activity Recognition using Large Simulated Dataset for Model Pretraining

1st Vinay Joshi†
Undergraduate, Dept. of Computer Science
UNC Charlotte
Charlotte, USA
vjoshi9@charlotte.edu

2nd Shengkai Xu
Dept. of Computer Science
UNC Charlotte
Charlotte, USA
sxu7@charlotte.edu

3rd Qiming Cao
Electrical and Computer Engineering
Purdue University
West Lafayette, USA
cao393@purdue.edu

4th Yi Zhu
Dept. of Computer Science
Wayne State University
Detroit, USA
yzhu39@wayne.edu

5th Pu Wang
Dept. of Computer Science
UNC Charlotte
Charlotte, USA
pu.wang@charlotte.edu

6th Hongfei Xue*
Dept. of Computer Science
UNC Charlotte
Charlotte, USA
hongfei.xue@charlotte.edu

***Abstract*—Human activity recognition (HAR) is crucial for real-world applications such as healthcare, surveillance, and smart homes. Among sensing technologies, millimeter wave (mmWave) sensors stand out due to their contactless nature, high sensitivity, and ability to operate in low-light environments while preserving privacy. However, the scarcity of mmWave sensing data limits the generalizability of mmWave-based HAR systems. To address this, we propose mmAP, a data augmentation and pretraining framework that synthesizes a large mmWave dataset using human mesh data, followed by pretraining a robust and general mmWave heatmap encoder using a multi-modal masked autoencoder framework using the synthesized data. We enhance the model's robustness with heatmap-specific data perturbations and perform task-specific fine-tuning on a small real-world dataset. The experiment results over the baseline demonstrate the effectiveness of the proposed mmAP framework.**

***Index Terms*—mmWave Sensing, Human Activity Recognition, Data Synthesis, Masked Model Pretraining, Multi-modality**

I. Introduction

As the demand for intelligent systems that enhance human life continues to grow, human activity recognition (HAR) has become essential for understanding human behavior, supporting applications in healthcare [1], surveillance [2], smart homes [3], etc. Among various sensing technologies, wireless sensing has gained significant attention due to its contactless nature, offering a new paradigm in human activity monitoring. The mechanism of wireless sensing relies on the interaction between wireless signals and the sensing target: as signals propagate, their characteristics, such as phase, amplitude, and frequency, are affected by the target. By analyzing these changes, target-related information, including human activity or gesture, can be captured and analyzed to infer activities.

†Undergraduate author.
* Corresponding author.

Among all the wireless sensing techniques, millimeter wave (mmWave) sensors, operating at high frequencies (30-300 GHz), have emerged as a promising solution for HAR. Their large bandwidth and short wavelength enable high sensitivity, precision, and advanced capabilities such as beamforming, providing significant advantages over traditional low-frequency sensing methods like Wi-Fi, UWB, and LoRa. Notably, compared to camera-based solutions, mmWave-based sensing technology excels in poor lighting or complete darkness, penetrates obstacles, and addresses privacy concerns inherent to visual data capture. Unlike wearable sensors, mmWave systems eliminate the discomfort and inconvenience of wearing extra devices, offering a seamless and non-intrusive approach to HAR. Additionally, the low cost, compact size, and low power consumption of mmWave radar devices make them well-suited for integration into environments such as homes [4] and robots [5], promoting widespread adoption.

Despite its promising potential, the generalizability of mmWave-based HAR remains limited, primarily due to the scarcity of available mmWave sensing data. Collecting such data is a labor-intensive and time-consuming process, requiring specialized hardware and software, extensive participant involvement, and precise synchronization and calibration, which renders large-scale data collection extremely challenging. To address this challenge, we propose mmAP, a mmWave-based data augmentation and pretraining framework designed to enhance model performance for mmWave-based HAR tasks.

