Towards Building a Foundation Model for Wireless Sensing: A Pilot Study

Omer Gokalp Serbetci*
Ming Hsieh Dept. of Elec. and Computer Eng.
University of Southern California
Los Angeles, CA, 90089
serbetci@usc.edu

Aditya V. Padaki
Amazon Lab126
Sunnyvale, CA, 94089
appadaki@amazon.com

Prasad K. Shamain
Amazon Lab126
Sunnyvale, CA, 94089
prasadks@amazon.com

Koushik A. Manjunatha
Amazon Lab126
Sunnyvale, CA, 94089
koushiam@amazon.com

Abstract

This paper presents a pilot study toward a foundation model for wireless sensing using FMCW radar. We propose a transformer-based architecture trained on data from mmWave sensor with self-supervised objectives designed to capture temporal–spatial signal characteristics without labels. Our framework introduces strategies for handling sparse channel representations and provides a unified normalization across diverse radar configurations, enabling task-agnostic representation learning. Experimental evaluation on presence detection shows that the learned embeddings generalize effectively and achieve competitive performance compared to pretrained models such as DINO and LWM. These results validate the feasibility of RF foundation models and highlight their potential to advance physical AI through adaptable and label-efficient wireless sensing systems.

1 Introduction

Wireless sensing builds on a long legacy of radar and indoor localization research and has emerged as a transformative technology that leverages radio frequency (RF) signals to perceive and understand the physical world without requiring dedicated sensors or cameras. Early radar systems in the 1970s employed microwave Doppler sensors for motion detection [1], and later in the 1990s UWB systems were used for through-wall imaging and indoor localization, achieving centimeter-level accuracy but requiring dedicated hardware [2, 3]. A pivotal shift occurred in 2010s when Adib and Katabi demonstrated that commodity Wi-Fi radios could enable through-wall sensing, detecting human motion and gestures using MIMO techniques [4]. In parallel, Frequency-modulated continuous-wave (FMCW) radar was initially developed for automotive and industrial applications. More recently, miniaturized millimeter-wave FMCW chips have enabled consumer-grade gesture sensing, most prominently in Google's Project Soli [5]. FMCW radar systems have since been successfully applied to indoor human activity sensing, including fall detection, skeletal pose estimation, and non-contact vital-sign monitoring, leveraging micro-Doppler features, point clouds, and phase-based signal processing [6–8]. Together, these developments have enabled wireless sensing to emerge as an important modality for device-free perception across consumer electronics, vehicles, and smart environments.

*Work done while the author was an intern at Amazon Lab126.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: AI4NextG.

Limitations of Current Approaches. Traditional wireless sensing approaches face significant scalability challenges that limit their practical deployment [9]. First, collecting labeled data for each specific use-case, environment, or scenario is expensive and time-consuming, often requiring weeks of data collection campaigns to achieve satisfactory performance [10, 11]. Second, cross-scenario data utilization remains severely limited—datasets collected for one task (e.g., gesture recognition) cannot effectively be leveraged for other sensing applications (e.g., presence detection), even when using the same hardware configuration. This leads to a fragmented ecosystem where small, task-specific models are trained independently on small datasets, resulting in limited generalizability and suboptimal resource utilization. Current state-of-the-art methods predominantly employ deep learning-based neural networks that function as specialized mappings from preprocessed data frames to task-specific probability distributions [12]. While these approaches have demonstrated success in controlled environments, they suffer from the fundamental limitation of requiring extensive manual feature engineering and task-specific model design for each new application domain [9].

Foundation Model Paradigm. Recent advances in foundation models have revolutionized multiple domains by demonstrating remarkable capabilities in learning generalizable representations from large-scale datasets [13–15]. In computer vision, models like DINO [16] have shown exceptional performance through self-supervised learning on image data, while masked autoencoders (MAE) [17] demonstrate the power of reconstruction-based objectives. In natural language processing, transformer-based models like BERT [18] and GPT [13] have achieved remarkable success through pre-training on large text corpora. Similarly, the wireless sensing domain has recently begun exploring foundation model approaches. [19] introduced the Large Wireless Localization Model (LWLM) for positioning tasks in 6G networks, employing spatial-frequency masked channel modeling and contrastive learning on channel state information (CSI) data. Similarly, recent work has explored masked autoencoders for wireless signal processing [20] and contrastive learning approach is employed in [21]. However, these approaches are typically do not address the broader challenge of unified representation learning across diverse radar-based sensing applications.

