Yi Zhou[1], Miguel López-Benítez[1], Limin Yu[1], and Yutao Yue[1]

[1]Affiliation not available

March 18, 2024

Text2Doppler: Generating Radar Micro-Doppler Signatures for Human Activity Recognition via Textual Descriptions

Yi Zhou, Miguel López-Benítez, Limin Yu*, and Yutao Yue*

***Abstract*—Radar-based Human Activity Recognition (HAR) is popular because of its privacy and contactless sensing capabilities. However, a major challenge in this area is the lack of large and diverse datasets. In response, we present a novel framework that uses generative models to transform textual descriptions into motion data, thereby simulating radar signals. This approach significantly enriches the realism and diversity of the dataset, especially for infrequent but critical activities such as falls and abnormal walking. Textual descriptions capture the semantic complexity and ambiguity of actions, thereby improving intra-class diversity. Our framework scales the data generation process and improves simulation fidelity by controlling gait variation, multi-viewpoint adaptation and background noise modelling. The simulated micro-Doppler dataset can be used for model comparison and transfer learning to improve recognition in real-world data, even when available data samples are scarce. Our approach significantly mitigates the challenge of data shortages, enabling significant advances in activity recognition with limited samples.**

***Index Terms*—radar simulation, text-driven motion synthesis, human activity recognition**

I. Introduction

Radar sensing, known for its privacy and contactless sensing capabilities, has become a growing area of interest in Human Activity Recognition (HAR). The sensing pipeline can be divided into two paradigms: one based on high resolution point clouds, the other using Doppler velocity patterns [1], [2]. Radar point clouds are typically sparse due to the low angular resolution of radar and the weak reflection from the human body. Improving spatial resolution typically requires a larger aperture, which increases sensor size, power consumption and cost. A promising alternative approach is to use motion signatures for classification. Since frequency modulated continuous wave (FMCW) radar can measure Doppler velocity at high resolution, the micro-Doppler distribution of non-rigid body motion over time can serve as a distinctive motion signature for human activity.

A number of recent studies [1], [2] have harnessed the power of deep learning to classify radar micro-Doppler spectrograms. The prerequisite for deep learning is a high quality dataset. For example, the computer vision task relies on the rich resources of images collected from the Internet for training. In contrast, radar datasets are often orders of magnitude smaller than vision datasets. Data diversity is also critical for generalisation. Radar data collection typically involves volunteers performing a predefined set of activities. While this may be sufficient for simple activities such as walking, it is insufficient for less common activities such as falling or abnormal gait. These instructional activities tend to be unnatural, lack contextual richness and fail to represent the diverse range of human movement.

To address the challenges of data collection, several studies have focused on high-fidelity simulation for radar-based HAR tasks. This process involves two main steps. The first is the acquisition and extraction of skeletal information of the human pose, which can be accurately captured using marker-based motion capture systems [3] or estimated from other modalities such as video [4], [5] and IMU [6]. The second step involves radar simulation and signal processing to produce the micro-Doppler spectrogram. Sometimes post-processing [3] is applied to improve accuracy, using generative models to match the distribution of simulated and real recorded data. Although these methods expand the potential sources of human activity data, they still fall short in capturing critical activities such as falling, which are infrequent in daily life and therefore underrepresented in data collected by other modalities.

To overcome these difficulties, our work combines text-to-motion generation and physics-based simulation to generate the high-fidelity synthetic dataset. Specifically, we exploit knowledge from large language models (LLMs) and the power of generative models to transform textual descriptions into human motion. We then use simulation to generate high-fidelity radar micro-Doppler spectrograms. Our innovative approach aims to bridge the generalisation gap in the detection of rare but critical human activities such as falls and abnormal walking.

Manuscript received August 29, 2023; revised . This work received financial support from Jiangsu Industrial Technology Research Institute (JITRI) and Wuxi National Hi-Tech District (WND). (Corresponding author: Yutao Yue.)