The proposed framework incorporates several key components to address the challenges of mmWave-based HAR. First, to mitigate data scarcity, we perform cross-modality data synthesis by generating a large mmWave dataset from a large-scale human mesh dataset. Following the approach of mmGPE [6] and mmCLIP [7], which simulates mmWave signal reflections on the surface of the human body, we enhance this method by utilizing a large-scale human mesh dataset [8], rather than relying on motion capture data [9], [10], to simulate more diverse and general activities. Second, to exploit the potential of this synthetic dataset, we pretrain the model encoder using a multi-modal masked autoencoder with a high masking rate. Based on the synthesized data, we generate multiple types of signal heatmaps that capture different physical attributes, such as range, velocity, and angle information of the subject. These heatmaps are treated as distinct data modalities, allowing us to leverage a multi-modal masked autoencoder [11] to train a robust heatmap encoder. To adapt the autoencoder for heatmap inputs, we use a temporal masking strategy, masking heatmaps along the time dimension to enable the encoder to learn temporal signal information effectively. Third, to bridge the gap between simulated and real-world data, we conduct task-specific fine-tuning using a small real-world dataset, incorporating a classification head for activity recognition. Lastly, to further enhance model robustness, we utilize heatmap-specific data perturbation techniques during both pretraining and fine-tuning phases. Data perturbation, a common strategy in deep learning, involves artificially expanding the training dataset through various transformations to improve robustness and reduce overfitting. However, applying data perturbation to mmWave signals presents unique challenges, as these signals lack the intuitive interpretability of images. In our framework, we apply three types of data perturbations based on the properties of signal data: adding Gaussian noise, vertical rotation of heatmaps, and temporal cropping and expansion. These techniques introduce variability, helping the model better generalize to real-world scenarios.

In conclusion, the main contributions of this paper are summarized as follows:

- We introduce mmAP, a novel mmWave-based data augmentation and pretraining framework aimed at improving model performance for mmWave-based HAR tasks.
- We develop a method to train a robust mmWave heatmap encoder using a multi-modal masked autoencoder combined with data perturbation techniques.
- We validate the effectiveness of the proposed mmAP framework by creating a real-world human activity recognition testbed utilizing commercial off-the-shelf (COTS) mmWave devices and collected real-world data.

II. Related Works

A. Human Activity Recognition using mmWave

Recent advancements in deep learning have driven the development of numerous mmWave-based sensing systems to "observe, detect, analyze, and interact" with the human body, enabling precise and pervasive human sensing capabilities [12]. The task of mmWave-based human activity recognition is extensively explored by [13]–[15], which leverages signal processing techniques to transform signal attributes into formats suitable for neural networks. Despite their success, these methods commonly face the limitation of requiring extensive real data to model training, which poses significant challenges in real-world scenarios where large-scale data collection is often impractical. Different from their works, our proposed framework leverages the synthesized data for the model pertaining which reduces the data required during the training of the classifier.

B. Masked Model Pretraining

Masked Image Modeling (MIM) [11], [16], [17] has become a powerful self-supervised pretraining strategy, inspired by Masked Language Modeling (MLM) [20]–[22] in NLP. In MIM, models are trained to reconstruct missing parts of an image by learning from the unmasked regions. Among the MIM approaches, Masked Autoencoders (MAE) [16] is a leading approach, where a transformer-based model is trained to predict masked image patches, using a high masking ratio. This forces the model to learn generalizable global or context features, which can be effectively fine-tuned for tasks like classification and segmentation. Beyond MAE, MultiMAE [11] extends this approach by handling multiple modalities (e.g., RGB, depth, and segmentation maps) and tasks simultaneously. MultiMAE trains the model to reconstruct masked patches across different modalities, making it more versatile and efficient in multimodal learning environments. This multi-task, multi-modal learning strategy improves the generalization capabilities of the model, especially in scenarios requiring multimodal input. These pretraining strategies demonstrate strong transfer learning potential, allowing models to be fine-tuned on various downstream tasks with minimal labeled data. In our paper, by leveraging a similar masked modeling strategy, our proposed model can capture rich, generalizable features that are crucial for tasks where labeled data is scarce or expensive to obtain.

C. Signal Data perturbation

Unlike images in computer vision, wireless signals contain significant variability due to their interactions with the environment, device placement, and user movements. Thus, perturbating wireless data must go beyond simple data transformations and should consider the physical properties and environmental characteristics of the signals. Several studies have developed unique data perturbation approaches tailored for wireless sensing applications. In RFBoost [18], a physical data perturbation framework was proposed for WiFi sensing, aiming to address data scarcity by leveraging the inherent data diversity of wireless signals. Specifically, RFBoost utilizes techniques based on time-frequency spectrograms of WiFi signals, exploring data diversity across different subcarriers, antennas, and time windows. By generating multiple informative spectrograms and effectively mixing them, RFBoost can significantly boost the dataset size, thereby improving model performance without additional data collection. For mmWave-based wireless sensing, a different approach to data perturbation has been proposed, as seen in the DI-Gesture system [19]. This method employs transformations like geometric translations, scaling, and noise elimination to simulate different distances, angles, and movement speeds of human gestures. The perturbation framework is carefully designed to account for specific characteristics of mmWave signals, such as varying angular resolution and radiated power, making it particularly effective for enhancing robustness in challenging scenarios like extreme sensing angles.

