DeepSenseMoE: Harnessing Power of Time Series Foundation Models for Few-Shot Human Activity Recognition

Zenan Fu[1*], Dongzhou Cheng[2*], Lei Zhang[1†], Wenbo Huang[3 4†], Zhenghao Chen[5†], Hao Wu[6]

[1]Nanjing Normal University, Nanjing 210023, Jiangsu, China
[2]Shanghai Innovation Institute, Shanghai 200231, China
[3]Southeast University, Nanjing 211189, Jiangsu, China
[4]Institute of Science Tokyo, Tokyo 152-8550, Japan
[5]The University of Newcastle, Callaghan, NSW 2308, Australia
[6]Yunnan University, Kunming 650500, China
{231812011, leizhang}@njnu.edu.cn, 240108390137@sii.edu.cn, wenbohuang1002@outlook.com, zhenghao.chen@newcastle.edu.au, haowu@ynu.edu.cn

Abstract

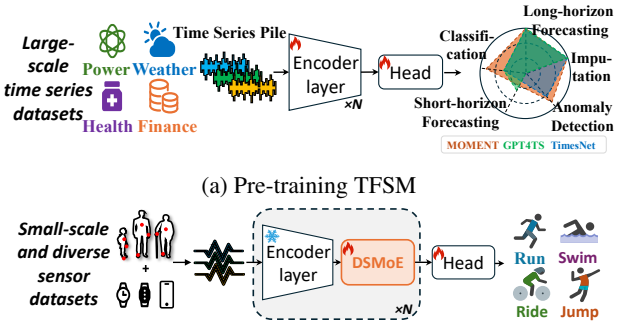
Recent advances in Time Series Foundation Models (TSFMs) have fundamentally revolutionized general time series analysis across domains like finance, retail, weather, and power. However, how to unlock the hidden capacity of general-purpose TSFMs for wearable activity recognition still remains largely unexplored, given severe sensor annotation scarcity and highly heterogeneous sensor data. To address these challenges, we propose DeepSenseMoE—a novel multi-scale convolution-based Mixture of Experts (MoE) module for parameter-efficient fine-tuning of general-purpose TSFMs to sensor-based activity recognition. DeepSenseMoE integrates three key innovations: (1) Multi-scale convolutional experts with different filter sizes responsible for capturing varying sensor contexts; (2) Shared-expert isolation mechanism compressing common activity knowledge into a single shared expert while reducing redundancy among routed experts; and (3) Hierarchical supervised contrastive alignment guiding experts to further learn discriminative activity features. Extensive experiments on three challenging HAR benchmarks demonstrate DeepSenseMoE's superiority, achieving up to 9.5% accuracy gains over state-of-the-art under few-shot and full-supervised settings, with only $<1\%$ additional trainable parameters. We hope that this work may lay a solid foundation to accelerate development and deployment of powerful TSFMs in data-scarce wearable activity recognition tasks while reducing the reliance on labeled sensor data.

Code — https://github.com/FuZenan/DeepSenseMoE

1 Introduction

During the past decade, human activity recognition (HAR) has garnered significant attention. It leverages a variety of sophisticated deep learning models including Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Transformers to analyze and infer human activities through wearable sensors attached to different body locations, offering a wide range of real-world applications like health management, fitness tracking, sports performance analysis, and gesture recognition (Chen et al. 2021). In essence, sensor-based activity recognition can be seen as a multivariate time series classification task. Recently, inspired by outstanding success of foundation models (FMs) in modalities like text and image (Awais et al. 2025; Min et al. 2023), the concept of time series foundation models (TSFMs) has emerged as a new research direction, which aims to harness the potential of FM paradigm by leveraging large-scale time series datasets to learn generalizable representations across a multitude of general time series tasks like finance, power, weather, and health (Das et al. 2024; Zhou et al. 2023; Goswami et al. 2024). Despite promising potential of TSFMs, we notice that this evolution seems mostly limited to general time series tasks (e.g., stock prices for different companies or weather data from various locations), and has been notably absent in the field of HAR.

Now, there are still two ongoing challenges that severely hinder the development of foundation model paradigm in wearable activity recognition: 1) **Sensor annotation scarcity**. Though huge amounts of multivariate sensor readings are being produced by wearable devices, labeling sen-



(a) Pre-training TFSM

(b) DeepSenseMoE Fine-tuning for TFSM

Figure 1: Unlocking hidden capacity of general-purpose TSFMs for wearable activity recognition.

*These authors contributed equally.
†Corresponding authors.

sor data is harder than labeling image or text, which requires an annotator to meticulously annotate sensor readings. While it is easy to distinguish a picture of a dog from a cat instantly, telling subtle differences between IMU sensor signals for upstairs and downstairs may be challenging. This is a tedious and time-consuming process, causing severe sensor annotation scarcity. Thus, most publicly available HAR datasets remain small in scale; 2) **High heterogeneity in sensor data**. Wearable sensor data is highly heterogeneous. Given multiple sensor devices and varying on-body recording locations, their resulting distributions may be astonishingly diverse (Chen et al. 2021). Even the same actions recorded at identical sensors and locations but with different persons also produce vastly different sensor readings, due to respective behavior patterns. Although numerous small-scale HAR datasets have been released, they are mostly gathered under different sensor settings and collection protocols including various sensor modalities (e.g., accelerometer, gyroscope, and magnetometer), body locations (e.g., wrist, arm, leg), subject-variability (e.g., men, women, elderly, children), which remain highly diverse or heterogeneous in dataset structure (i.e., distribution).