Yi Zhou is with the Institute of Deep Perception Technology, JITRI, Wuxi 214000, China, and also with the XJTLU-JITRI Academy of Industrial Technology, Xi'an Jiaotong-Liverpool University, Suzhou 215123, China (e-mail: zhouyi1023@tju.edu.cn)

Miguel López-Benítez is with the Department of Electrical Engineering and Electronics, University of Liverpool, Liverpool L69 3GJ, UK, and also with the ARIES Research Centre, Antonio de Nebrija University, 28040 Madrid, Spain (email: m.lopez-benitez@liverpool.ac.uk)

Limin Yu is with the Department of Electrical and Electronic Engineering, School of Advanced Technology, Xi'an Jiaotong-Liverpool University, Suzhou 215123, China (email: limin.yu@xjtlu.edu.cn)

Yutao Yue is with the Thrust of Artificial Intelligence and Thrust of Intelligent Transportation, The Hong Kong University of Science and Technology (Guangzhou), Guangzhou 511400, China; Institute of Deep Perception Technology, JITRI, Wuxi 214000, China, and also with the XJTLU-JITRI Academy of Industrial Technology, Xi'an Jiaotong-Liverpool University, Suzhou 215123, China and Department of Mathematical Sciences, University of Liverpool, Liverpool L69 7ZX, UK (email: ytyue@ustc.edu)

The main contributions of this letter can be summarized as follows:

1) **Scalable Text-Based Motion Generation:** We propose an innovative pipeline for generating motion data from textual descriptions. This method significantly improves scalability and introduces a diverse range of motions that are aligned with common sense.
2) **Improved Simulation Fidelity:** Our research enhances the fidelity of motion simulation by integrating several key elements:
   - The design of motion prompts is carefully crafted to take into account both the context of the motion and fine-grained variation.
   - Gait and viewpoint adaptations are implemented to increase within-class diversity.
   - Background noise is modelled and added to the simulation data using a generative model.

The remainder of this letter is organised as follows. Section II presents the proposed Text2Doppler pipeline. Section III presents the two applications of the simulation dataset, one for model comparisons and the other for transfer learning to real-world data. Finally, Section IV concludes the paper and summarizes the whole work.

II. Text2Doppler simulation pipeline

As shown in Fig. 1, the overall pipeline consists of several parts: First, we use an LLM to generate motion descriptions at scale. Second, we use a pre-trained text-to-motion model to generate human skeletons. Next, we control gait speed and view direction to further diversify the dataset. We then use a radar simulation module to generate micro-Doppler spectrograms. Finally, a background adaptation module is used to augment the data with a real-world noise distribution.

A. Motion Description Generation via LLM

Our simulation pipeline starts with the automatic generation of motion descriptions for daily indoor human activities using LLMs at scale. To avoid oversimplified motion descriptions, we provide a database of motion properties in the prompts as contextual information. These properties include elements such as action, direction, body part involved, objects interacted with, and a descriptive narrative, as detailed in Table I. The inclusion of these fine-grained properties significantly enriches the motion descriptions, resulting in more detailed and diverse motion simulations.

The benefits of using LLMs are twofold: First, it allows for easy scalability of the generation process. Second, LLMs allow us to incorporate 'common sense' into our simulations. Common sense in this context refers to the intuitive knowledge and understanding of the world that most people possess, which facilitates reasonable judgements and actions in everyday life. In the action recognition domain, pre-trained LLMs with common sense capabilities are guided to generate more realistic and contextually appropriate motion descriptions for a wide range of activities. For example, LLMs can be prompted to describe the abnormal gait characteristic of Parkinson's disease, a difficult task for a non-expert without specific medical knowledge.