III. Methodology

A. Overview

In this section, we present an overview of our proposed mmAP framework, designed to address the data scarcity challenges inherent in mmWave-based HAR by integrating cross-modality data synthesis, multi-modal masked autoencoder pretraining, task-specific fine-tuning, and heatmap-based data perturbation. The framework begins by synthesizing a large-scale mmWave radar signal dataset from an extensive 3D human mesh dataset [8], compensating for the limited availability of real-world data. This synthetic dataset is then used for pretraining the model's encoder with a multi-modal masked autoencoder [11], utilizing an 83% masking rate to encourage learning robust cross-modal representations by predicting missing information from other modalities. To further enhance robustness, we apply heatmap-specific data perturbation techniques—including Gaussian noise addition, vertical rotation, and temporal cropping and expansion—during both the pretraining and fine-tuning phases. Finally, we perform task-specific fine-tuning using a small real-world dataset, incorporating a classification head for activity recognition to bridge the gap between simulated and real-world data, ensuring accurate recognition and classification of human activities.

B. Data Synthesis

The objective of this paper is to develop a robust and efficient framework for Human Activity Recognition (HAR) using mmWave radar signals. Collecting real-world mmWave radar data, however, is labor-intensive, requiring specialized hardware/software setups, subject participation, and meticulous environment preparation. These demands pose significant challenges in building a comprehensive mmWave radar dataset for training a robust human activity recognition model. To address this issue, we propose to leverage cross-modal signal synthesis which can simulate large-scale realistic synthesized mmWave radar signals from extensive 3D human mesh data using the physical simulator. The 3D human mesh data represents the human body through a collection of 3D triangulated faces in virtual space, and it can be sourced from extensive motion capture datasets or derived from large video datasets using mesh estimation algorithms [8]. Recent works [6], [7] have also demonstrated the feasibility of synthesizing realistic wireless signals from the 3D human mesh data. With this foundation, we are able to build a large synthesized mmWave radar dataset for human activity recognition tasks, serving as the data source for model pre-training.

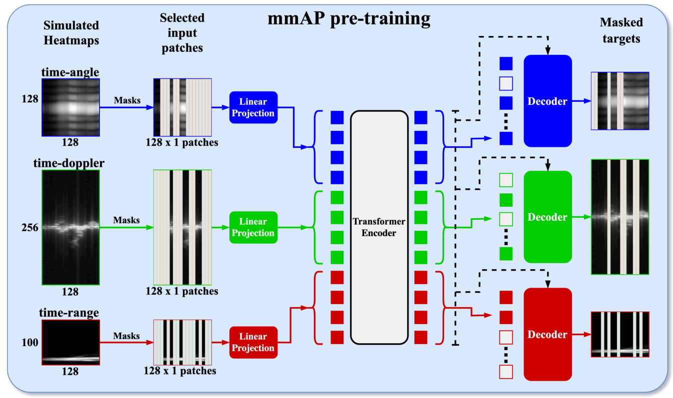

Fig. 1. A simulation heatmap dataset was utilized as input for pre-training the MultiMAE architecture, which processes multiple time-based heatmap modalities such as time-angle, time-doppler, and time-range. The data is divided into patches on the time dimension and linearly projected into tokens of a fixed dimension before being encoded using a Transformer. Task-specific decoders reconstruct the masked-out patches by initially performing a cross-attention step from queries to the encoded tokens, followed by processing with a shallow Transformer. The queries are composed of mask tokens, shown in gray, with task-specific encoded tokens added at their respective positions.

C. Model Pretraining

To effectively capture the activity pattern and integrate different input modalities from the mmWave signal, we follow the idea of MultiMAE [11], a masked-autoencoder-like unsupervised pretraining network designed to process masked multi-modal data as input and reconstruct the masked patches. This methodology allows the network to capture the intricate features of human activities from mmWave signals, enhancing the model's ability to recognize and classify diverse activities accurately.