Given that the two major challenges persist, the current HAR research still remains highly fragmented in the absence of common standardization across small-scale datasets. On one hand, the lack of massive and unified datasets makes it impractical to train activity foundation models from scratch. On the other hand, most existing TSFMs are often pre-trained for general-purpose time series tasks, not customized for activity-specific domain (Das et al. 2024). This creates a crucial gap: *Can we harness the power of existing TSFMs pre-trained from general data-rich time series datasets tailored to wearable sensor stream for data-scarce activity recognition tasks?* To close the gap, in this paper, we make the first attempt to leverage existing TSFMs pre-trained from data-rich general time series datasets, and adapt them to activity-specific tasks with scarce and heterogeneous sensor data. Figure 1 illustrates the motivation behind our work. However, the majority of previous TSFMs are closed-source, resulting in limited access to the model itself. Therefore, in this paper, we mainly focus on the first family of open-source TSFM namely MOMENT (Goswami et al. 2024), and fine-tune it with activity-related knowledge.

Since wearable sensor data is highly scarce and heterogeneous, direct usage of MOMENT will considerably underperform under data-scarce scenario, due to lacking homogeneity between general time series datasets and activity-related target datasets. To empower general-purpose MOMENT with strong activity sensing capability, a straightforward strategy is to fine-tune it on downstream activity-specific tasks. Despite superior performance in computer vision (CV) and natural language processing (NLP), conventional full fine-tuning strategy often involves updating all model weights with high training cost (Awais et al. 2025; Min et al. 2023), which makes it unsuitable for wearable activity recognition under resource-constrained scenario. An alternative solution is to incorporate activity-related domain knowledge into MOMENT by Parameter-Efficient Fine-Tuning (PEFT) adapters. However, existing adapters (Hu et al. 2022; Houlsby et al. 2019) are mostly designed for text and image having unified format, not purposely designed for heterogeneous sensor signals. Consequently, they often simply compress the upstream features with linear projection, where the fixed layer parameters cannot be well fine-tuned to fully match the varying distribution of diverse activity recognition tasks with scarce publicly available sensor data.

To resolve this problem, we introduce DeepSense-MoE, a new multi-scale convolution-based Mixture of Experts (MoE) architecture, which serves as a parameter-efficient fine-tuning module to adapt powerful general-purpose TSFMs to downstream activity recognition tasks with varying sensor signals. DeepSenseMoE incorporates three principal strategies: 1) **Multi-Scale Convolutional Experts**. Prior works claim that CNNs inherently possess different cognition of sensor features at different filter scales (Chen et al. 2021). Thus, we first explore the use of CNN as experts, and leverage the transferring capability of multi-scale convolutional filters referred to as convolution-based MoE to capture varying sensor contexts for heterogeneous sensor data modeling; 2) **Shared Expert Isolation**. Intuitively, heterogeneous sensor knowledge should be learned respectively by different experts, so that each expert maintains a certain level of specialization. Given high variability in sensor input space, conventional routing strategy may make multiple experts converge in acquiring too similar knowledge in their respective parameters, weakening expert specialization while increasing parameter redundancy among experts. Inspired by recent success of DeepSeek-MoE in language modeling (Dai et al. 2024), we address this by isolating a specific expert designated as shared expert that remains always activated, in charge of capturing and consolidating common knowledge across varying sensor contexts. By compressing common sensor knowledge into this shared expert, the parameter redundancy can be effectively mitigated among routed experts; 3) **Hierarchical Supervised Contrastive Alignment**. By introducing layer-wise contrastive learning between shared experts and routed experts, we treat expert outputs from the same activity label as positive pairs, and those from different activity labels as negative pairs, which forces all experts to learn increasingly discriminative features, further enhancing expert specialization. Our contributions are summarized as follows:

- **New perspective**: Given that wearable sensor data is scarce and heterogeneous, how to harness the power of TSFMs pre-trained from data-rich general time series datasets and tailor them to sensor stream remains largely unexplored. To the best of our knowledge, **this paper is the first to unlock the hidden capacity of general-purpose TSFMs for wearable activity recognition**.
- **Architectural innovation**: To tackle high heterogeneity in wearable sensor data, we introduce DeepSense-MoE: a parameter-efficient routed MoE architecture, which features three core innovations: multi-scale convolutional experts, shared expert isolation mechanism, and hierarchical supervised contrastive alignment, effectively bridging distribution gap between general pre-training time-series data and activity-specific sensor streams.

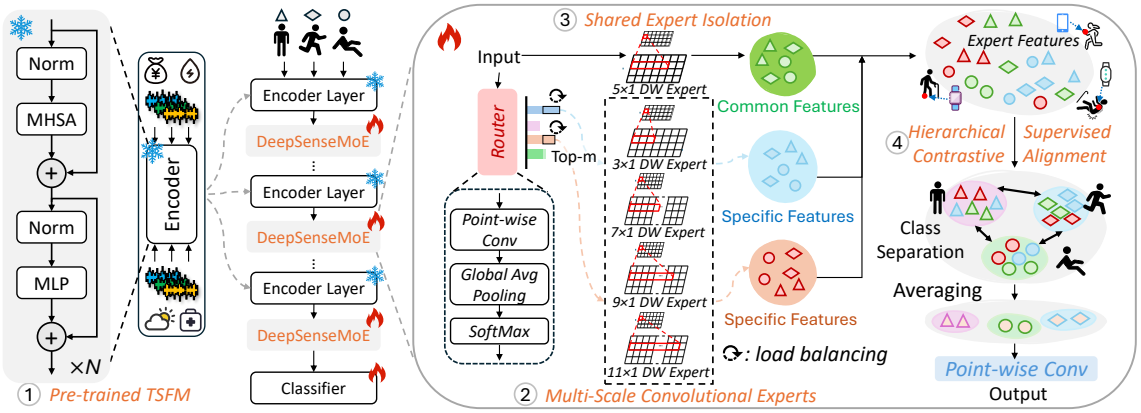


Figure 2: An overview of our proposed DeepSenseMoE: (1) Pre-trained MOMENT backbone, (2) Multi-scale convolutional experts (MSCE), (3) Shared expert isolation (SEI), and (4) Hierarchical supervised contrastive alignment (HSCA).