TABLE I
Database of Properties

Properties	Items
Action	walk, fall, sit, bend, lie down, stand up, climb, stretch, squat, dance, run, jump, turn
Direction	left, right, clockwise, counterclockwise, anticlockwise, forward, back, up, down, straight
Body Part	arm, foot, feet, hand, leg, waist, knee
Object	stair, chair, floor, ball, handrail
Description	slowly, carefully, fast, careful, slow, quick, happily, angry, sad, happily, angrily, sadly

B. Text to Motion Generation

In the next step, we directly use the pre-trained model MoMask [7] for text-to-motion generation. MoMask uses a generative masked modelling framework to convert textual descriptions into 3D human motions. It is pre-trained on the HumanML3D dataset [8], which contains 14,616 motion clips, each annotated with 3-4 descriptive sentences. MoMask excels at generating high-quality and diverse 3D human motion that closely matches the input textual descriptions.

As a state-of-the-art text-to-motion generation model, MoMask is capable of generating vivid motion sequences from complex and detailed motion descriptions. However, our examination of the generated motions revealed certain shortcomings. For example, we found that falling motions were sometimes exaggerated, and sometimes these motions were completely absent from the generated results. To mitigate this problem, we have adopted a simple but effective method that focuses on analysing the head trajectory to detect falls. This approach helps to identify whether a fall has occurred in the motion sequence. In addition, we visually inspect the animations and eliminate any unrealistic movements.

C. View Control and Gait Speed Control

The different motion descriptions for a given activity class address the semantic diversity of real-world data. In addition, we address some physics-based diversity, including the different viewing angles and different walking speeds for a given motion sequence. In the traditional pipeline, this physics-based diversity is approximated at the radar received data stage, where the reflection signals are mixed. In comparison, introducing diversity at the skeleton stage offers significant advantages by maintaining physical fidelity and interpretability.

Regarding the change in viewpoint, it is important to note that, as radar primarily measures radial velocity, different viewpoints can lead to variations in the motion pattern. As shown in Fig. 2 (a), to account for viewpoint changes, we first determine the main direction of the motion trajectory using Principal Component Analysis (PCA). The first two principal components represent the direction of motion. We then calculate the new viewpoint location based on the desired angular difference, ensuring that the distance from the start point of the motion remains unchanged.

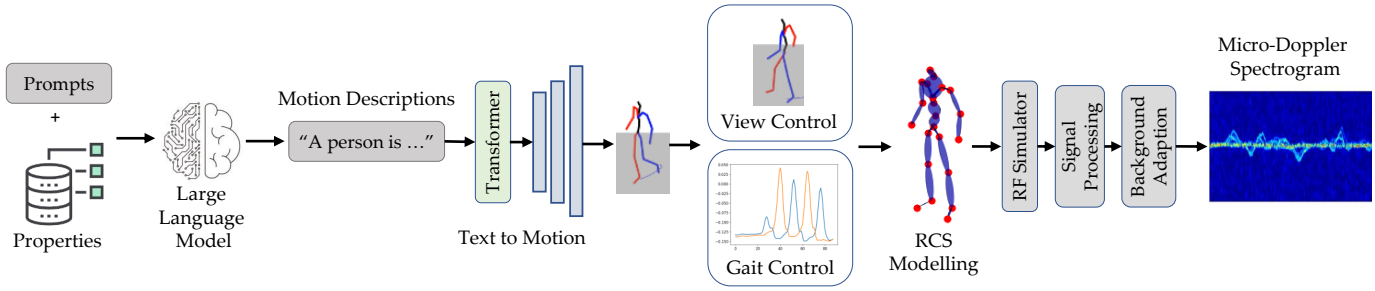

Fig. 1. Text2Doppler simulation pipeline

For gait speed control, as shown in Fig. 2 (b), we start by retrieving the foot trajectories of both feet. By identifying the peaks, we can divide the whole trajectory into distinct segments. These segments are then interpolated to either speed up or slow down the motion.

