Learning Unsupervised Representations for Sensing Humans with mmWave Radars

Nakul Singh [*1] **Aadesh Madnaik** [*1]

Abstract

mmWave radar data is rich in information related to humans like pose and movement. We also observe that collection of data is easy, however labelling the data is extremely hard owing to the fact that this data is not directly human interpretable unlike images and videos. In this project, we have built an unsupervised framework to extract spatial and temporal features of the data which are critical to downstream supervised tasks like pose estimation, tracking and identifying individuals. Using the latent space features, we are able to build more effective supervised learning frameworks that require 50 times fewer training samples than prior work while maintaining accuracy.

1. Introduction

millimeter-wave (mmWave) sensing (Abdu et al., 2021) is a relatively new modality of sensing which relies on electromagnetic waves whose frequency is in the order of 60 GHz and wavelength is in the order of millimetres. By virtue of the relatively smaller wavelength in comparison to tradition communication technology (below 6 GHz), mmWave sensing enables high spatial resolution at large distances. This technology has been commonly deployed on cars for use in early-warning systems (Liu et al., 2017) and military applications (Jr. et al., 1997).

Advances in sensing hardware have enabled finer resolution so the uses have expanded to human sensing. Some prospective deployments include rehabilitation systems (An & Ogras, 2021), gait analysis (Alanazi et al., 2022; Wu et al., 2021), sleep tracking (Zhou et al., 2021), and non-contact vital monitoring (Yang et al., 2016) as illustrated in Fig. 1. mmWave radars designed to be used at home are becoming

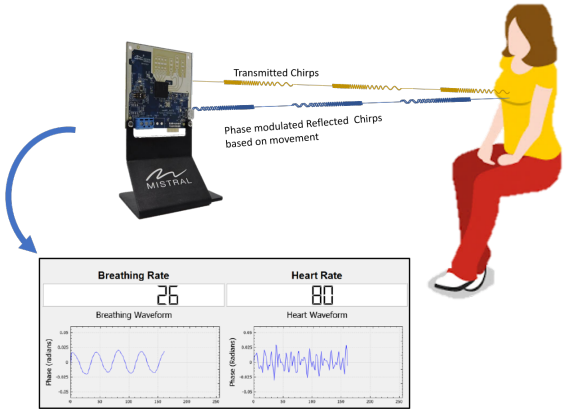

Figure 1. Vital monitoring with mmWave (Subramani, 2020)

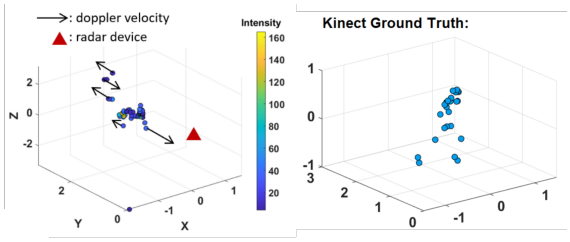

Figure 2. Sample point cloud frame illustrating doppler velocity, intensity and ground truth obtain using inertial sensors

quite popular with examples like Amazon Halo Rise (Giordano, 2023) being available for retail purchase for about $100. Fundamentally, all of these applications rely on high density point clouds which not only hold spatial information, but also contain Doppler velocity and reflection intensity as features. Each point cloud frame is structured as a set of 5D data points where the dimensions of the data point represent

$$(x, y, z, \text{Doppler}, \text{Intensity}).$$

Here, we observe a critical problem. There are several *distinct* use cases without a unifying framework. While the modality of the data remains the same, the application it is tailored to can vastly differ. Additionally, given the high dimensionality of the data, and the high frame-rate associated with such a motion-capture system, the input complexity in-

[*]Equal contribution [1]Department of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta GA. **AUTHORERR: Missing \icmlcorrespondingauthor.**

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

creases exponentially. Underlying the large amount of data is a relatively simple data generation mechanism. Since all of the measurements originate from humans in predictable and correlated movements, there is a low-dimensional latent structure to the data. Given the aforementioned thesis, we ask the question: *Is it possible to learn the underlying latent space and use it effectively in an application?*

In order to answer this question, we resort to developing a feature learning model with latent space capabilities which will help us in the goal of performing task-agnostic feature extraction. Building on this goal, we show the effectiveness of the framework by employing the learned features to estimate the 3D pose of humans.