- **Superior performance**: Extensive experiments on three challenging HAR benchmarks show that DeepSenseMoE achieves substantial accuracy gains of up to 9.5% over state-of-the-art under few-shot and full-supervised settings, with only $<1\%$ additional trainable parameters. We provide empirical analyses for every specific design choice. This work lays a solid foundation to accelerate development and deployment of TSFMs in activity recognition while mitigating sensor label reliance.

2 Related Works

2.1 Human Activity Recognition

A variety of machine learning algorithms always play a dominant role in wearable activity recognition. Earlier works often relied on traditional machine learning algorithms such as SVM, KNN, and Random Forests (Bulling, Blanke, and Schiele 2014), requiring manual feature engineering. Deep learning models ranging from CNNs (Yang et al. 2015) to hybrid CNN-RNN architectures (Ordóñez and Roggen 2016) and attention-based Transformer (Gao et al. 2023) have significantly prompted rapid development of sensor-based activity recognition, which enable an end-to-end learning to automate sensor feature representation learning. However, they remain fundamentally limited due to heavy reliance on supervised learning and annotated sensor datasets. Due to sensor annotation scarcity, the paradigm of self-supervised representation learning has increasingly gained attention in the context of HAR (Logacjov 2024). Despite promising progress, these models are typically trained on individualized small-scale and diverse datasets. The HAR research still remains highly fragmented, where findings from one dataset may fail to generalize to others.

2.2 Time Series Foundation Models

Inspired by outstanding success of foundation model paradigms in CV and NLP, these techniques have recently been extended to time series data, referred to as time series foundation models (TSFMs). For instance, GPT4TS (Zhou et al. 2023) has introduced a pre-trained Transformer backbone from another modality (e.g, BERT) for the time series forecasting. Based on multi-periodicity characteristic, TimesNet (Wu et al. 2023) handled intricate temporal variations through converting 1D time series data into 2D space for general time series analysis. TimesFM (Das et al. 2024) has proposed a decoder-only attention model purposely optimized for time series forecasting tasks. MOIRAI (Woo et al. 2024) has designed a masked encoder-based Transformer architecture for universal time series forecasting. In particular, MOMENT has released the first family of open-source large-scale TSFM, which is pre-trained via a self-supervised masked prediction objective from scratch, and may serve as a foundation framework for diverse time series analysis tasks (Goswami et al. 2024). Recent advances in general-purpose TSFMs have paved a new way to look at wearable activity recognition problem. However, until now, there is no existing literature on fine-tuning time series-based foundation models for wearable activity recognition.

3 Methodology

3.1 Preliminary

To ensure effective adaptation for wearable activity recognition, it is crucial to identify one high-performance TSFM that suits this task. Prior work (Goswami et al. 2024) has conducted a systematic comparison of three state-of-the-art TSFMs: MOMENT (Goswami et al. 2024), GPT4TS (Zhou et al. 2023), and TimesNet (Das et al. 2024) on five general time series tasks. Among them, MOMENT consistently performs the best, especially in time series classification. On this basis, we adopt the open-source MOMENT-small model as our backbone. It is an encoder-only Transformer architecture pre-trained via a self-supervised masked prediction objective on a large corpus of time series data spanning diverse domains like weather, healthcare, power, and finance, which holds the potential of generalizing to downstream activity recognition tasks. Notably, the original MOMENT pre-training time series datasets fundamentally exclude these activity-centric datasets used in our study, ensuring fair and unbiased evaluation.

We employ an adapter-tuning paradigm with only a small

number of activity-specific parameters introduced into the pre-trained MOMENT backbone (Houlsby et al. 2019). Given a sensor dataset $D = \{(x_i, y_i)\}_{i=1}^{S}$, the optimization objective under standard full fine-tuning may be written as:

$$\theta^* = \arg\min_{\theta} \mathcal{L}(D; \theta), \tag{1}$$

where θ denotes the entire model parameters, $\mathcal{L}$ is the training loss. In contrast, while keeping the N-layer MOMENT encoder θ_F frozen, PEFT only optimizes the parameters of N-layer adapter ω, thus significantly reducing training cost:

$$\omega^* = \arg\min_{\omega} \mathcal{L}(\mathcal{D}; \theta_F, \omega). \tag{2}$$

3.2 Method Overview

Figure 2 presents an overview of our DeepSenseMoE with three core strategies: multi-scale convolutional experts (MSCE), shared expert isolation (SEI), and hierarchical supervised contrastive alignment (HSCA).

Multi-Scale Convolutional Experts. While adapter-based PEFT methods have proven effective in CV and NLP (Houlsby et al. 2019), their heavy reliance on linear projection layers renders suboptimal performance due to highly heterogeneous time series characteristic in wearable sensor signals. Consequently, the frozen MOMENT backbone pretrained on generic time series datasets may produce feature biases and distributional misalignment, degrading downstream performance. To remedy this, we directly embed a sparse MoE architecture with convolutional experts within each adapter: Multi-Scale Convolutional Experts (MSCE). MSCE comprises k parallel depth-wise convolutional branches, each with kernel size $(1, K_i)$ for $K_i \in \{3, 5, \ldots, 2k+1\}$, followed by point-wise aggregation. Let $x \in \mathbb{R}^{B \times C_{\text{in}}^D \times H \times W}$ be a sensor input feature map. We can first calculate its multi-scale output responses:

$$f_{\text{dw}}(x) \;=\; \frac{1}{k} \sum_{i=1}^{k} \big(\omega_{\text{dw}}^{i} \,\hat{\otimes}\, x\big), \tag{3}$$

where $\omega_{\text{dw}}^{i} \in \mathbb{R}^{C_{\text{in}}^D \times 1 \times K_i \times 1}$ and $\hat{\otimes}$ denotes depth-wise convolution. Then these multi-scale outputs can be fused via a residual point-wise convolution operation:

$$f_{\text{ms}}(x) \;=\; x \;+\; \omega_{\text{pw}} \,\overline{\otimes}\, f_{\text{dw}}(x), \tag{4}$$

where $\omega_{\text{pw}} \in \mathbb{R}^{C_{\text{in}}^P \times 1 \times 1 \times C_{\text{out}}^P}$ and $\overline{\otimes}$ denotes point-wise convolution. This design can significantly enhance expressive diversity and model robustness under distribution shift.