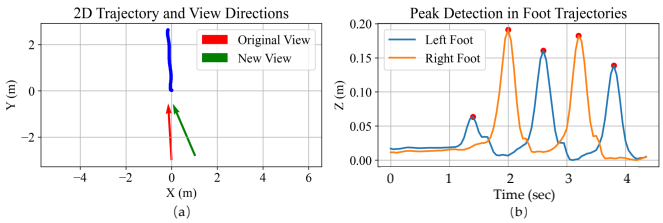

Fig. 2. Diversify the motion data (a) Change the viewing angle (b) Detect and change the gait period

D. Radar Data Simulation

The generated motion data is stored as BVH (BioVision Hierarchy format) files. These files contain the motion data, which is a sequence of frames, each of which contains the position and orientation of each body segment in the skeleton. In line with other gesture simulators [3], [9], we approximate the body segments as ellipsoids. Assume that each body part segment can serve as an ellipsoid parameterized by two semi-axes of equal length a and a principal semi-axis of length c. The radar cross section (RCS) σ of the i-th body segment can then be approximated by the following equation [9]:

$$\sigma_i = \frac{\pi a_i^4 c_i^2}{\left(a_i^2 \sin^2(\psi_i) + c_i^2 \cos^2(\psi_i)\right)^2} \tag{1}$$

where ψ_i describes the aspect angle of the principal axis.

The intermediate frequency (IF) signal for each body segment can then be calculated, and the total radar response is obtained by superimposing all the responses. Suppose we use a FMCW radar with carrier frequency f_c, bandwidth B, chirp repetition time T_{chirp}, chirp duration T_c and speed of light c. The IF signal for the l-th chirp can be modelled as

$$s_{\mathrm{IF},l}(t) = A \exp\left(2\pi \mathrm{j}\left[\frac{2f_c R}{c} - \left(\frac{2f_c v_{rad}}{c}\right) \cdot l \cdot T_{chirp} + \left(\frac{2BR}{cT_c}\right) \cdot t\right]\right) \tag{2}$$

For each body segment, the range R and the radial velocity v_{rad} can be calculated from the locations of the scattering points and the radar. The amplitude A can be calculated from the radar equation as

$$A = \sqrt{P_r G_{IF}} = \sqrt{P_t G_t G_r \frac{\lambda^2 \sigma A_e}{4\pi R^4} G_{IF}} \tag{3}$$

where P_r is the received power, P_t is the transmitted power, G_t is the gain of the transmitting antenna, G_r is the gain of the receiving antenna, λ is the wavelength of the radar signal, σ is the RCS of the target, A_e is the effective aperture of the receiving antenna and G_{IF} is the gain of the IF amplifier.

Once the IF signal has been obtained, the next step is to perform analog-to-digital (ADC) sampling for each received channel. The data acquired by the ADC is then organized into a three-dimensional tensor. A range Fast Fourier Transform (FFT) is then performed along the fast time dimension to obtain the range profile, followed by a Moving Target Indication (MTI) filter to remove static clutter. Finally, a Short-Time Fourier Transform (STFT) is applied to extract the micro-Doppler spectrogram, given by

$$\mathrm{S}(t, f_d) = \mathrm{STFT}_{f \in w}(\mathrm{FFT}_{range}(R_{ADC})) \tag{4}$$

where w is the window size, t is the time index and f_d is the Doppler index.

E. Background Noise Generation

The above simulations are noise-free and only take into account the reflections from the human. Therefore, we need to add background noise to the generated data to improve the generalisation. The background noise for the micro Doppler spectrogram can be introduced by system noise, sensor noise and environmental factors. Instead of explicitly modelling these complex noises, we learn the background noise in a supervised manner. Specifically, we use real data collected in empty rooms to train a VQ-VAE [10] to learn the background reflections. As shown in Fig. 3, we add randomly generated background noise to the Doppler spectrogram during the data simulation process.

To model occlusions, such as those caused by room furnishings, we apply a temporal mask to the micro-Doppler spectrogram. Specifically, we randomly select windows in the temporal dimension and mask them by replacing them with background noise.