2. Methodology

In this section, we first talk about our approach to the problem by discussing the advantages as well as the inefficiencies of prior work. Following the discussion, we explain our approach to the problem and then delve into the specifics of each of the individual pieces.

2.1. Prior Work

The problem of working with point clouds is not new. There are several works in literature, most notably PointNet and its variants (Qi et al., 2017a;b), which have built neural network architectures to natively support the geometric data structure of point clouds (in contrast to voxels or grids), and respect the permutation invariance of points in the input. These models have been quite successful in classification or semantic segmentation tasks. However, this functionality comes at a cost – all of the data used in training must be labelled. When it comes to building a task-specific model using mmWave point clouds, it falls short because creating labelled data is expensive, and at times impossible.

An example of a neural network which works with mmWave point clouds to estimate the 3D pose (skeleton model illustrated in Fig. 2) of individuals is MARS (An & Ogras, 2021). Fundamentally, this model takes the high density point cloud, which might contain up to a thousand individual points, and uses CNNs to visualize the skeleton model. The model was trained on 24,000 labelled frames which may not always be available! Ground truth labels were collected over the span of a few hours by placing inertial sensors on each of the 19 joints.

Realizing that the task of collecting ground truth labels for RF data is quite difficult, TGUL (Li et al., 2022) and RF-URL (Song et al., 2022) introduced the concept of an unsupervised framework which learns the latent space representation of RF data. The extracted features are low-dimensional and extremely useful in downstream supervised tasks. However, the model does not support 5D point clouds because it works on a different type of RF data. So, in our work, we use the same core concepts of TGUL and apply them to work on mmWave point clouds. Ultimately, we aim to balance the pros and cons of prior work by first building an unsupervised, generalized framework which learns the latent space representation without the need for labels. Next, we show the utility of the low-dimensional latent space by using the features to reconstruct the 3D skeleton model.

2.2. Overall Approach

Our overall approach is illustrated in Fig. 3. The first step in the overall approach is to learn the mapping to low-dimensional latent features of the mmWave point cloud. To learn this mapping, we draw inspiration from PointNet (Qi et al., 2017a) which is an autoencoder. Additionally, the autoencoder must be built in a way which is invariant to permutations of the point cloud. This condition is necessary because we are trying to understand the motion of humans, and the individuals could be facing any direction, not just the mmWave radar head-on. The next step is to choose a supervised task. There are several open problems in this domain which span reconstructing a skeleton model, classifying different poses, identifying people, etc. We pick the task of reconstructing a skeleton model primarily because of the availability of labelled data.

2.3. Approach for Learning Features

It is crucial to recognize that the processing of point clouds needs to remain unaffected by transformations that maintain the spatial relationships among points, such as rotation and translation. Additionally, it is important to note that the point clouds obtained from mmWave radar are not meshes or voxels; rather, they constitute an unordered collection of points. Also, these points have noisy features attached to them which are important to the physical world like doppler velocity (which measures the velocity vector of that point in space) and reflection intensity (which measures the relative size of the object and its material composition). The final latent features must encode this information too while ignoring the noise.

The solution PointNet++ came up with is to aggregate local and global information. PointNet captures local features using functions that operate only on one point at a time. To capture global features and relationships among points, PointNet++ aggregates local information from all the points. This is typically done by taking the maximum or average of the features computed by the MLPs across all points. This process is illustrated in Fig. 4 and expressed using a general function defined on a point set by applying a symmetric (permutation invariant) function on the elements in the set