Specifically, the inputs to the model are the three mmWave heatmap modalities [7]: $H_A \in \mathcal{R}^{A\times T}$, $H_D \in \mathcal{R}^{D\times T}$, and $H_R \in \mathcal{R}^{R\times T}$, which represent the time-angle heatmap, time-doppler heatmap, and time-range heatmap, respectively. The x-axis of these heatmaps represents time, while the y-axis captures the angle, doppler, and range information relevant to the subject's activity. Each heatmap is essentially a temporal concatenation of observations from consecutive timestamps. Different from the image-based patching solutions [11], [16], we segment the heatmaps into smaller vertical patches along the time dimension and selectively mask portions of them, taking into account all three heatmap modalities. This approach ensures that the physical patterns within each patch are preserved. The unmasked patches are then linearly projected into patch embeddings. These embeddings are subsequently processed by a transformer encoder designed to capture the inter-patch relationships. Considering that the self-attention layer in the transformer is order-agnostic, we incorporate positional encodings for each patch before feeding the patches to the transformer encoder.

The output embeddings from the transformer encoder are directed to modality-dependent decoders that aim to reconstruct the masked patches based solely on the observations from the unmasked patches. This cross-modal self-supervised reconstruction process compels the model to learn cross-modal relationships by effectively predicting missing information in one modality using the available information from other modalities, cultivating a robust feature representation for activity recognition from the heatmap modalities.

D. Data perturbation

While the large synthetic dataset substantially mitigates the data scarcity challenge associated with mmWave radar datasets, additional data perturbation strategies can be employed to enhance data diversity and improve model generalizability. Inspired by image perturbation techniques commonly used in computer vision—such as adding noise, cropping, and flipping—to train more robust deep learning models, we propose three tailored perturbation strategies for mmWave radar heatmaps. These strategies are designed to preserve the physical validity of the heatmaps while introducing meaningful variations: adding random Gaussian noise, row shifting, and cropping & stretching. For the random Gaussian noise strategy, we generate individual Gaussian noise for each input modality heatmap and add this noise to the original heatmap. This introduces variability and mimics real-world signal disturbances, enhancing the model's noise tolerance. Row shifting involves applying a cyclic shift along the y-axis of the time-range heatmap. This adjustment reflects real-world scenarios where the subject may perform the same activity at different locations, while maintaining consistent patterns in time-angle and time-doppler. Augmenting the time-range heatmap in this manner increases the model's robustness to variations in activity distance across training and testing environments. The cropping & stretching perturbation involves cropping a segment along the time dimension of the heatmap and resizing it to the original shape. This technique addresses variations in activity speed, as it simulates performing the same activity at different paces. This not only improves the model's ability to handle temporal variations but also its overall robustness and generalizability.

E. Fine-tuning

The objective of this section is to fine-tune the model, initially pre-trained with a large augmented synthetic dataset, using a smaller set of real-world collected mmWave radar signals for a specific task. While the augmented synthetic dataset effectively trains the model to robustly represent mmWave radar signals, a discernible gap remains between the simulated data and real-world datasets, which may impede the model's ability to accurately capture real-world radar features. Additionally, the model in the pre-training stage is class-agnostic; it is designed to generate feature representations of the mmWave radar signals without the capability to predict specific class labels. Consequently, it is essential to incorporate an additional stage that leverages real-world collected labeled data to fine-tune the pre-trained model for a specific downstream task.

Specifically, we utilize the pre-trained model as a foundation, initializing parameters and specifically training the last classification head with real-world data. This strategy allows us to capitalize on the strong feature representability from the pre-trained model to generate feature embeddings, subsequently fine-tuning a classification head with minimal parameters to predict the class labels. By focusing the fine-tuning process solely on the classification head, we enhance efficiency and mitigate the risk of catastrophic forgetting that could arise from fine-tuning the entire model.