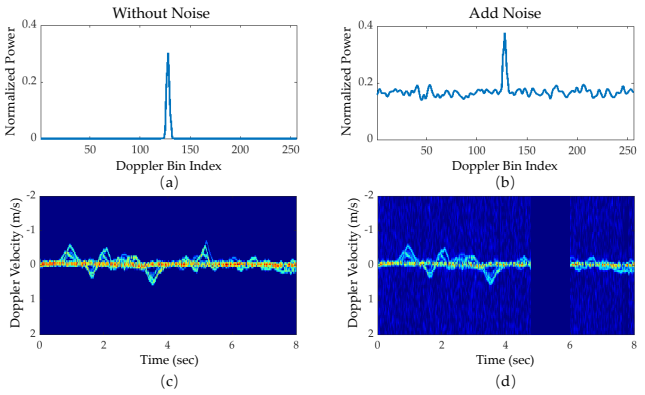

Fig. 3. Adding background noise: (a) and (b) are the Doppler value distribution for a single time step without and with noise. (c) and (d) are the spectrogram without and with data augmentation

III. Experiments and Results

A. Dataset Specification

We design 8 classes of daily activities for the dataset, including normal walking, abnormal walking, running, falling, sitting, bending, jumping and dancing. For each, we generate 200 motion descriptions per class and 20 data samples per description. We then manually inspect the generated motion video and remove the ambiguous data, resulting in a filtered high quality dataset with a size of 19,165 samples. We also generate a simple version of the dataset with the same size and data distribution, but less motion diversity, by generating the data samples from the same motion descriptions.

Due to the randomness in the generation, the motion data generated for each description will vary in length, direction and detailed content. In particular, for the safety critical activities such as falling and abnormal walking, we design diverse motion descriptions to account for the intra-class variance of these activities. For example, for falling, we consider three types of falls, including slipping, tripping and collapsing, as shown in Fig. 4. For abnormal walking, we describe different walking styles, such as Parkinson's gait, waddling gait, fixation gait, ataxic gait, scissor gait and stepping gait.

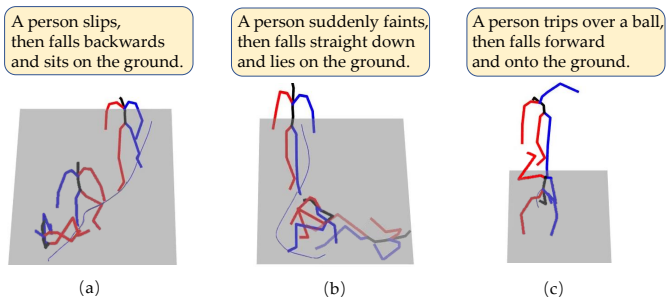

Fig. 4. Examples of different types of falling

For the simulation we configure the radar according to Table II. For the signal processing, the range FFT size is set to 256. For the STFT, the sample points and window size are set to 256. A significant overlap of 200 samples between windows is implemented to increase the resolution in the time-frequency spectrum, which is essential to capture subtle changes in the Doppler signatures. The simulation code is all written in Python, unifying the programming language of the entire text-to-Doppler pipeline. To speed up processing, we use multi-threaded programming. Using a 3.1 GHz CPU and configuring the simulator to run 16 processes in parallel, it takes about 20 seconds to generate 10 seconds of radar data.

TABLE II
Radar Configuration

Parameter	Value
Operating Frequency (GHz)	77
Frequency Slope (MHz/μs)	46.397
Sample Rate (ksps)	6847
Periodicity Per Frame(ms)	100
Number of ADC per Chirp	256
Number of Chirp per Frame	128

B. Model Comparisons

First, we evaluate the performance of the models on the Text2Doppler dataset. The test models include VGG-7 [11], ResNet-18 [12], ViT-tiny [13], CRNN [14], Conv1D-LSTM and Conv1D-ALSTM-FCN [1]. According to Table III, it can be observed that for the simple case, all evaluated models achieve a remarkably high accuracy. This observation suggests that datasets without motion diversity are insufficient to effectively benchmark the capabilities of these models. Conversely, in the context of more challenging datasets, it is evident that the larger convolutional networks significantly outperform both the lightweight RNN-based models and the ViT models. The RNN-based models may struggle to capture the complex motion patterns due to their lightweight 1D convolutional encoder, while the ViT models should require a larger dataset to learn the inductive bias.