$$f(\{x_1, \cdots, x_n\}) \approx g(h(x_1), \cdots, h(x_n)),$$

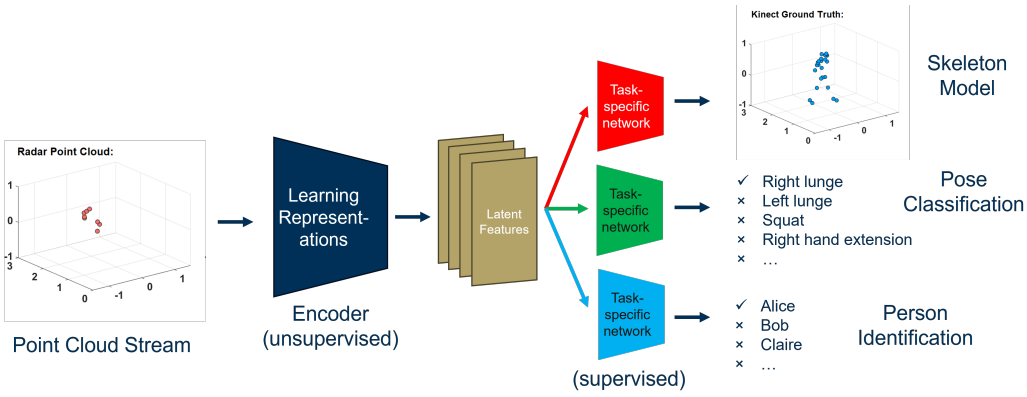

Figure 3. Learning the latent space representation using an encoder. Using the compressed feature map in downstream supervised tasks.

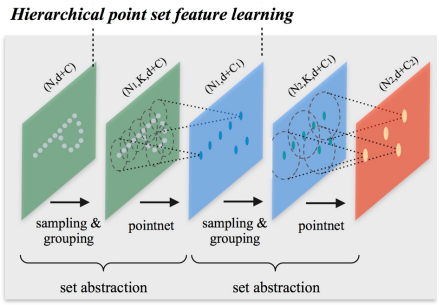

Figure 4. Model Architecture of PointNet++

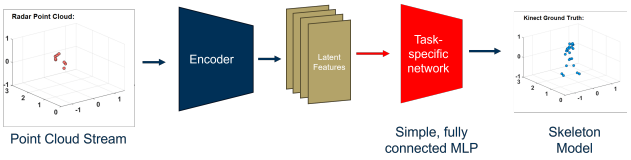

Figure 5. Inferring skeleton model from point cloud

where $f : 2^{\mathbb{R}^N} \to \mathbb{R}, h : \mathbb{R}^N \to \mathbb{R}^K$ and $g : \mathbb{R}^K \times \cdots \times \mathbb{R}^K \to \mathbb{R}$ is a symmetric function.

In an autoencoder framework, we try to reconstruct the same point cloud as the input. The reconstruction loss for point clouds is a new distance metric called Chamfer loss defined as

$$d_{CD}(S_1, S_2) = \sum_{x \in S_1} \min_{y \in S_2} \|x - y\|_2^2 + \sum_{y \in S_2} \min_{x \in S_1} \|x - y\|_2^2.$$

We observe that this loss is permutation invariant since it is defined as an aggregate over the set of points. In addition to this, we add another mean-squared error loss term for the doppler and intensity features. After training the autoencoder, we discard the decoder because only the encoder is of importance to us. This encoder, given a point cloud, returns the latent space representation which will be used in downstream tasks.

As a design choice, we set the number of input points in one frame to be 196. We observe that picking the 196 points with the largest intensity also describes human motion the best and also helps get rid of noisy reflections from surrounding objects. Practically, not all frames have 196 points. To preserve the spatial structure of the points without disturbing the model, we choose to pad the data with randomly selected points from the same frame. The size of the latent space is set to 64 to accommodate for all possible intricate movements of the targets.

2.4. Approach for Task-Specific Learning

We have picked the task of creating a skeleton model of a given point cloud. As we have observed in Fig. 2, there are very few human-interpretable features that are preserved in the point clouds. So, the task of mapping them to a skeleton model is crucial because a skeleton model forms the basis for a lot of tasks like pose classification and live-action computer generated graphics. There are 19 pre-defined joints, each with 3 coordinates, so a total of 57 quantities to be inferred from a point cloud that can be up to 196 points dense. To do so, we use a simple fully connected two-layer MLP as illustrated in Fig. 5. The loss under consideration is a mean-squared error loss between the ground truth labels and the predicted skeleton model defined as

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^{n} \|x_i - \hat{x}_i\|^2.$$

In case of another task, a very similar process can be followed. All that is required is accurate labelled data. Similar to the mapping from a 64-dimensional latent space to 57-dimensional skeletons, we could have another MLP which maps to another set of labels.