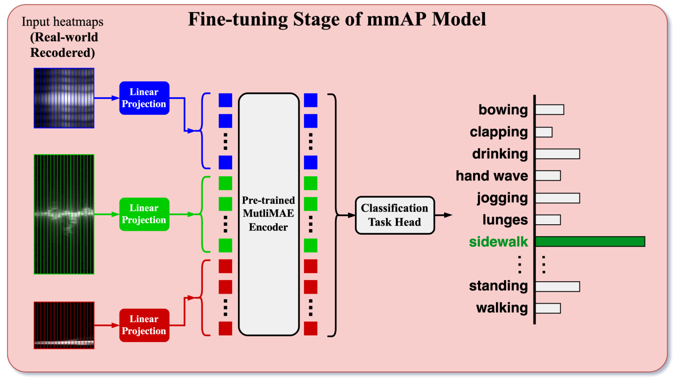

Fig. 2. A real-world mmWave dataset comprising 12 distinct activity classes was employed. An additional linear task-specific classification head was fine-tuned to predict the distribution of possible activities performed corresponding to multi-modality input heatmaps. The classification head was in conjunction with the pre-trained encoder parameters.

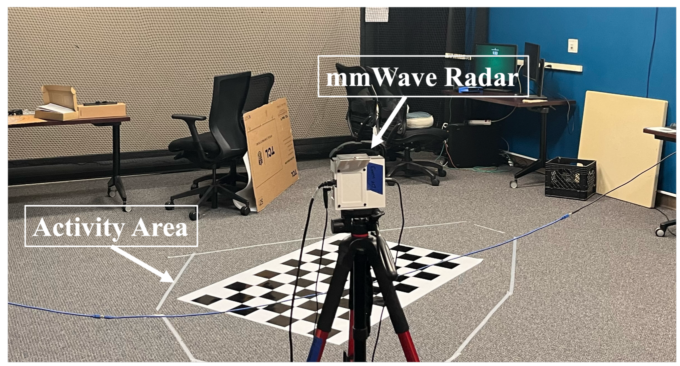

Fig. 3. The real-world testbed and the data collection scenario, where the activity area and the mmWave radar are indicated.

IV. Experiments

A. Data Collection & Model Complexity

As illustrated in Figure 3, to train our mmAP model, we conducted a data collection with 2 volunteers participating in 12 activities within a laboratory setting. The radar was positioned 1 meter above the ground, with activities performed 3 meters away. The activities included walking clockwise and anticlockwise, standing, sitting down and standing up from a chair, walking sideways, lunges, jogging, hand waving, drinking water, clapping, waving both hands, and bowing. Overall, we collected 256 data samples for each activity and used half of them for training and half of them for evaluation. All experimental procedures were conducted in compliance with institutional IRB policies.

TABLE I
ACTIVITY CLASSIFICATION RESULTS

Model Setting	Top-1 Accuracy (↑)
Plain Model	78.71%
Plain Model + Signal Perturbation	51.69%
Pre-training + Fine-tuning	82.23%
mmAP (time-angle only)	81.51%
mmAP (time-doppler only)	82.22%
mmAP (time-range only)	81.70%
mmAP	**84.54%**

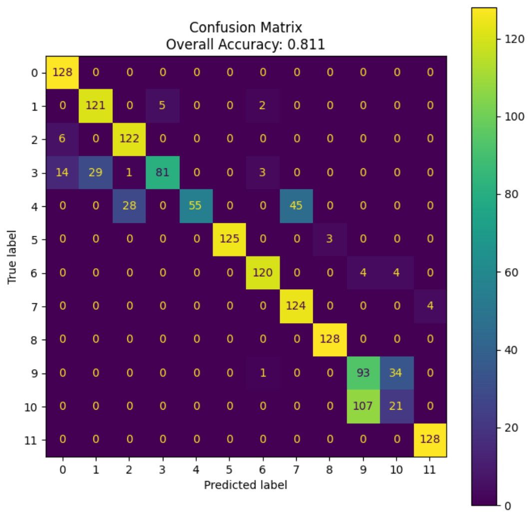

Fig. 4. The confusion matrix for the final mmAP model setting comprising pre-training on simulated data, followed by finetuning on real data for the entire model, with signal perturbation applied during data loading for both phases demonstrates enhanced classification accuracy and balanced performance across all activity classes.