Our Approach: FMWS. In this work, we present FMWS (Foundation Model for Wireless Sensing), a pilot study exploring the feasibility of building a foundation model specifically designed for wireless sensing applications using FMCW radar sensors. Our approach addresses the critical data engineering challenges inherent in wireless sensing, including handling non-uniform data shapes across different radar configurations, developing effective tokenization strategies for radar signal representations, and designing domain-specific self-supervised learning objectives. We implement a transformer-based architecture trained on a small-scale dataset of over 300,000 frames collected from diverse radar sensing scenarios, including gesture recognition, presence detection and other applications. Our methodology incorporates novel data processing strategies for sparse radar channel representations and a unified framework for normalizing diverse radar configurations through patch-based tokenization.

Our key contributions include: (1) a comprehensive data engineering pipeline that handles the heterogeneous nature of radar data across different configurations and applications, (2) a transformer-based architecture with domain-specific training objectives including masked reconstruction, contrastive learning, and temporal prediction tasks, (3) experimental validation demonstrating improved performance over traditional task-specific approaches and competitive results against established foundation models adapted from other domains, and (4) empirical evidence showing the effectiveness of foundation model embeddings for downstream tasks including presence detection.

The results of this pilot study demonstrate that foundation models represent a promising direction for advancing wireless sensing technology, potentially enabling more robust, scalable, and versatile sensing systems that can adapt to new environments and applications with minimal retraining while significantly reducing the data collection burden for new deployments.

2 FMWS: Foundation Model for Wireless Sensing

FMCW Radar Signal Model. We use a linear-chirp FMCW radar with chirp duration T_c and chirp rate $\mu = B/T_c$, where B is the bandwidth. The complex baseband transmit signal is $s_{\mathrm{tx}}(t) = \exp\{j2\pi(f_0 t + \frac{1}{2}\mu t^2)\}$ for $0 \le t < T_c$, with start frequency f_0 and carrier f_c (wavelength $\lambda = c/f_c$). A target at range R and radial velocity v induces delay $\tau = 2R/c$ and Doppler $f_D = 2vf_c/c = 2v/\lambda$, giving $s_{\mathrm{rx}}(t) = \alpha \exp\{j2\pi[f_0(t-\tau) + \frac{1}{2}\mu(t-\tau)^2]\} \exp\{j2\pi f_D t\}$, where α captures path loss

and reflectivity. Dechirping (mixing with $s_{\text{tx}}^*(t)$ and low-pass filtering) yields the IF signal

$$s_{\text{IF}}(t) = \alpha \, \exp\{j2\pi[(\mu\tau + f_D)t - f_0\tau - \tfrac{1}{2}\mu\tau^2]\}.$$

The beat frequency within an up-chirp is $f_b = \mu\tau + f_D = (B/T_c)\,(2R/c) + 2v/\lambda$, where the first term relates to range and the second to velocity. Using up/down chirps allows range/velocity separation via $f_b^{\uparrow} = \mu\tau + f_D$ and $f_b^{\downarrow} = -\mu\tau + f_D$.

In a burst of N_c chirps, each chirp is separated by the pulse repetition interval (PRI) T_r (slow-time sampling period), so the m-th chirp starts at time $t = mT_r$. Doppler induces a slow-time phase term $\exp\{j2\pi f_D m T_r\}$. For L targets, the IF signal is

$$s_{\text{IF}}(m,t) = \sum_{\ell=1}^{L} \alpha_\ell \exp\{j2\pi[(\mu\tau_\ell + f_{D,\ell})t - f_0\tau_\ell - \tfrac{1}{2}\mu\tau_\ell^2]\} \cdot \exp\{j2\pi f_{D,\ell} m T_r\}. \tag{1}$$