3. Results

The choice of the downstream supervised task was significantly influenced by the datasets available. Unfortunately, there are not many publicly available datasets for mmWave radar sensing where the data is collected from commercial off-the-shelf radars. Additionally, if the data is available, it may not be present in several modalities to establish ground truth accurately. The most promising and extensive dataset available is mRI (An et al., 2022) where 3D human poses are synchronously recorded in three modalities (mmWave, RGB-D, and inertial sensors). This formed the basis for all our testing. The mRI dataset consists of 24,000 frames recorded using synchronized feeds of 20 subjects performing rehabilitation exercises like limb extensions, lunges, and squats.

First, we split the dataset of 24,000 frames into two sets, one each for supervised and unsupervised training. We keep 500 frames for supervised training and 8000 frames for supervised validation. The remaining 15,500 frames are used for training the unsupervised auto-encoder (with 500 frames reserved for unsupervised validation at each epoch). We reserve a smaller number of frames in the validation set for unsupervised training since this task is more training intensive and implicit model validation of the encoder occurs while evaluating the validation split in the supervised training.

We ran the model on Google Colab where we trained the autoencoder for 50 epochs over 15,500 frames. The autoencoder mapped 196 points in the point cloud to a 64-dimensional latent space vector.

After training the unsupervised model, we fixed the weights and called it the task-agnostic network. Now, for the task of inferring the skeleton model, we chose to train only the supervised MLP on the remaining 500 frames. Using fifty times fewer frames, we achieved accuracy in joint prediction which matched the baseline MARS (An & Ogras, 2021). The dataset includes recording durations which have long periods of We make sure that the dataset is balanced to include a variety of poses.

In the baseline, MARS, using a fully supervised framework, the authors achieved an average localization error of 10.37 cm. Using our unsupervised pipeline we were able to achieve an average localization error of 10.58 cm.

Furthermore, we can deduce that the network learns features by the following experiment. We randomly initialize another network with the same architecture as the PointNet autoencoder. Now, we train two MLPs on the same data over 50 epochs, however, they use the features learned using the trained autoencoder and the untrained autoencoder respectively. The validation loss in the untrained model takes a clear hit because it does not learn features relevant to the

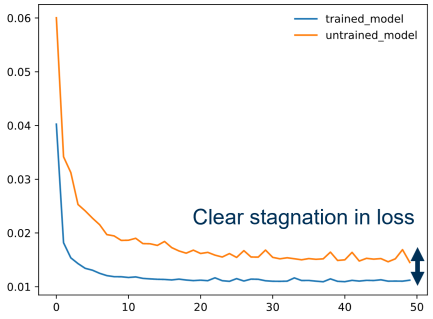

Figure 6. Comparison of validation loss for trained and untrained models

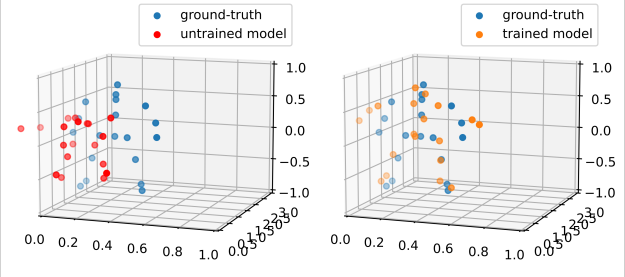

Figure 7. Sample projections of 3D skeleton model compared to ground truth

skeleton model, or even human point clouds in general. This is illustrated in Fig. 6. Ultimately, we observe that a network which has compressed the input into a smaller dimension requires significantly fewer samples to produce the same accuracy. In our case, the final downstream task requires 50 times fewer frames to train! A few samples are illustrated in Fig. 7 where we see that the untrained model is unable to accurately predict the skeleton model while the trained model which uses the learned features can.

4. Conclusions

In this work, we discuss a novel approach for mmWave point clouds of combining unsupervised and supervised learning. It uses an autoencoder inspired by PointNet, emphasizing permutation invariance to preserve spatial relationships. The autoencoder handles unordered point clouds with noisy features. The approach reconstructs a 3D skeleton model, utilizing a 64-dimensional latent space. The task-specific learning employs a two-layer MLP for mapping, adaptable to other tasks with labeled data.