MoE with Shared Expert Isolation. Building on MoE's dynamic routing paradigm (Dai et al. 2024), we treat each depth-wise convolution branch as a distinct 'expert' and introduce a lightweight convolutional router for input-dependent routing selection. Given $x \in \mathbb{R}^{B \times C \times H \times W}$, the router applies a k-channel convolution followed by global average pooling and softmax to produce expert scores $\boldsymbol{\alpha} = [\alpha_1, \ldots, \alpha_k]$. To enforce sparsity, we select the top-m experts per sample. If $\mathcal{S}_b$ denotes the chosen expert indices for sensor sample b, the dynamic output can be formulated as:

$$y_{\text{dyn}}^{(b)} \;=\; \frac{1}{m} \sum_{i \in \mathcal{S}_b} \alpha_i^{(b)} \, E_i\big(x^{(b)}\big), \tag{5}$$

where $E_i(x) = \omega_{\text{dw}}^{i} \,\hat{\otimes}\, x$. The top-$m$ gating strategy ensures that each sensor input is adaptively routed through a small subset of specialized convolutional experts. By such sparse expert activation, our convolution-based MoE architecture can dynamically adapt its routing path to better match the target domain, which implicitly aligns feature distributions between original pre-training time series datasets and diverse activity-related sensor datasets, thereby improving model generalization without modifying backbone structure. To further improve parameter efficiency and reduce redundancy among experts, we complement these k routed experts with a *shared depth-wise convolution expert* $E_{\text{sh}}(x)$ (fixed kernel size). The final per-sample output evolves as:

$$y^{(b)} \;=\; \frac{1}{m+1} \, E_{\text{sh}}\big(x^{(b)}\big) \;+\; \frac{1}{m+1} \sum_{i \in \mathcal{S}_b} \alpha_i^{(b)} \, E_i\big(x^{(b)}\big). \tag{6}$$

While these routed experts specialize in varying sensor contexts, the shared expert isolation mechanism is responsible for capturing and compressing common activity knowledge into a single shared expert, which can effectively reduce parameter redundancy among routed experts. Both of them can better balance specialization and generalization.

In fact, our MSCE with shared expert isolation may still suffer from expert collapse: the router always selects a subset of experts, preventing others from enough training. Inspired by Switch Transformers's load-balancing (Fedus, Zoph, and Shazeer 2022), we mitigate this risk by regularizing the router to equalize expert usage. Let $\alpha_i^{(b)}$ denotes the per-sample softmax score, the marginal utilization probability of expert i over a batch of samples B is given as:

$$P_i \;=\; \frac{1}{B} \sum_{b=1}^{B} \alpha_i^{(b)}, \quad i = 1, \ldots, k. \tag{7}$$

We may minimize the *KL divergence* to make P_i come close to the uniform distribution $U_i = 1/k$, i.e.,

$$\mathcal{L}_{\text{bal}} = \sum_{n=1}^{N} \text{KL}\big(U \,\|\, P\big) = \sum_{n=1}^{N} \sum_{i=1}^{k} \frac{1}{k} \log \frac{1/k}{P_i^{(n)}}, \tag{8}$$

which is used to balances input allocation among all routed experts, ensuring full utilization of the MSCE's capacity.

Hierarchical Supervised Contrastive Alignment. While ensuring sufficient training on experts, the load balance loss does not exhibit a preference for any specific expert, making the routing process appear somewhat random. Since each expert receives sensor inputs allocated by such random routing, the sensor content learned by all experts may not differ significantly, contradicting with our original intention of employing MSCE to increase diversity. To mitigate such random routing caused by load balance, we introduce supervised contrastive learning, which encourages both shared and routed experts to learn distinct features by disentangling their feature subspaces with activity labels. For sensor samples x_p of the same activity label drawn from different experts, we regard their representations $\big(z_p^{(i)} =$

$E_i(x_p)$, $z_p^{(i)} = E_i(x_p)\big)$ routed to the expert i as positive pair, avoiding data augmentation to maintain the model streamlined, while all embeddings $E_i(x_r)$ from experts with other labels serve as negative pair. The supervised contrastive loss *without augmentation* is then defined as:

$$\mathcal{L}_{\text{sca}} = -\sum_{i=1}^{m+1}\sum_{p\in\mathcal{S}}\frac{1}{|\mathcal{P}_i|}\sum_{p\in\mathcal{P}_i} \log\frac{\exp\left(z_p^{(i)}\cdot z_p^{(i)}/\tau\right)}{\sum_{i=1}^{m+1}\sum_{r\in\mathcal{N}_{(p)}}\exp\left(z_p^{(i)}\cdot E_i(x_r)/\tau\right)}, \tag{9}$$

where $\mathcal{P}_i$ indexes all the positive pairs for label i from all m selected experts and $\mathcal{N}_{(p)}$ denotes the set of samples from index $p \neq r$. Aggregating this loss over all N layers yields the final hierarchical supervised contrastive loss,

$$\mathcal{L}_{\text{hsca}} = \sum_{n=1}^{N}\mathcal{L}_{\text{sca}}^{(n)}. \tag{10}$$