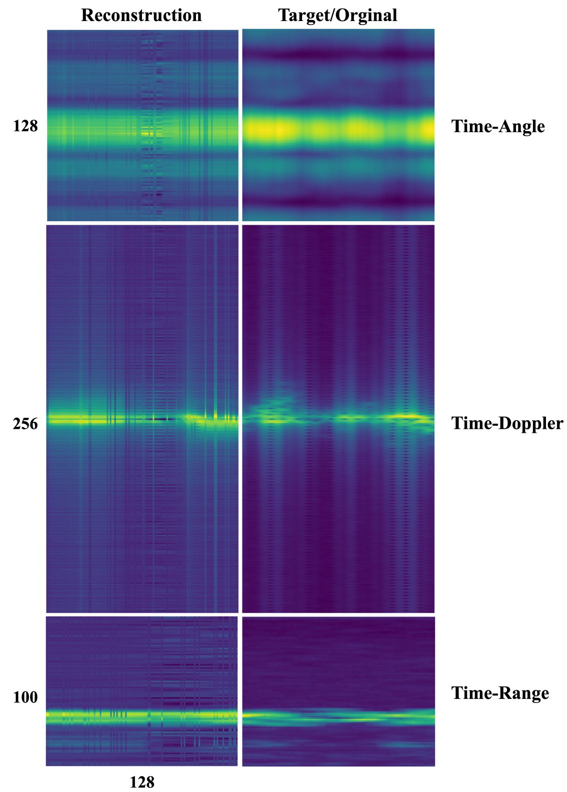

Fig. 5. The figure demonstrates the mmAP network's pre-training objective. Original heatmaps serve as input and reconstruction targets, with variable heights and fixed 128 as widths for each modality. Patches are dynamically selected and partially masked. Validation heatmaps from a simulation dataset, along with loss values, verify the model's learning accuracy in self-supervised conditions. The left side displays the final reconstructed results from the trained model on real-world data.

B. Model Settings

Plain Model: This baseline approach trains the model using only real-world collected data.

Plain Model + Signal Perturbation: This approach integrates data perturbation into the training process using real-world collected data.

Pre-training + Fine-tuning: This approach utilizes a large synthetic dataset for model pre-training, followed by fine-tuning with real-world data, without incorporating data perturbation.

mmAP Single Modality Only: This baseline is similar to the baseline Pre-training + Fine-tuning, however, only a single modality is used as the model input.

mmAP: This is the full implementation of our proposed mmAP framework, which includes synthetic data pre-training, real-world data fine-tuning, and data perturbation strategies.

C. Activity Recognition Results

In this section, we evaluate the overall performance of our proposed mmAP framework and compare it with baseline approaches, as detailed in Table I. Our mmAP framework significantly outperforms the baseline models, achieving the best results among the configurations tested. It can be seen that the baseline Plain Model, which trains solely with limited real-world mmWave radar data, does not perform well. Due to the limited real-world training data, the signal perturbation performed on the plain model impairs the model performance. In contrast, our pre-training and fine-tuning scheme substantially surpasses the Plain Model, demonstrating the effectiveness of using a large synthesized dataset for pre-training to generate robust feature representations. This approach also employs real-world data for fine-tuning, effectively mitigating the simulation-to-reality gap. Additionally, in the proposed mmAP model, signal perturbation has been demonstrated to be a useful factor in enhancing model performance, significantly increasing the model's robustness against variations in activity paces and locations. Furthermore, the performance of the mmAP framework decreases when only one modality heatmap is used as input. This outcome highlights the effectiveness of the mmAP network, which strengthens multi-modal integration, underscoring the importance of leveraging multiple modalities for improved accuracy and robustness.

D. Qualitative Reconstruction Results

In this section, we qualitatively evaluate the reconstruction results from the mmAP model, as illustrated in Figure 5. The rows from top to bottom display the time-angle, time-doppler, and time-range heatmaps, respectively. The columns from left to right depict the reconstructed heatmap alongside the original heatmap. The reconstructed signals notably retain the key semantic information from the original images, despite a high masking rate. This demonstrates the model's capability to capture essential semantic structures from the real-world heatmap, which facilitates robust feature representation learning from a limited dataset.

V. Conclusion

To address the scarcity of mmWave sensing data that limits the generalizability of existing HAR models, we proposed mmAP, a novel mmWave-based data augmentation and pre-training framework designed to enhance model performance for mmWave-based HAR tasks. By combining cross-modality data synthesis, multi-modal masked autoencoder pretraining, task-specific fine-tuning, and heatmap-based data perturbation, mmAP effectively trains a robust and effective activity recognition model. Our evaluations demonstrate that mmAP significantly enhances model robustness and accuracy, thus advancing the potential of mmWave-based HAR for real-world applications.