The study uses the mRI dataset for mmWave radar sensing. An autoencoder trained on 15,500 frames creates a task-agnostic network. A supervised MLP on 500 frames infers the skeleton model with accuracy comparable to a fully supervised baseline. The unsupervised pipeline achieves a localization error of 10.58 cm. An experiment shows the importance of feature learning for the skeleton model.

References

Abdu, F. J., Zhang, Y., Fu, M., Li, Y., and Deng, Z. Application of deep learning on millimeter-wave radar signals: A review. *Sensors*, 21(6), 2021. ISSN 1424-8220. doi: 10.3390/s21061951. URL https://www.mdpi.com/1424-8220/21/6/1951.

Alanazi, M. A., Alhazmi, A. K., Alsattam, O., Gnau, K., Brown, M., Thiel, S., Jackson, K., and Chodavarapu, V. P. Towards a low-cost solution for gait analysis using millimeter wave sensor and machine learning. *Sensors*, 22(15), 2022. ISSN 1424-8220. doi: 10.3390/s22155470. URL https://www.mdpi.com/1424-8220/22/15/5470.

An, S. and Ogras, U. Y. Mars: mmwave-based assistive rehabilitation system for smart healthcare. *ACM Transactions on Embedded Computing Systems (TECS)*, 20(5s): 1–22, 2021.

An, S., Li, Y., and Ogras, U. mri: Multi-modal 3d human pose estimation dataset using mmwave, rgb-d, and inertial sensors. *Advances in Neural Information Processing Systems*, 35:27414–27426, 2022.

Giordano, M. Review: Amazon halo rise. 2023. URL https://www.wired.com/review/amazon-halo-rise/.

Jr., D. D. F., McMillan, R. W., Currie, N. C., Wicks, M. C., and Slamani, M.-A. Sensors for military special operations and law enforcement applications. In Watkins, W. R. and Clement, D. (eds.), *Targets and Backgrounds: Characterization and Representation III*, volume 3062, pp. 173 – 180. International Society for Optics and Photonics, SPIE, 1997. doi: 10.1117/12.276674. URL https://doi.org/10.1117/12.276674.

Li, T., Fan, L., Yuan, Y., and Katabi, D. Unsupervised learning for human sensing using radio signals. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 3288–3297, 2022.

Liu, G., Zhou, M., Wang, L., Wang, H., and Guo, X. A blind spot detection and warning system based on millimeter wave radar for driver assistance. *Optik*, 135:353–365, 2017. ISSN 0030-4026. doi: https://doi.org/10.1016/j.ijleo.2017.01.058. URL https://www.sciencedirect.com/science/article/pii/S0030402617300797.

Qi, C. R., Su, H., Mo, K., and Guibas, L. J. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 652–660, 2017a.

Qi, C. R., Yi, L., Su, H., and Guibas, L. J. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in neural information processing systems*, 30, 2017b.

Song, R., Zhang, D., Wu, Z., Yu, C., Xie, C., Yang, S., Hu, Y., and Chen, Y. Rf-url: unsupervised representation learning for rf sensing. In *Proceedings of the 28th Annual International Conference on Mobile Computing And Networking*, pp. 282–295, 2022.

Subramani, S. Using mmwave radar for vital signs monitoring. 2020. URL https://www.embedded.com/using-mmwave-radar-for-vital-signs-monitoring/.

Wu, J., Wang, J., Gao, Q., Pan, M., and Zhang, H. Path-independent device-free gait recognition using mmwave signals. *IEEE Transactions on Vehicular Technology*, 70 (11):11582–11592, 2021.

Yang, Z., Pathak, P. H., Zeng, Y., Liran, X., and Mohapatra, P. Monitoring vital signs using millimeter wave. In *Proceedings of the 17th ACM international symposium on mobile ad hoc networking and computing*, pp. 211–220, 2016.

Zhou, T., Xia, Z., Wang, X., and Xu, F. Human sleep posture recognition based on millimeter-wave radar. In *2021 Signal Processing Symposium (SPSympo)*, pp. 316–321. IEEE, 2021.