FiLM-Conditioned Multimodal Learning for Radar-Based Human Activity Recognition

Sheriff Murtala[1], Soojung Hur[1], Ingyu Lee[1], and Gyu Sang Choi[2*]

[1]School of Computer and Engineering
[2]Information and Communication Engineering
Yeungnam University

Abstract

Radar-based human activity recognition (HAR) increasingly leverages multiple signal representations such as range-Doppler spectrograms, range-angle heatmaps, and point clouds. However, fusing these heterogeneous modalities remains challenging due to representation-specific feature distributions. We propose FiLM-FusionNet, a multimodal deep learning architecture that employs Feature-wise Linear Modulation (FiLM) to adapt intermediate features according to representation context, enabling robust and efficient fusion. Experimental results demonstrate that FiLM conditioning significantly improves classification accuracy compared to single-modality baselines and conventional fusion strategies. Ablation studies confirm that FiLM and multimodal integration provide complementary gains, supporting accurate radar-based HAR for real-world applications.

Ⅰ. Introduction

Human activity recognition (HAR) using radio frequency (RF) and millimeter-wave (mmWave) radar has gained significant momentum due to its privacy-preserving nature, robustness to lighting or occlusion, and suitability for smart environment applications. Device-free sensing surveys emphasize that deep learning and transfer learning have become central to RF-based HAR, especially in addressing challenges such as domain shift, multimodal variability, and generalization across environments [1]. Similarly, recent AIoT-focused reviews highlight the need for scalable, edge-deployable architectures capable of handling practical sensing constraints in real-world IoT systems [2].

In radar based HAR specifically, earlier benchmarking and community-driven challenges played an important role in establishing evaluation standards. The Human Activity Radar Challenge introduced a unified assessment framework using the Glasgow "Radar Signatures of Human Activities" dataset, demonstrating the progress and limitations of existing algorithms, especially under varied activity types and subject conditions [3], [4]. Parallel efforts have expanded multimodal sensing datasets. For example, OPERAnet integrates RF and vision-based sources to encourage cross-sensor and cross-representation research in human activity monitoring [5]. The Open Radar Initiative further addressed fragmentation in radar datasets by promoting standardized acquisition procedures and a publicly accessible micro-Doppler benchmark [6].

Deep learning models designed for radar HAR continue to evolve toward more compact, accurate, and robust architectures. Tan *et al.* proposed a two-stream CNN- BiGRU model that leveraged magnitude and phase features for micro-Doppler classification, demonstrating the importance of combining spatial and temporal cues [7]. Ayaz *et al.* performed a systematic comparison of CNN architectures including MobileNetV2, VGG-16/19, and ResNet-50 across STFT, SPWVD, and time-range representations, reporting strong performance efficiency trade-offs, relevant for real-time radar HAR [8]. More recently, Wu *et al.* introduced RadMamba, a radar-adapted Mamba State-Space architecture achieving competitive accuracy with drastically reduced parameters, demonstrating the feasibility of efficient on-device radar HAR [9].

Despite these advancements, cross-device generalization in mmWave HAR remains largely unsolved. Most existing datasets rely on a single radar type or fixed configuration, making it difficult to evaluate how modulation schemes, antenna geometry, center frequencies, and data representations influence performance. This limitation directly motivated the creation of the MM-DCDR dataset, which systematically captures variability across radar hardware and signal representations. MM-DCDR contains 352k frames collected from 11 subjects performing 8 common actions at 1 m and 2 m distances, using TI AWR1843 (77- 81 GHz FMCW) and IMAGEVK-74 (63- 67 GHz SFCW) radars. These sensors differ substantially in array configuration (AWR1843: 3 TX × 4 RX; IMAGEVK-74: 20 TX × 20 RX), modulation format, frame rate, angular resolution, and sampling parameters [10]. The dataset also offers three complementary representations, range- Doppler maps, range- angle heatmaps, and point clouds, enabling researchers to systematically study representation-device interactions.

The corresponding MM-DCDR benchmark paper further emphasizes that, although many HAR models optimize well for a single radar system, they do not generalize across devices or representation domains,

underscoring the need for architectures that explicitly incorporate device-aware processing [10]. This naturally motivates the development of models equipped with mechanisms such as sensor conditioning, representation fusion, and device-invariant feature learning to tackle cross-radar heterogeneity.

Contribution. In response, this work introduces MM-DCDR-FusionNet, a multimodal deep learning framework designed for HAR. The model integrates three parallel encoders for range-Doppler, range-angle, and point-cloud data, combined with a sensor-conditioned Feature-wise Linear Modulation (FiLM) module that adapts features based on the radar source. This architecture is aligned with current trends in radar HAR toward multimodal fusion, compact representations, and device-aware adaptation.