Objective Function. Finally, our MSCE with shared expert isolation mechanism can be jointly optimized with the primary activity classification loss (i.e., $\mathcal{L}_{\text{cls}}$), load balance loss (i.e., $\mathcal{L}_{\text{bal}}$), hierarchical supervised contrastive loss (i.e., $\mathcal{L}_{\text{hsca}}$), resulting in the overall training objective:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{cls}} + \lambda_1\mathcal{L}_{\text{bal}} + \lambda_2\mathcal{L}_{\text{hsca}}. \tag{11}$$

4 Experiment

4.1 Dataset

To comprehensively evaluate the effectiveness of our method, we conduct experiments on three publicly available and widely-employed HAR benchmarks including HHAR (Stisen et al. 2015), MotionSense (Malekzadeh et al. 2018), and PAMAP2 (Reiss and Stricker 2012), which encompass a set of diverse human activities captured under various sensor modalities, body locations, and sampling rates, thus providing a rigorous testbed for real-world HAR evaluation. To ensure fair comparisons with previous works (Xia et al. 2024a; Ma et al. 2021), we follow the same data preprocessing pipeline. Specifically, raw sensor signals are uniformly segmented into overlapping time-series windows, each containing 500 consecutive sensor readings with a 50% overlap rate between adjacent windows. We randomly assign 80% of the resulting segments as training set and hold out the remaining 20% for testing. During training, the unlabeled training data is reserved for validation. More dataset details can be found in Supplementary Material.

4.2 Comparative Methods

We compare the proposed DeepSenseMoE against a comprehensive set of state-of-the-art HAR approaches, spanning from fully supervised to semi-/unsupervised paradigms. Specifically, the supervised baselines include DCNN (Yang et al. 2015), TCN (Bai, Kolter, and Koltun 2018) and ConformerHAR (Kim et al. 2022). For semi-supervised and unsupervised learning, we evaluate against FixMatch (Sohn et al. 2020), SimCLR (Chen et al. 2020), MDC (Ma et al. 2021) TS2ACT (Xia et al. 2024a) and Vi2ACT (Xia et al. 2024b). All baseline implementations follow their original configurations or standard practices to ensure comparisons.

4.3 Implementation Details

We place our DeepSenseMoE module after each MOMENT encoder layer, while keeping its pre-trained backbone weights frozen. The entire model was then trained for up to 200 epochs using the Adam optimizer with a weight decay of 5×10^{-4}. Following standard contrastive learning practices (Khosla et al. 2020), we set the temperature τ in the contrastive loss to 0.07. The regularization coefficients λ_1 and λ_2 in Eq. 11 are set to 0.2 and 0.5, respectively. Learning rates are chosen from $\{10^{-2}, 10^{-3}\}$ based on validation performance, and the batch size is fixed at 8. To ensure statistical reliability, all results are averaged over three independent runs. Experiments are conducted on a workstation equipped with three NVIDIA GeForce RTX 3090 GPUs.

4.4 Main Results

HAR Benchmarks. We evaluate DeepSenseMoE against eight competitive baselines under four few-shot settings (1, 5, 10, and 20-shot) and fully supervised settings, adhering to the same protocol from Xia et al. (2024a). Here, '1-shot's indicates one labeled sample per class, etc. Among the comparing baselines, Vi2ACT (Xia et al. 2024b) leads as state-of-the-art while TS2ACT (Xia et al. 2024a) ranks second-best in few-shot activity recognition. Table 1 shows DeepSenseMoE consistently outperforms both supervised and semi-/unsupervised baselines across most datasets and settings, surpassing Vi2ACT by significant margins. In challenging 1-shot mode, it attains 75.1% accuracy on HHAR, 83.1% on MotionSense, and 74.4% on PAMAP2, yielding up to 7.0% accuracy gains over Vi2ACT. Notably, the performance gap widens in the 5-shot setting, where DeepSenseMoE attains 82.1% accuracy on HHAR, 91.7% on MotionSense, and 90.5% on PAMAP2, outperforming TS2ACT by up to 9.5%. Though the performance gap tends to narrow beyond the 5-shot setting, DeepSenseMoE still maintains consistent superiority. For instance, it obtains 96.6% accuracy on HHAR, 98.9% on MotionSense, and 98.6% on PAMAP2 under full supervision, refreshing current state-of-the-art with fewer trainable parameters. Unlike Vi2ACT and TS2ACT that rely on laborious cross-modality data augmentation (e.g., activity video generation and manual image search), DeepSenseMoE eliminates the need for such costly procedure. *Against baselines, our method delivers not only unprecedented performance but also data-efficient scalability from extreme low-data to full-data regimes*.

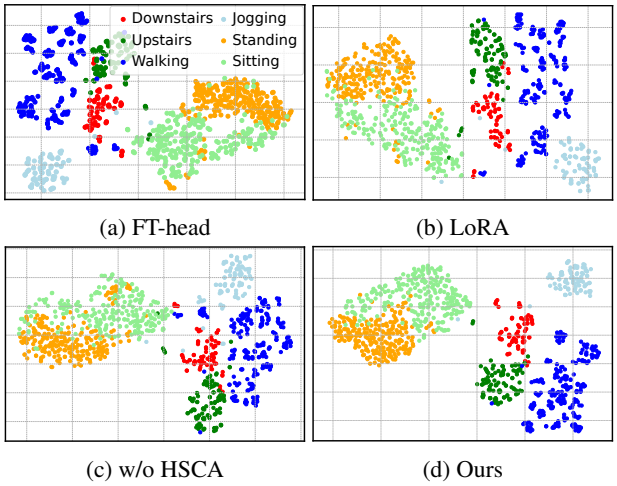
Fine-Tuning Benchmarks. Until now, there has been no similar works that explore fine-tuning pre-trained TSFMs for few-shot activity recognition. We present the first systematic comparisons with various mainstream fine-tuning strategies. Note that the fine-tuning strategies with extra architecture still require fine-tuning the classification head of MOMENT. Compared to traditional fine-tuning and popular PEFT baselines such as Adapter (Houlsby et al. 2019)