TABLE III
Classification Performance

Model	Accuracy(%)		Params(G)	FLOPs(M)
	Text2D Simple	Text2D Hard		
Conv1D-LSTM	98.26	80.80	0.008	0.111
CRNN	99.65	86.95	0.415	0.895
Conv1D-ALSTM-FCN	99.48	82.79	0.104	0.127
VGG-7	98.44	87.87	2.095	0.298
ResNet-18	98.96	91.26	1.824	11.180
ViT-Tiny	99.48	79.64	1.433	7.261

From the confusion matrix shown in Fig. 5, it is clear that the improvement in performance of the convolutional network is primarily due to its improved ability to distinguish between walking and abnormal walking, and between falling and dancing. This distinction is critical as the detection of abnormal walking and falling is often the core functionality of commercial radar sensor applications. Consequently, while a more compact RNN-based model may be sufficient for detecting simple daily activities, the use of a larger convolutional network proves more effective in accurately identifying complex movements, such as dancing, and more nuanced activities, such as abnormal walking.

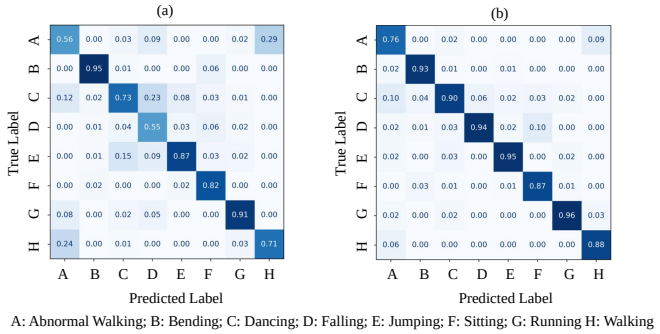

Fig. 5. Confusion matrix: (a) for CRNN and (b) for ResNet-18

C. Transfer Learning to Real-World Data

In this section, we investigate the effectiveness of applying a network pre-trained on a simulated radar dataset to real-world data, particularly when the real-world dataset is limited in size. For this analysis, we adopt the dataset described in [2], which uses a 77 GHz mmwave radar to collect 11 types of activity with only 60 samples per class. Direct training on this dataset proved challenging as the training process was unstable and did not converge, probably due to the limited size of the dataset. To overcome this problem, we adopt a transfer learning approach. First, we pre-train the model on a simulated dataset, followed by fine-tuning on the real dataset. For the fine-tuning phase, we use 80% of the data, reserving the remaining 20% for evaluation purposes.

Specifically, we use a pre-trained ResNet-18 model as a baseline to investigate the effectiveness of four different fine-tuning strategies, as described in [15]. These include linear probing, where only the linear layer is tuned; layer fine-tuning, where the last convolutional layer is tuned in addition to the linear layer; skip connection layer (SCL) fine-tuning, where a skip connection is introduced from the tuned layer to the penultimate layer; and full model fine-tuning.

The results of our transfer learning experiments, as detailed in Table IV, show that linear probing struggles to reach the expected performance levels and exhibits slow convergence rates. Conversely, fine-tuning the encoder layers significantly improves accuracy. The best results, characterised by fast convergence, are achieved by full model tuning. These results suggest that pre-training on simulated data is advantageous for stabilising convergence and transferring prior knowledge to the real-world dataset. However, if the encoder is trained only on simulated data, it may not effectively bridge the domain gap to real-world data. Thus, our analysis suggests that full fine-tuning of the model turns out to be the most effective strategy for such a transfer learning paradigm.