With a fast-time sampling interval T_s, we denote n as the fast-time sample index within a chirp and m as the slow-time chirp index. The discrete-time IF signal becomes

$$s_{\text{IF}}[m,n] = \sum_{\ell=1}^{L} \alpha_\ell \exp\{j2\pi[(\mu\tau_\ell + f_{D,\ell})\, nT_s - f_0\tau_\ell - \tfrac{1}{2}\mu\tau_\ell^2]\} \cdot \exp\{j2\pi f_{D,\ell} m T_r\}. \tag{2}$$

Here $n = 0, \ldots, N_s - 1$ indexes range samples (fast time) and $m = 0, \ldots, N_c - 1$ indexes chirps (slow time). This 2-D data structure forms the input to range–Doppler processing, where a 2-D FFT over n and m yields joint range–velocity estimates.

Data Structure and Preprocessing. After downconversion and sampling, the radar produces a single frame and we stack N_f frames to get complex-valued data tensor $\boldsymbol{S} \in \mathbb{C}^{N_f \times N_a \times N_c \times N_s}$, where N_a the number of RF channels, N_c the number of chirps per frame (slow time), and N_s the number of fast-time samples per chirp. Each element $\boldsymbol{S}_{i,a,m,n}$ corresponds to a discrete-time sample of the IF signal in (2) for frame i, RF channel a, chirp m, and fast-time index n. Range information is extracted by applying an N_s-point FFT along the fast-time dimension n for each (i,a,m): $\tilde{\boldsymbol{S}}_{i,a,m} = \frac{\text{FFT}(\boldsymbol{S}_{i,a,m,:})}{N_s}$, which maps fast time to range bins with resolution $\Delta R = c/(2B)$. To suppress static clutter, Moving Target Indication (MTI) filtering is performed by subtracting the previous frame i for each antenna a, chirp m, and range bin k, $\boldsymbol{Z}_{i,a,m,k} = \tilde{\boldsymbol{S}}_{i,a,m,k} - \tilde{\boldsymbol{S}}_{i-1,a,m,k}$.

Each $\boldsymbol{Z}_{i,a} \in \mathbb{C}^{N_c \times N_s}$ is divided into non-overlapping patches $\boldsymbol{P}_{i,a,p,q} \in \mathbb{C}^{P_c \times P_s}$, flattened into $\boldsymbol{p}_{i,a,p,q} \in \mathbb{C}^{P_c P_s}$, where P_c is the patch size along slow time, and P_s is the patch size along fast time. Then, stacked across antennas to form tokens

$$\boldsymbol{t}_{i,p,q} = [\text{Re}(\boldsymbol{p}_{i,1,p,q}), \text{Im}(\boldsymbol{p}_{i,1,p,q}), \ldots, \text{Re}(\boldsymbol{p}_{i,N_a,p,q}), \text{Im}(\boldsymbol{p}_{i,N_a,p,q})]^\top \in \mathbb{R}^{2N_a P_c P_s}. \tag{3}$$

We denote the number of tokens for sample i as L_i. To accommodate heterogeneity across radar configurations—where the number of patches per frame may differ—tokens are zero-padded within a batch to ensure consistent sequence lengths. The resulting token sequence is represented as $\boldsymbol{x}_i \in \mathbb{R}^{L_i \times d}$, where the token dimension is given by $d = 2N_a P_c P_s$, with N_a denoting the number of RF channels, P_c the number of chirps per patch, and P_s the number of samples per patch.

Energy-Weighted Two-Level Masking. Given the tokenized radar frame $\boldsymbol{x}_i = [\boldsymbol{t}_{i,1}, \ldots, \boldsymbol{t}_{i,L_i}]^\top \in \mathbb{R}^{L_i \times d}$, where each token $\boldsymbol{t}_{i,j} \in \mathbb{R}^d$ encodes spatio-temporal information from a patch, we apply a two-stage masking strategy. First, we compute the energy of each token as $E_{i,j} = \|\boldsymbol{t}_{i,j}\|_2^2$. Tokens with higher energy typically correspond to patches containing stronger target returns, while low-energy tokens often represent noise or clutter. To emphasize informative regions, the probability of selecting token j for masking is weighted by its normalized energy:

$$P_{\text{sel}}(i,j) = \frac{E_{i,j}}{\frac{1}{L_i}\sum_{k=1}^{L_i} E_{i,k}}\, p_{\text{patch}}, \tag{4}$$

where p_{patch} is the global patch selection rate. This energy-weighted sampling ensures that the model learns to reconstruct the most informative parts of the scene rather than focusing on uninformative

background. In the second stage, for each selected token, individual elements are masked independently with probability p_{mask}. A masked element is assigned a value of zero (representing missing information in the input) $\boldsymbol{m} \in \mathbb{R}^d$ in the model's internal representation:

$$\tilde{t}_{i,j,u} = \begin{cases} m_u, & \text{with prob. } p_{\text{mask}}, \\ t_{i,j,u}, & \text{otherwise}, \end{cases} \tag{5}$$

where $u \in \{1, \ldots, d\}$ indexes token dimensions. This produces the masked input $\tilde{\boldsymbol{x}}_i \in \mathbb{R}^{L_i \times d}$, which is fed into the transformer encoder. The encoder is thus trained to infer high-energy, informative features from partial observations, improving its ability to model key radar returns while ignoring redundant or low-energy background.

Model Architecture. A classification token (CLS) is prepended to the token sequence $\tilde{\boldsymbol{x}}_i \in \mathbb{R}^{L_i \times d}$. This CLS token is a deterministic vector of length d with all entries initialized to 0.5, serving as a global representation of the input frame. Deterministic positional embeddings $\boldsymbol{PE} \in \mathbb{R}^{(L_i+1) \times d}$ are then added to encode spatial–temporal ordering. The resulting model input is $\boldsymbol{x}_i^{\text{in}} = [\boldsymbol{t}_{\text{CLS}}, \tilde{\boldsymbol{x}}_i] + \boldsymbol{PE}$.

The model consists of two main components. The first is the foundation encoder, composed of stacked transformer encoder layers, which processes $\boldsymbol{x}_i^{\text{in}}$ to produce contextualized representations $\boldsymbol{H}_i \in \mathbb{R}^{(L_i+1) \times d_{\text{model}}}$. The first row of $\boldsymbol{H}_i$ corresponds to the CLS token embedding, and the remaining L_i rows correspond to the encoded patch tokens.

The second component consists of three decoder heads, each implemented as a linear projection and applied only to the non-CLS tokens. The reconstruction head predicts the masked token values: $\hat{\boldsymbol{x}}_i^{\text{recon}} = \text{MLP}_{\text{recon}}(\boldsymbol{H}_i^{\text{patch}})$, where $\boldsymbol{H}_i^{\text{patch}} \in \mathbb{R}^{L_i \times d_{\text{model}}}$ contains only patch token embeddings. The temporal prediction heads output the previous and next frame patches, respectively: $\hat{\boldsymbol{x}}_i^{\text{prev}} = \text{MLP}_{\text{prev}}(\boldsymbol{H}_i^{\text{patch}})$ and $\hat{\boldsymbol{x}}_i^{\text{next}} = \text{MLP}_{\text{next}}(\boldsymbol{H}_i^{\text{patch}})$. The encoder thus learns a shared, task-agnostic radar representation, while the decoder heads map these features to specific reconstruction and temporal prediction objectives.

Reconstruction loss (energy-weighted masked MSE). Let $\boldsymbol{x}_i = [\boldsymbol{x}_{i,1}, \ldots, \boldsymbol{x}_{i,L_i}]^\top \in \mathbb{R}^{L_i \times d}$ be the zero-padded ground-truth token sequence for frame i, and $\tilde{\boldsymbol{x}}_i$ the masked version where selected elements are set to zero. The set $\mathcal{M}$ contains indices of masked tokens, with energy $E_{i,j} = \|\boldsymbol{x}_{i,j}\|_2^2$ and per-frame mean $\bar{E}_i$. The reconstruction head outputs $\hat{\boldsymbol{x}}_i^{\text{recon}}$, and the loss is

$$\mathcal{L}_{\text{recon}} = \frac{1}{|\mathcal{M}|} \sum_{(i,j) \in \mathcal{M}} \frac{E_{i,j}}{\bar{E}_i} \left\| \hat{\boldsymbol{x}}_{i,j}^{\text{recon}} - \boldsymbol{x}_{i,j} \right\|_2^2. \tag{6}$$

Energy weighting emphasizes reconstruction of high-information patches while down-weighting low-energy (often noise-dominated) regions.