Acknowledgment

This work is supported in part by the US National Science Foundation under grant NSF IIS-2348427.

References

[1] Serpush, Fatemeh, et al. "Wearable sensor-based human activity recognition in the smart healthcare system." Computational intelligence and neuroscience 2022.1 (2022): 1391906.

[2] Hussain, Altaf, et al. "Vision transformer and deep sequence learning for human activity recognition in surveillance videos." Computational Intelligence and Neuroscience 2022.1 (2022): 3454167.

[3] Niu, Hao, et al. "Multi-source transfer learning for human activity recognition in smart homes." 2020 IEEE international conference on smart computing (SMARTCOMP). IEEE, 2020.

[4] Jin, Feng, Arindam Sengupta, and Siyang Cao. "mmfall: Fall detection using 4-d mmwave radar and a hybrid variational rnn autoencoder." IEEE Transactions on Automation Science and Engineering 19.2 (2020): 1245-1257.

[5] Zhao, Peijun, et al. "Heart rate sensing with a robot mounted mmwave radar." 2020 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2020.

[6] Xue, Hongfei, et al. "Towards generalized mmwave-based human pose estimation through signal augmentation." Proceedings of the 29th Annual International Conference on Mobile Computing and Networking. 2023.

[7] Cao, Qiming, et al. "mmCLIP: Boosting mmWave-based Zero-shot HAR via Signal-Text Alignment." Proceedings of the 22nd ACM Conference on Embedded Networked Sensor Systems. 2024.

[8] Mahmood, Naureen, et al. "AMASS: Archive of motion capture as surface shapes." Proceedings of the IEEE/CVF international conference on computer vision. 2019.

[9] Xue, Hongfei, et al. "mmMesh: Towards 3D real-time dynamic human mesh construction using millimeter-wave." Proceedings of the 19th Annual International Conference on Mobile Systems, Applications, and Services. 2021.

[10] Xue, Hongfei, et al. "M4esh: mmwave-based 3d human mesh construction for multiple subjects." Proceedings of the 20th ACM Conference on Embedded Networked Sensor Systems. 2022.

[11] Bachmann, Roman, et al. "Multimae: Multi-modal multi-task masked autoencoders." European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2022.

[12] Zhang, Jia, et al. "A survey of mmWave-based human sensing: Technology, platforms and applications." IEEE Communications Surveys & Tutorials (2023).

[13] Singh, Akash Deep, et al. "Radhar: Human activity recognition from point clouds generated through a millimeter-wave radar." Proceedings of the 3rd ACM Workshop on Millimeter-wave Networks and Sensing Systems. 2019.

[14] Zeng, Xianlin, Yiming Shi, and Anfu Zhou. "Multi-har: Human activity recognition in multi-person scenes based on mmwave sensing." 2022 IEEE 8th International Conference on Computer and Communications (ICCC). IEEE, 2022.

[15] Liu, Haipeng, et al. "Real-time arm gesture recognition in smart home scenarios via millimeter wave sensing." Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies 4.4 (2020): 1-28.

[16] He K, Chen X, Xie S, et al. Masked autoencoders are scalable vision learners[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022: 16000-16009.

[17] Hondru V, Croitoru F A, Minaee S, et al. Masked Image Modeling: A Survey[J]. arXiv preprint arXiv:2408.06687, 2024.

[18] Hou, Weiying, and Chenshu Wu. "RFBoost: Understanding and Boosting Deep WiFi Sensing via Physical Data Augmentation." Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies 8.2 (2024): 1-26.

[19] Li, Yadong, et al. "Towards domain-independent and real-time gesture recognition using mmwave signal." IEEE Transactions on Mobile Computing 22.12 (2022): 7355-7369.

[20] Ghazvininejad M, Levy O, Liu Y, et al. Mask-predict: Parallel decoding of conditional masked language models[J]. arXiv preprint arXiv:1904.09324, 2019.

[21] Devlin J. Bert: Pre-training of deep bidirectional transformers for language understanding[J]. arXiv preprint arXiv:1810.04805, 2018.

[22] Liu Y. Roberta: A robustly optimized bert pretraining approach[J]. arXiv preprint arXiv:1907.11692, 2019.