Method	Year	HHAR					MotoinSense					PAMAP2				
		1-shot	5-shot	10-shot	20-shot	Full	1-shot	5-shot	10-shot	20-shot	Full	1-shot	5-shot	10-shot	20-shot	Full
DCNN	2015	27.5	34.7	45.7	69.3	90.6	35.5	39.5	62.2	73.3	88.3	32.0	45.3	53.8	64.7	90.5
TCN	2018	32.4	43.2	53.2	71.9	91.5	38.7	47.2	68.9	79.2	90.9	35.8	52.0	61.2	72.1	93.5
ConformerHAR	2022	36.4	49.9	66.8	74.5	94.2	45.5	53.3	73.1	83.3	94.3	36.3	61.2	71.2	77.2	96.4
FixMatch	2020	52.6	61.9	76.6	82.2	91.4	58.2	65.2	78.8	85.4	91.3	56.6	66.5	77.2	85.2	94.2
SimCLR	2020	50.6	54.6	71.8	75.8	91.8	56.3	59.8	76.2	80.9	91.5	46.9	63.1	75.9	80.9	94.5
MDC	2021	68.2	70.6	74.9	75.5	90.3	72.7	72.8	73.2	84.6	87.8	64.1	70.3	80.6	84.6	89.2
TS2ACT	2024	71.1	75.2	82.2	88.7	93.3	74.7	77.3	87.4	92.6	94.4	69.2	78.0	90.1	94.3	97.1
Vi2ACT	2024	**75.5**	78.5	86.5	91.3	95.2	76.1	82.2	90.2	94.1	96.7	73.3	86.0	92.2	95.5	97.9
Ours	2025	75.1 ∇0.4	**82.1** Δ3.6	**87.1** Δ0.6	**93.3** Δ2.0	**96.6** Δ1.4	**83.1** Δ7.0	**91.7** Δ9.5	**94.1** Δ3.9	**97.3** Δ3.2	**98.9** Δ2.2	**74.4** Δ1.1	**90.5** Δ4.5	**95.3** Δ3.1	**97.4** Δ1.9	**98.6** Δ0.7

Table 1: Accuracy (%) comparisons between DeepSenseMoE and state-of-the-art methods. Bold is best and underline is second. Δ indicates accuracy gains over the second-best method.

Method	Param	HHAR					MotionSense					PAMAP2				
		1-shot	5-shot	10-shot	20-shot	Full	1-shot	5-shot	10-shot	20-shot	Full	1-shot	5-shot	10-shot	20-shot	Full
Direct	-	22.2	22.2	22.2	22.2	22.2	2.1	2.1	2.1	2.1	2.1	15.3	15.3	15.3	15.3	15.3
FT-Head	0.037M	67.5	80.0	82.3	91.9	93.5	75.5	84.8	90.1	94.3	98.0	70.9	88.4	93.6	95.5	97.4
FT-Full	35.374M	65.1	73.0	81.3	90.2	91.6	69.5	89.7	90.6	91.6	94.3	59.6	72.4	87.6	94.2	94.2
Adapter	0.238M	63.0	80.3	82.6	92.0	93.0	68.6	82.9	91.9	93.8	98.1	62.4	80.0	93.4	95.7	98.3
LoRA	0.266M	66.2	76.5	81.4	88.8	93.1	65.6	87.1	91.2	94.3	98.1	64.7	86.7	91.0	93.1	97.5
Ours	0.043M	**75.1** Δ7.6	**82.1** Δ1.8	**87.1** Δ4.5	**93.3** Δ1.3	**96.6** Δ3.1	**83.1** Δ7.6	**91.7** Δ2.0	**94.1** Δ2.2	**97.3** Δ3.0	**98.9** Δ0.8	**74.4** Δ3.5	**90.5** Δ2.1	**95.3** Δ1.7	**97.4** Δ1.7	**98.6** Δ0.3

Table 2: Accuracy (%) comparisons between DeepSenseMoE and mainstream fine-tuning strategies. Bold is best and underline is second. Δ indicates accuracy gains over the second-best method. 'Direct' denotes direct usage of pre-trained MOMENT.



(a) FT-head (b) LoRA

(c) w/o HSCA (d) Ours

Figure 3: Visualization of sensor embeddings with t-SNE.

and LoRA (Hu et al. 2022), DeepSenseMoE achieves notably higher accuracy across all three datasets. As shown in Table 2, under 1-shot setting, it outperforms Adapter and LoRA by 14.5% and 17.5% on MotionSense, respectively, despite fewer trainable parameters. Results again highlight the strength of our DeepSenseMoE in enhancing model adaptation under data-scarce scenario. Remarkably, even in full-supervised setting, our method remains competitive, achieving consistent accuracy improvements with minimal overhead, underscoring its efficacy and scalability.

Visualization. We perform t-SNE visualization on the features extracted from the last encoder layer of MOMENT under 5-shot setting on MotionSense dataset. As seen in Figure 3, fine-tuning classification head (i.e., FT-Head) alone fails to form compact clusters, indicating its limited discriminative power. LoRA shows improved structure but still suffers from class entanglement. Our method, even without hierarchical supervised contrastive alignment, can produce more compact and separated clusters. With our hierarchical supervised contrastive alignment, the classification boundaries become even clearer, confirming the synergy between expert specialization and contrastive supervision in enhancing activity class-wise separability in low-data regime.