Contrastive loss (InfoNCE on CLS with targeted augmentations). For each frame i, we first form a primary masked token view $\tilde{\boldsymbol{x}}_i^{(1)}$ from the MTI-filtered data. To construct the secondary view, we apply an augmentation operator $\mathcal{A}$ directly on the underlying complex-valued range-FFT data $\tilde{\boldsymbol{s}}_i$ (after MTI filtering) to produce $\tilde{\boldsymbol{s}}_i^{(2)} = \mathcal{A}(\tilde{\boldsymbol{s}}_i)$. The augmentation $\mathcal{A}$ consists of: (i) additive complex Gaussian noise, $\tilde{\boldsymbol{s}} \leftarrow \tilde{\boldsymbol{s}} + \boldsymbol{\epsilon}$, $\boldsymbol{\epsilon} \sim \mathcal{CN}(\boldsymbol{0}, \sigma^2 \mathbf{I})$; (ii) random phase injection, $\tilde{\boldsymbol{s}} \leftarrow \tilde{\boldsymbol{s}} \odot e^{j\phi}$, with $\phi \sim \text{Unif}[-\phi_{\max}, \phi_{\max}]$; and (iii) fast-time circular shift by Δ samples, $\tilde{\boldsymbol{s}}(m, \cdot) \leftarrow \tilde{\boldsymbol{s}}(m + \Delta, \cdot)$, where m indexes the slow-time axis. After augmentation, $\tilde{\boldsymbol{s}}_i^{(2)}$ is patch-tokenized in the same way as the primary view to produce $\tilde{\boldsymbol{x}}_i^{(2)}$. Both views are passed through the encoder, yielding raw output CLS tokens $\boldsymbol{h}_{i,\text{CLS}}^{(1)}$ and $\boldsymbol{h}_{i,\text{CLS}}^{(2)} \in \mathbb{R}^{d_{\text{model}}}$. After ℓ_2-normalization, the InfoNCE loss over frames $\mathcal{B}_F$ is

$$\mathcal{L}_{\text{contrastive}} = -\frac{1}{|\mathcal{B}_F|} \sum_{i \in \mathcal{B}_F} \log \frac{\exp\Big(\text{sim}(\boldsymbol{h}_{i,\text{CLS}}^{(1)}, \boldsymbol{h}_{i,\text{CLS}}^{(2)})/\tau\Big)}{\sum_{i' \in \mathcal{N} \cup \{i\}} \exp\Big(\text{sim}(\boldsymbol{h}_{i,\text{CLS}}^{(1)}, \boldsymbol{h}_{i',\text{CLS}}^{(2)})/\tau\Big)}, \tag{7}$$

where $\text{sim}(\boldsymbol{u}, \boldsymbol{v}) = \boldsymbol{u}^\top \boldsymbol{v}$ is cosine similarity after normalization, $\tau > 0$ is a temperature parameter, and $\mathcal{N}$ indexes all negatives in the batch. The denominator includes the positive pair, following the standard InfoNCE formulation, while the augmentations ensure invariance to small perturbations.