TABLE IV
PERFORMANCE IN REAL WORLD DATA

Method	Average Accuracy (%)	Average Number of Epochs
Linear Probing	70.31	74
Layer Fine Tuning	77.08	52
SCL Fine Tuning	77.60	15
Full Model Fine Tuning	84.72	18

IV. CONCLUSION

In this study, we present a text-to-radar simulation framework designed to enable large-scale, high-fidelity radar simulations that are particularly suited to HAR tasks. A notable advantage of this model is its ability to significantly improve data collection for less commonly observed activities, such as falling and abnormal walking. Our results suggest that the simulated dataset serves as a useful benchmark for model performance and can act as an effective pre-training dataset, facilitating rapid adaptation to real-world datasets. Future research could focus on speeding up the simulation process and considering motion interferences, such as pet motion, and room interferences, such as table fans, swinging curtains, etc.

REFERENCES

[1] Y. Zhou, M. López-Benítez, L. Yu, and Y. Yue, "Improving performance with feature enhancement nd ranking constraints for radar-based human activity recognition," in *IET International Radar Conference*, 2023, pp. 1–8.

[2] S. Z. Gurbuz, M. M. Rahman, E. Kurtoglu, T. Macks, and F. Fioranelli, "Cross-frequency training with adversarial learning for radar micro-doppler signature classification (rising researcher)," in *Radar Sensor Technology XXIV*, vol. 11408. SPIE, 2020, pp. 58–68.

[3] S. Vishwakarma, W. Li, C. Tang, K. Woodbridge, R. Adve, and K. Chetty, "Simhumalator: An open-source end-to-end radar simulator for human activity recognition," *IEEE Aerospace and Electronic Systems Magazine*, vol. 37, no. 3, pp. 6–22, 2021.

[4] K. Ahuja, Y. Jiang, M. Goel, and C. Harrison, "Vid2doppler: Synthesizing doppler radar data from videos for training privacy-preserving activity recognition," in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 2021, pp. 1–10.

[5] K. Deng, D. Zhao, Q. Han, Z. Zhang, S. Wang, A. Zhou, and H. Ma, "Midas: Generating mmwave radar data from videos for training pervasive and privacy-preserving human sensing tasks," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 7, no. 1, pp. 1–26, 2023.

[6] S. Bhalla, M. Goel, and R. Khurana, "Imu2doppler: Cross-modal domain adaptation for doppler-based activity recognition using imu data," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 5, no. 4, pp. 1–20, 2021.

[7] C. Guo, Y. Mu, M. G. Javed, S. Wang, and L. Cheng, "Momask: Generative masked modeling of 3d human motions," *arXiv preprint arXiv:2312.00063*, 2023.

[8] C. Guo, S. Zou, X. Zuo, S. Wang, W. Ji, X. Li, and L. Cheng, "Generating diverse and natural 3d human motions from text," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 5152–5161.

[9] N. Kern, J. Aguilar, T. Grebner, B. Meinecke, and C. Waldschmidt, "Learning on multistatic simulation data for radar-based automotive gesture recognition," *IEEE Transactions on Microwave Theory and Techniques*, vol. 70, no. 11, pp. 5039–5050, 2022.

[10] A. Van Den Oord, O. Vinyals *et al.*, "Neural discrete representation learning," *Advances in neural information processing systems*, vol. 30, 2017.

[11] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[12] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[13] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[14] B. Shi, X. Bai, and C. Yao, "An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 11, pp. 2298–2304, 2016.

[15] X. Li, S. Liu, J. Zhou, X. Lu, C. Fernandez-Granda, Z. Zhu, and Q. Qu, "Principled and efficient transfer learning of deep models via neural collapse," *arXiv preprint arXiv:2212.12206*, 2022.