4.5 Quantitative Analysis

Ablation Study. To deeper understand the contributions of three core components including multi-scale convolutional experts, shared expert isolation, and hierarchical supervised contrastive alignment in DeepSenseMoE, we present their main ablation study results in Table 3, which are divided into four scenarios: In **case 1**, all three core components are disabled. Particularly, we exclude the MoE architecture via bypassing its router to uniformly employ all convolutional expert weights. Without dynamic expert routing, the model completely loses the ability to ensure expert specialization, which performs the worst among all four cases; In **case 2**, we add the router for every DeepSenseMoE layer to activate our MoE architecture, while keeping other two components still disabled. Comparing to case 1, it results in a notable accuracy gain across all few-shot set-

Case	M	S	C	1-shot	5-shot	10-shot	20-shot	Full
1	✗	✗	✗	76.9	87.1	90.2	95.2	97.6
2	✓	✗	✗	80.1	88.7	91.2	96.2	98.0
3	✓	✓	✗	81.3	89.5	92.0	96.6	98.2
4	✓	✓	✓	**83.1**	**91.7**	**94.1**	**97.3**	**98.9**

Table 3: Main ablation study on MotionSense dataset. 'M', 'S', and 'C' denote MSCE, SEI, and HSCA respectively.

Method	1-shot	5-shot	10-shot	20-shot	Full
w/o HSCA	81.3	89.5	92.0	96.6	98.2
DIS	81.4	90.2	92.5	96.8	98.5
INF	81.6	90.1	92.9	97.1	98.2
Ours	**83.1**	**91.7**	**94.1**	**97.3**	**98.9**

Table 4: The effectiveness of different contrastive learning strategies on MotionSense dataset.

tings, highlighting the effectiveness of our sophisticatedly designed MoE architecture in handling heterogeneous sensor data; In **case 3**, while only excluding hierarchical supervised contrastive alignment, we further enable the shared expert isolation mechanism to capture generalizable patterns that regularize routed expert outputs and mitigate redundancy. Despite further performance improvements, it still lags far behind our full-version DeepSenseMoE; In **case 4**, when all three components are activated, our full model consistently achieves the best performance across all few-shot settings, confirming their synergistic benefit.

Effectiveness of Multi-Scale Convolutional Experts. To assess the effectiveness of our multi-scale convolutional experts (MSCE), we conduct an ablation study, where these convolutional experts are replaced with two-layer feed-forward networks referred to as linear experts, while keeping all other settings unchanged. As shown in Figure 4, this modification leads to a dramatic accuracy drop across various few-shot settings. Interestingly, the performance gap tends to increasingly widen, as the number of labeled sensor samples decreases. The findings indicate the clear advantage of multi-scale convolutions over linear mappings, which can improve the expressive richness of individual expert while capturing complex and heterogeneous sensor dynamics, especially under data-scarce scenario.

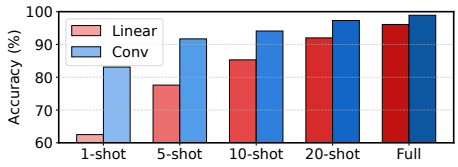


Figure 4: Performance comparison between convolutional and linear experts on Motionsense dataset.

Effectiveness of Hierarchical Supervised Contrastive Alignment. We further investigate alternative contrastive regularization objectives, such as Dispersive (Wang and He 2025) and InfoNCE (He et al. 2020). Though Table 4 shows these variants still provide accuracy improvements over the baseline without any contrastive loss (i.e.,w/o HSCA), they consistently underperform our hierarchical supervised contrastive loss. This underscores the superiority of our method in effectively harnessing limited label information to deliver substantial performance gains under few-shot settings.

Hyperparameter Sensitivity Analysis. Figure 5 shows a sensitivity analysis on the regularization coefficients λ_1 and λ_2 in Eq. 11, which governs both load-balance loss and hierarchical supervised contrastive loss, respectively. By analyzing their impact under the 5-shot setting on MotionSense dataset, we find that the model performance can remain stable across a broad value range, demonstrating the robustness of our method to hyperparameter selection. Notably, while reaching the optimal performance, we can clearly see λ_2 (around 0.5) is greater than λ_1 (around 0.2), indicating that the hierarchical supervised contrastive loss plays a more critical role in guiding expert specialization and enhancing discriminative representation. This is particularly vital in few-shot scenarios, where promoting expert specialization and inter-class separability are essential for generalization.

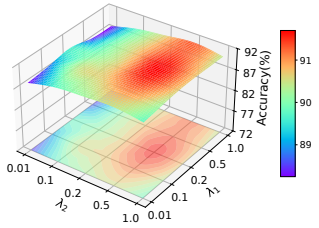


Figure 5: Hyperparameter sensitivity analysis of λ_1 and λ_2.

5 Conclusion

How to unlock the hidden capacity of powerful TSFMs pre-trained from general time series tasks (e.g., finance, retail, power, health, and weather) for wearable activity recognition remains largely unexplored. In this work, we address this challenge by introducing a new Parameter-Efficient Fine-Tuning (PEFT) module named DeepSense-MoE, which incorporates three principal strategies including multi-scale convolution-based MoE, shared expert isolation, and hierarchical supervised contrastive alignment, aiming to leverage rich time series feature representations learned by general-purpose TSFM namely MOMENT during pre-training, while integrating intricate sensor knowledge through targeted fine-tuning for activity-specific tasks. Extensive experiments demonstrate that DeepSenseMoE can achieve substantial accuracy gains over state-of-the-art methods under various few-shot and full-shot scenarios with remarkably fewer trainable parameters, verifying its effectiveness, efficiency, and scalability, especially in low-data regimes. Additionally, we provide empirical analyses for every specific design choice to facilitate better application of DeepSenseMoE in wearable activity recognition. We hope that this work may serve as a solid foundation to support progress of TSFM+HAR in future work.

Acknowledgements

The authors would like to appreciate all participants of peer review. The work was supported by the National Natural Science Foundation of China under Grant 62373194, in part by Startup Funds from The University of Newcastle, Australia.