Temporal prediction loss. The previous and next frame heads output $\hat{\boldsymbol{x}}_i^{\text{prev}} \in \mathbb{R}^{(L_i-1)\times d}$ and $\hat{\boldsymbol{x}}_i^{\text{next}} \in \mathbb{R}^{(L_i-1)\times d}$ for frame i. Let $\Delta \geq 1$ denote the temporal offset (e.g., $\Delta = 1$ for immediate neighbors). Binary indicators $\nu_{i-\Delta}$ and $\nu_{i+\Delta}$ ensure that reference frames exist (avoiding sequence boundaries or padding). The temporal loss is

$$\mathcal{L}_{\text{temporal}} = \frac{1}{|\mathcal{B}_F|} \sum_{i \in \mathcal{B}_F} \left[\nu_{i-\Delta} \frac{\|\hat{\boldsymbol{x}}_i^{\text{prev}} - \boldsymbol{x}_{i-\Delta}\|_F^2}{L_i d} + \nu_{i+\Delta} \frac{\|\hat{\boldsymbol{x}}_i^{\text{next}} - \boldsymbol{x}_{i+\Delta}\|_F^2}{L_i d} \right], \tag{8}$$

where $\mathcal{B}_F$ is the set of frames in the batch and $\|\cdot\|_F$ denotes the Frobenius norm. This loss encourages frame embeddings to be temporally predictive of spatially aligned patches at offset Δ, thereby capturing motion dynamics. The total loss is a weighted sum of three components: $\mathcal{L}_{\text{total}} = \lambda_1 \mathcal{L}_{\text{recon}} + \lambda_2 \mathcal{L}_{\text{contrastive}} + \lambda_3 \mathcal{L}_{\text{temporal}}$.

3 Experiments

Experiments are conducted on a large-scale radar dataset acquired with mmWave sensors across diverse scenarios, totaling ≈750,000 frames. Standard preprocessing includes DC removal, range FFT, MTI filtering, and patch-based tokenization. The foundation model is pre-trained on the full dataset and evaluated on task-specific separate datasets with train/validation/test splits. The encoder is a 4-layer transformer with $d_{\text{model}} = 256$, hidden dimension 512, patch sizes $P_c = P_s = 4$, and token dimension $d = 2N_a P_c P_s = 96$ ($N_a = 3$). Training uses AdamW with learning rate 5×10^{-5}, batch size 64, and 150 epochs. Loss weights are $\lambda_1 = 0.4$, $\lambda_2 = 0.4$, $\lambda_3 = 0.2$; masking rates are $p_{\text{patch}} = 0.35$ and $p_{\text{mask}} = 0.75$.

Downstream Task: Presence Detection. Presence detection is formulated as binary classification: $y = 0$ (absence) vs. $y = 1$ (presence of at least one person). The model must differentiate subtle temporal signatures (e.g., micro-movements) from static clutter. Labels are frame-level, requiring instantaneous inference from single-frame radar data. The pre-trained encoder is frozen, and a lightweight classifier is attached. The encoder's CLS token $\mathbf{H}^{\text{CLS}} \in \mathbb{R}^{d_{\text{model}}}$ is used as the input to a four-layer MLP with ReLU nonlinearities. The MLP outputs class logits, which are transformed into probabilities via a softmax layer. The model is trained using the cross-entropy loss:

$$\mathcal{L}_{CE} = -\tfrac{1}{N} \sum_{i=1}^{N} \Big[y_i \log \hat{y}_{i,1} + (1 - y_i) \log \hat{y}_{i,0} \Big]. \tag{9}$$

Baselines and Benchmarks. We compare FMWS (≈1M parameters) against three baselines. The first is a task-specific transformer (≈0.4M parameters) trained from scratch solely on the presence detection task, serving as a lower-bound reference without pre-training. The second is the Large Wireless Model (LWM) [20], a pre-trained wireless foundation model (≈0.7M parameters) that we fine-tune on our radar dataset. The third is DINOv2-Small [22] (≈21M parameters), a vision transformer originally trained on large-scale image corpora, which we adapt to radar data by mapping measurements into image-like representations and fine-tuning accordingly.