Ⅱ. Method

FiLM-FusionNet is a multimodal deep learning framework designed to integrate heterogeneous radar representations for human activity recognition (HAR). The central challenge addressed is the mismatch in feature distributions across representations such as range-Doppler (RD) and range-angle (RA) heatmaps, and radar point clouds. To mitigate this, Feature-wise Linear Modulation (FiLM) is employed to adapt intermediate features dynamically, enabling robust fusion across modalities while maintaining a compact computational footprint.

The architecture comprises three parallel encoders tailored to distinct radar representations. The RD encoder utilizes a convolutional backbone augmented with squeeze-and-excitation (SE) blocks [11] to emphasize informative micro-Doppler bands and suppress less relevant activations. The RA encoder mirrors this design to capture spatial structure and angular cues characteristic of RA heatmaps. For point cloud input, a PointNet-style multilayer perceptron [12] aggregates per-point attributes-geometric coordinates, Doppler, and intensity into a fixed length embedding through global max pooling. Each encoder outputs a compact representation that preserves the salient information required for downstream classification.

Cross-modality alignment is facilitated by FiLM-based conditioning applied to intermediate feature maps in the RD and RA streams. Channel-wise affine transformations are generated from learnable embeddings that encode representation context, allowing features to be modulated according to the statistical characteristics of each modality. This conditioning enhances compatibility between modality embeddings in the shared latent space without reliance on device specific metadata, thereby focusing the adaptation on representation differences rather than hardware identities.

Following conditioning, modality embeddings are concatenated into a unified latent vector and processed by a lightweight multilayer perceptron that balances capacity and efficiency before producing class logits via a softmax layer. Inputs are normalized to fixed dimensions to enable batched training: RD and RA heatmaps are resized via interpolation, and point clouds are padded or subsampled to a consistent number of points. Data augmentation is intentionally omitted to isolate the effects of FiLM conditioning and multimodal fusion on recognition performance.

The overall design emphasizes modularity and compactness, enabling consistent training and inference across heterogeneous radar representations without reliance on data augmentation. To isolate the impact of representation-aware conditioning and multimodal fusion, inputs are standardized in size, and training employs a cross-entropy objective with Adam optimization.

III. Result

Performance evaluation is conducted on radar datasets, MM-DCDR, which includes eight activity classes. The overall accuracy and per-class accuracy are reported to capture variability across different motion patterns. Comparisons include single-modality baselines and the proposed FiLM-conditioned multimodal model, followed by ablation studies that examine the impact of FiLM layers and modality fusion. These results provide a comprehensive view of how representation-aware conditioning improves classification accuracy and robustness across heterogeneous radar representations.

All experiments were conducted on the MM-DCDR dataset comprising eight activity classes. The FiLM-FusionNet model and single-modality baselines were trained for 10 epochs using the Adam optimizer with a learning rate of 1×10^{-3}, batch size of 16, and input resolution of 128×128 for image-based representations for the heatmaps. Point cloud inputs were standardized to a fixed number of points, and no data augmentation was applied to isolate the effect of FiLM conditioning and multimodal fusion. The loss function was cross-entropy.

From Table 1 and 2, FiLM-FusionNet achieves the highest accuracy across all eight classes, with most activities exceeding 98% and four classes reaching 100%. The RD only variant remains competitive and even surpasses the full model on UpDown and TurnChair, underscoring the strong discriminative power of spectrogram features for periodic or angular motions. SparsePointNet [13] consistently trails the FiLM-based models, especially on SitDown and TurnChair, highlighting the limitations of point-cloud-only representations for fine-grained temporal patterns. Overall, FiLM conditioning and multimodal fusion deliver superior robustness and accuracy compared to single-modality approaches under standardized training conditions.

Table 1. Overall classification performance on MM-DCDR (8 classes).