References

Awais, M.; Naseer, M.; Khan, S.; Anwer, R. M.; Cholakkal, H.; Shah, M.; Yang, M.-H.; and Khan, F. S. 2025. Foundation models defining a new era in vision: a survey and outlook. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Bai, S.; Kolter, J. Z.; and Koltun, V. 2018. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv preprint arXiv:1803.01271*.

Bulling, A.; Blanke, U.; and Schiele, B. 2014. A tutorial on human activity recognition using body-worn inertial sensors. *ACM Computing Surveys*.

Chen, K.; Zhang, D.; Yao, L.; Guo, B.; Yu, Z.; and Liu, Y. 2021. Deep learning for sensor-based human activity recognition: Overview, challenges, and opportunities. *ACM Computing Surveys*.

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. 2020. A simple framework for contrastive learning of visual representations. In *ICML*.

Dai, D.; Deng, C.; Zhao, C.; Xu, R.; Gao, H.; Chen, D.; Li, J.; Zeng, W.; Yu, X.; Wu, Y.; et al. 2024. DeepSeekMoE: Towards Ultimate Expert Specialization in Mixture-of-Experts Language Models. In *ACL*.

Das, A.; Kong, W.; Sen, R.; and Zhou, Y. 2024. A decoder-only foundation model for time-series forecasting. In *ICML*.

Fedus, W.; Zoph, B.; and Shazeer, N. 2022. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*.

Gao, Z.; Wang, Y.; Chen, J.; Xing, J.; Patel, S.; Liu, X.; and Shi, Y. 2023. MMTSA: Multi-modal temporal segment attention network for efficient human activity recognition. In *IMWUT/Ubicomp*.

Goswami, M.; Szafer, K.; Choudhry, A.; Cai, Y.; Li, S.; and Dubrawski, A. 2024. MOMENT: A Family of Open Time-series Foundation Models. In *ICML*.

He, K.; Fan, H.; Wu, Y.; Xie, S.; and Girshick, R. 2020. Momentum contrast for unsupervised visual representation learning. In *CVPR*.

Houlsby, N.; Giurgiu, A.; Jastrzebski, S.; Morrone, B.; De Laroussilhe, Q.; Gesmundo, A.; Attariyan, M.; and Gelly, S. 2019. Parameter-efficient transfer learning for NLP. In *ICML*.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W.; et al. 2022. Lora: Low-rank adaptation of large language models. In *ICLR*.

Khosla, P.; Teterwak, P.; Wang, C.; Sarna, A.; Tian, Y.; Isola, P.; Maschinot, A.; Liu, C.; and Krishnan, D. 2020. Supervised contrastive learning. In *NeurIPS*.

Kim, Y.-W.; Cho, W.-H.; Kim, K.-S.; and Lee, S. 2022. Inertial-measurement-unit-based novel human activity recognition algorithm using conformer. *Sensors*.

Logacjov, A. 2024. Self-supervised learning for Accelerometer-based human activity recognition: A survey. In *IMWUT/Ubicomp*.

Ma, H.; Zhang, Z.; Li, W.; and Lu, S. 2021. Unsupervised human activity representation learning with multi-task deep clustering. In *IMWUT/Ubicomp*.

Malekzadeh, M.; Clegg, R. G.; Cavallaro, A.; and Haddadi, H. 2018. Protecting sensory data against sensitive inferences. In *PbD-DS*.

Min, B.; Ross, H.; Sulem, E.; Veyseh, A. P. B.; Nguyen, T. H.; Sainz, O.; Agirre, E.; Heintz, I.; and Roth, D. 2023. Recent advances in natural language processing via large pre-trained language models: A survey. *ACM Computing Surveys*.

Ordóñez, F. J.; and Roggen, D. 2016. Deep convolutional and lstm recurrent neural networks for multimodal wearable activity recognition. *Sensors*.

Reiss, A.; and Stricker, D. 2012. Introducing a new benchmarked dataset for activity monitoring. In *ISWC*.

Sohn, K.; Berthelot, D.; Carlini, N.; Zhang, Z.; Zhang, H.; Raffel, C. A.; Cubuk, E. D.; Kurakin, A.; and Li, C.-L. 2020. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. In *NeurIPS*.

Stisen, A.; Blunck, H.; Bhattacharya, S.; Prentow, T. S.; Kjærgaard, M. B.; Dey, A.; Sonne, T.; and Jensen, M. M. 2015. Smart devices are different: Assessing and mitigatingmobile sensing heterogeneities for activity recognition. In *SenSys*.

Wang, R.; and He, K. 2025. Diffuse and Disperse: Image Generation with Representation Regularization. *arXiv preprint arXiv:2506.09027*.

Woo, G.; Liu, C.; Kumar, A.; Xiong, C.; Savarese, S.; and Sahoo, D. 2024. Unified Training of Universal Time Series Forecasting Transformers. In *ICML*.

Wu, H.; Hu, T.; Liu, Y.; Zhou, H.; Wang, J.; and Long, M. 2023. TimesNet: Temporal 2D-Variation Modeling for General Time Series Analysis. In *ICLR*.

Xia, K.; Li, W.; Gan, S.; and Lu, S. 2024a. TS2ACT: Few-shot human activity sensing with cross-modal co-learning. In *IMWUT/Ubicomp*.

Xia, K.; Li, W.; Shao, Y.; and Lu, S. 2024b. Vi2ACT: Video-enhanced Cross-modal Co-learning with Representation Conditional Discriminator for Few-shot Human Activity Recognition. In *ACM MM*.

Yang, J.; Nguyen, M. N.; San, P. P.; Li, X.; and Krishnaswamy, S. 2015. Deep convolutional neural networks on multichannel time series for human activity recognition. In *IJCAI*.

Zhou, T.; Niu, P.; Sun, L.; Jin, R.; et al. 2023. One fits all: Power general time series analysis by pretrained lm. In *NeurIPS*.