For the task-specific transformer, raw patches are directly used without pretrained embeddings; tokens are constructed in the same way as FMWS but without a CLS token. LWM follows a similar pipeline, except RF channels are encoded as independent tokens rather than stacked within the same patch, reducing the per-token dimension from $2N_a P_c P_s$ to $2P_c P_s$ and thus increasing the number of input tokens relative to FMWS. DINOv2 accepts RGB images; we treat RF channels as pseudo-RGB channels, normalize them, and feed the resulting tensor of shape $3 \times N_c \times N_s$. Patching and flattening are handled internally by DINOv2. For all three models (FMWS, LWM, DINOv2), presence detection is performed using the CLS token passed through a four-layer MLP described earlier. In Fig. 1, the blue curve shows that raw patches can achieve more than 95% accuracy under supervised learning when sufficient labels are available. However, such large training datasets (≈2M labeled frames) are rarely accessible. We therefore evaluate our models in the low-data regime, comparing raw inputs against embeddings from frozen and fine-tuned pretrained models. We observe that fine-tuning consistently improves performance across all pretrained models. Without fine-tuning, DINOv2 performs best, likely due to its large training corpus and substantially larger backbone. In contrast, LWM performs poorly even after fine-tuning, providing an important insight: models

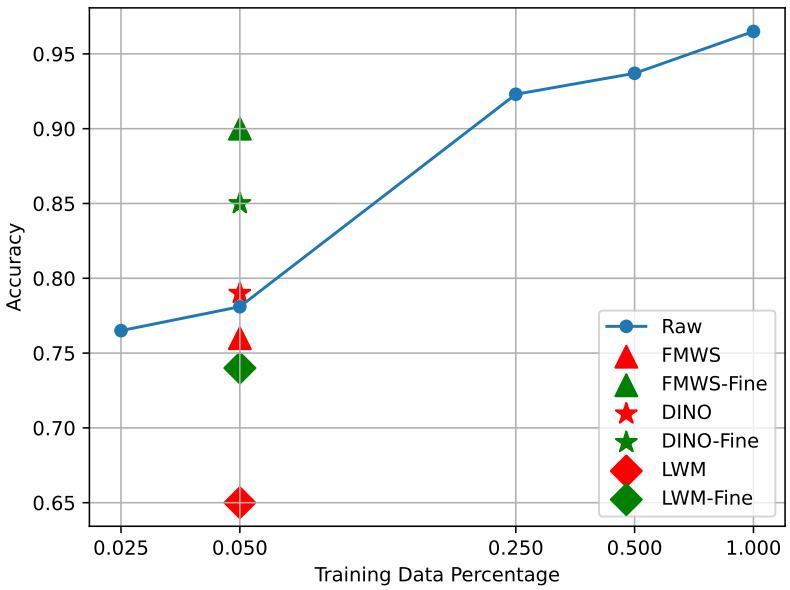

Figure 1: Comparison of test set accuracy

trained on different modalities may not effectively capture radar-specific features without proper input transformations. In our case, radar data were mapped to RGB-like images, enabling DINOv2 to leverage spectrogram-like patterns it may have encountered during training, whereas LWM was trained only on channel frequency responses. Our model shows limited gains without fine-tuning but, once fine-tuned, achieves accuracy comparable to that obtained with four times more labeled data. These results highlight the importance of pretrained models for downstream tasks where labeled data are scarce.

4 Conclusion

This paper introduced **FMWS**, a foundation model for mmWave radar sensing that unifies reconstruction, contrastive, and temporal predictive objectives within a transformer-based architecture. By leveraging pre-training on over 0.75 million radar frames, FMWS learns semantically rich and transferable representations that generalize across tasks and environments. Our formulation explicitly integrates energy-weighted reconstruction, cross-view contrastive learning using CLS tokens, and temporal consistency across variable frame shifts, enabling robust feature extraction without handcrafted priors. Comprehensive experiments on the presence detection task demonstrate that FMWS significantly outperforms task-specific models, handcrafted baselines, and cross-domain transfer methods such as LWM and DINOv2. The learned embeddings yield strong quantitative performance (e.g., 91% accuracy on presence detection). Looking ahead, we plan to scale FMWS to larger models and datasets to further improve embedding quality, and to evaluate its utility across a broader range of downstream tasks beyond presence detection.

References

[1] M. I. Skolnik. *Introduction to Radar Systems*. McGraw-Hill, 1970.

[2] T. E. McEwan. Ultra-wideband radar motion sensor. U.S. Patent 5,573,012, 1996.