Metric	FiLM-FusionNet	FiLM-FusionNet (RD)	SparsePoint Net [13]
Accuracy	**0.9890**	0.9787	0.8918
Macro F1	**0.9885**	0.9780	0.8901
Params (M)	4.9		**0.3-0.5**

Table 2. Per-class accuracy (%) for eight activities in the MM-DCDR dataset

Classes	FiLM-Fusion Net	FiLM-Fusion Net(RD)	Sparse PointNet [13]
Bowing	**0.9916**	0.9832	0.9333
Waving	**1.0000**	0.9957	0.9100
UpDOwn	0.9811	**0.9937**	0.9018
LieDown	**1.0000**	0.9710	0.9444
SitDown	**0.9960**	0.9598	0.8571
Stand	**1.0000**	0.9890	0.9216
TurnChair	0.9341	**0.9725**	0.8557
Squatting	**0.9840**	0.9720	0.9018

IV. Conclusion

This work introduced FiLM-FusionNet, a multimodal deep learning architecture for radar-based human activity recognition that leverages Feature-wise Linear Modulation (FiLM) to adapt intermediate features across heterogeneous representations. Evaluated on the MM-DCDR dataset, the proposed model achieved state-of-the-art accuracy (98.9%) and macro F1 (98.85%), outperforming both range-Doppler-only and point-cloud-only baselines. Per-class analysis confirmed the robustness of FiLM conditioning and multimodal fusion, particularly for complex activities such as chair transitions and squatting. The compact design and standardized training configuration demonstrate that representation-aware modulation can significantly enhance radar HAR performance without reliance on data augmentation, providing a strong foundation for future work on real-world sensing scenarios.

ACKNOWLEDGMENT

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (NRF-2021R1A6A1A03039493, NRF-2021R1A2 B5B02086773).

References

[1] Yang J., Xu Y., Cao H., Zou H., and Xie L. “Deep learning and transfer learning for device-free human activity recognition: A survey.” Journal of Automation and Intelligence (2022). (Device-free HAR survey).

[2] Qi W., Xu X., Qian K., Schuller B. W., et al. “A review of AIoT based human activity recognition: From application to technique.” IEEE Journal (Review) (2024). (AIoT review for scalable, deployable HAR).

[3] Yang S., Le Kernec J., Romain O., Fioranelli F., Cadart P., Fix J., et al. “The Human Activity Radar Challenge: Benchmarking Based on the ‘Radar Signatures of Human Activities’ Dataset From Glasgow University.” IEEE Journal of Biomedical and Health Informatics 27, no. 4 (2023): 1813- 1824.

[4] Fioranelli F., Shah S. A., Li H., Shrestha A., Yang S., and Le Kernec J. “Radar signatures of human activities.” University of Glasgow Research Data (Dataset) (2019). DOI: 10.5525/gla.researchdata.848.

[5] Bocus M. J., Li W., Vishwakarma S., Kou R., Tang C., Woodbridge K., et al. “OPERAnet, a multimodal activity recognition dataset acquired from radio frequency and vision-based sensors.” Scientific Data 9 (2022): Article 474.

[6] Gusland D., Christiansen J. M., Torvik B., Gurbuz S. Z., Fioranelli F., and Ritchie M. “Open Radar Initiative: Large Scale Dataset for Benchmarking of micro-Doppler Recognition Algorithms.” In IEEE Radar Conference (RadarConf21) (2021). Art. no. 9455239.

[7] Tan T.-H., Tian J.-H., Sharma A. K., Liu S.-H., and Huang Y.-F. “Human Activity Recognition Based on Deep Learning and Micro-Doppler Radar Data.” Sensors 24, no. 8 (2024): 2530.

[8] Ayaz F., Alhumaily B., Hussain S., Imran M. A., Arshad K., Assaleh K., and Zoha A. “Radar Signal Processing and Its Impact on Deep Learning-Driven Human Activity Recognition.” Sensors 25, no. 3 (2025): 724.

[9] Wu Y., Fioranelli F., and Gao C. “RadMamba: Efficient Human Activity Recognition through Radar-based Micro-Doppler-Oriented Mamba State-Space Model.” arXiv:2504.12039 (2025). (Accepted to IEEE Transactions on Radar Systems).

[10] Gao Y., Geng R., Zhang D., Hu Y., Lin H., and Chen Y. “MM-DCDR: A Benchmark of Device Configuration and Data Representation for mmWave-Based Human Sensing.” 16th International Conference on Wireless Communications and Signal Processing (WCSP), 801- 806, 2024.

[11] Hu, J., Shen, L., & Sun, G. (2018). Squeeze-and-Excitation Networks. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).

[12] Qi, C.R., Su, H., Mo, K., & Guibas, L.J. (2017). PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 652-660.

[13] Chuanwei Ding, Li Zhang, Haoyu Chen, Hong Hong, Xiaohua Zhu, and Francesco Fioranelli, “Sparsity-based human activity recognition with pointnet using a portable fmcw radar,” IEEE Internet of Things Journal, vol. 10, no. 11, pp. 10024- 10037, 2023.