[3] R. J. Fontana. Recent system applications of short-pulse ultra-wideband (UWB) technology. *IEEE Transactions on Microwave Theory and Techniques*, 52(9):2087–2104, 2004.

[4] Fadel Adib and Dina Katabi. See through walls with wifi! In *Proceedings of the ACM SIGCOMM 2013 Conference on SIGCOMM*, SIGCOMM '13, page 75–86, New York, NY, USA, 2013. Association for Computing Machinery. ISBN 9781450320566. doi: 10.1145/2486001.2486039. URL https://doi.org/10.1145/2486001.2486039.

[5] Jaime Lien, Nicholas Gillian, M. Emre Karagozler, Patrick Amihood, Carsten Schwesig, Erik Olson, Hakim Raja, and Ivan Poupyrev. Soli: ubiquitous gesture sensing with millimeter wave radar. *ACM Trans. Graph.*, 35(4), July 2016. ISSN 0730-0301. doi: 10.1145/2897824.2925953. URL https://doi.org/10.1145/2897824.2925953.

[6] J. Hasch, E. Topak, R. Schnabel, T. Zwick, R. Weigel, and C. Waldschmidt. Millimeter-wave technology for automotive radar sensors in the 77 GHz frequency band. *IEEE Transactions on Microwave Theory and Techniques*, 60(3):845–860, 2012.

[7] Y. Liu, Z. Chen, Y. Zhang, and M. Zhao. Fall detection using FMCW radar and deep learning. *IEEE Access*, 7:110749–110758, 2019.

[8] L. Tang, Y. Zhang, and P. Li. Human pose estimation using mmWave radar with point cloud representation. *IEEE Sensors Journal*, 21(5):6356–6367, 2021.

[9] Yongsen Ma, Gang Zhou, and Shuangquan Wang. Wifi sensing with channel state information: A survey. *ACM Comput. Surv.*, 52(3), June 2019. ISSN 0360-0300. doi: 10.1145/3310194. URL https://doi.org/10.1145/3310194.

[10] Wei Wang, Alex X. Liu, and Muhammad Shahzad. Gait recognition using wifi signals. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, UbiComp '16, page 363–373, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450344616. doi: 10.1145/2971648.2971670. URL https://doi.org/10.1145/2971648.2971670.

[11] Mingmin Zhao, Tianhong Li, Mohammad Abu Alsheikh, Yonglong Tian, Hang Zhao, Antonio Torralba, and Dina Katabi. Through-wall human pose estimation using radio signals. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7356–7365, June 2018. doi: 10.1109/CVPR.2018.00768.

[12] Mingmin Zhao, Yonglong Tian, Hang Zhao, Mohammad Abu Alsheikh, Tianhong Li, Rumen Hristov, Zachary Kabelac, Dina Katabi, and Antonio Torralba. Rf-based 3d skeletons. SIGCOMM '18, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450355674. doi: 10.1145/3230543.3230579. URL https://doi.org/10.1145/3230543.3230579.

[13] Tom Brown, Benjamin Mann, and Nick Ryder. Language Models are Few-Shot Learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

[14] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

[15] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

[16] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jegou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9630–9640, 2021. doi: 10.1109/ICCV48922.2021.00951.

[17] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners, 2021. URL https://arxiv.org/abs/2111.06377.

[18] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019. URL https://arxiv.org/abs/1810.04805.

[19] Guangjin Pan, Kaixuan Huang, Hui Chen, Shunqing Zhang, Christian Häger, and Henk Wymeersch. Large wireless localization model (lwlm): A foundation model for positioning in 6g networks. *arXiv preprint arXiv:2505.10134*, 2025.

[20] Sadjad Alikhani, Gouranga Charan, and Ahmed Alkhateeb. Large wireless model (lwm): A foundation model for wireless channels, 2025. URL https://arxiv.org/abs/2411.08872.

[21] Berkay Guler, Giovanni Geraci, and Hamid Jafarkhani. A Multi-Task Foundation Model for Wireless Channel Representation Using Contrastive and Masked Autoencoder Learning, 2025. URL https://arxiv.org/abs/2505.09160.

[22] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2023.