RobustHAR: Multi-scale Spatial-temporal Masked Self-supervised Pre-training for Robust Human Activity Recognition

Xiao Liu[1] , **Guan Yuan**[1,2*] , **Yanmei Zhang**[1] , **Shang Liu**[1] , **Qiuyan Yan**[1]

[1]School of Computer Science and Technology, China University of Mining and Technology
[2]Mine Digitization Engineering Research Center of the Ministry of Education
{liuxiao, yuanguan, ymzhang, shang, yanqy}@cumt.edu.cn

Abstract

Human activity recognition (HAR) is prone to performance degradation in real-world applications due to data missing between intra-sensor and inter-sensor channels. Masked modeling, as one mainstream paradigm of self-supervised pre-training, can learn robust representations across sensors in the data missing scenario by reconstructing the masked content based on the unmasked part. However, the existing methods predominantly emphasize the temporal dynamics of human activities, which limits their ability to effectively capture the spatial interdependencies among multiple sensors. Besides, different human activities often span across various spatial-temporal scales, which results in activity recognizer failing to capture intricate spatial-temporal semantic information. To address these issues, we propose RobustHAR, a new HAR model with multi-scale spatial-temporal masked self-supervised pre-training designed to improve model performance on the data missing context. RobustHAR involves three main steps: (1) RobustHAR constructs location-inspired spatial-temporal 3D-variation modeling to capture spatial-temporal correlated information in human activity data. (2) RobustHAR then designs multi-scale spatial-temporal masked self-supervised pre-training with semantic-consistent multi-scale feature co-learning for learning robust features at different scales. (3) Finally, RobustHAR fine-tunes the pretraining model with adaptive multi-scale feature fusion for human activity recognition. Extensive experiments on three public multi-sensor datasets demonstrate that RobustHAR outperforms existing state-of-the-art methods.

1 Introduction

Wearable sensor-based human activity recognition (WSHAR) has seen rapid development with the increasing availability of low-cost wearable devices. Compared with visual data [Liu *et al.*, 2024c; Liu *et al.*, 2024a; Zhou *et al.*, 2024] (such as RGB, depth, and skeleton), wearable sensors enable the recognition model focus more on motion information, with its robustness to environmental variations [Hong *et al.*, 2024]. Hence, WSHAR has attracted increasing research attention and demonstrates extensive applicability across diverse domains (including industrial control, and smart classrooms).

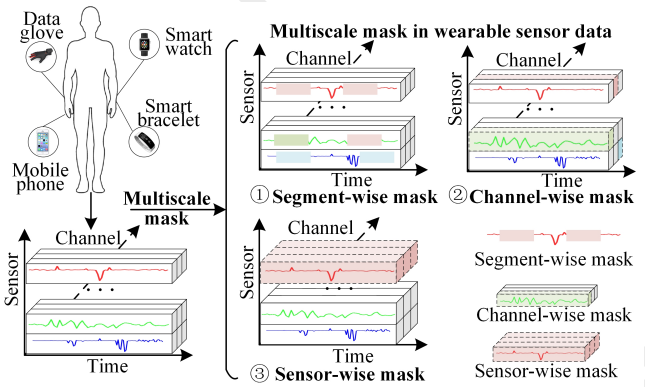

Figure 1: Multiscale masks in activity data from multiple wearable sensors can be classified into three types: segment-wise mask, channel-wise mask, and sensor-wise mask.

WSHAR is a standard time series classification task, owing to the temporal nature of the data generated by wearable sensors. Therefore, common time series classification methods like TimeNet [Wu *et al.*, 2023] and Informer [Zhou *et al.*, 2021] can enhance the performance of activity recognition to a certain extent. Nevertheless, traditional supervised learning methods are limited by extensive labeled datasets, which results in substantial labeling costs and potential inconsistencies. In practical application contexts, the collection and annotation of large-scale wearable sensor data often demand considerable time and financial resources, particularly in domains requiring specialized expertise. Furthermore, variations in labeling by different annotators can introduce inconsistencies that significantly affect the quality of model training, ultimately compromising its performance in real-world applications. To address the aforementioned issues, self-supervised pre-training can learn activity feature from large amounts of unlabeled data and then fine-tune for HAR with a limited set of labeled data.

*Corresponding author

Currently, contrastive learning [Zhang *et al.*, 2025a] and masked modeling [Haresamudram *et al.*, 2022] are two predominant approaches within self-supervised pre-training, widely used in WSHAR. Although contrastive learning can effectively extract features by data augmentation, it encounters considerable limitations in addressing data missing issues prevalent in real-world applications. This limitation arises from its reliance on complete sample pairs for training. Data missing compromises the construction of positive and negative pairs, leading to biased representations. In contrast, masked modeling can reconstruct the masked wearable sensor data based on the unmasked part. This approach facilitates the model's ability to learn robust activity representations without dependence on complete data. However, since semantic information of sensor-based action data is mainly contained in temporal variations [Dong *et al.*, 2024], the existing masked modeling often ignores spatial correlated information among different wearable sensors. Spatial-temporal masked modeling addresses this gap by integrating spatial and temporal dimensions, allowing the model to leverage spatial relationships while capturing temporal variations. Common spatial-temporal masked modeling (such as MaskCAE [Cheng *et al.*, 2024], STD-MAE [Gao *et al.*, 2024], ST-MAE [Miao *et al.*, 2024]) often separates temporal masking from spatial masking, which restricts the interaction between spatial and temporal information among wearable sensor data. Besides, different human activities often span across various spatial-temporal scales, which results in activity recognizer failing to capture intricate spatial-temporal semantic information, as shown in Fig.1.

Motivated by the above observations, we propose RobustHAR, a new HAR model with multi-scale spatial-temporal masked self-supervised pre-training. Firstly, RobustHAR extracts spatial-temporal correlated features among inter-sensor and intra-sensor by constructing location-inspired spatial-temporal 3D-variation modeling. Secondly, multi-scale spatial-temporal masked pre-training enables activity recognition model to learn robust feature across various temporal and spatial scales, thereby improving its capacity to capture the contextual information inherent in human activity. Finally, a small amount of labeled activity data is used to fine-tune the pre-training model for human activity recognition. Besides, Extensive experiments conducted on three public multi-sensor datasets demonstrate that RobustHAR outperforms the current state-of-the-art algorithms. The main contributions of our work are as follows:

- Location-inspired spatial-temporal 3D-variation modeling is constructed to mine spatial-temporal correlated information among human activity data.
- Multi-scale spatial-temporal masked pre-training with semantic-consistent multi-scale feature co-learning is designed for capturing robust spatial-temporal contextual information at different scales.
- RobustHAR fine-tunes the pre-training model with a small set of labeled data for human activity recognition by adaptive multiscale feature fusion.

2 Related Work

2.1 Wearable sensor-based human activity recognition

WSHAR can provide personalized solutions for human-computer interaction by leveraging real-time, continuous data collection from wearable sensors. Early human activity recognition models often rely on a single wearable sensor, such as accelerometers, gyroscopes, and magnetometers [Chen *et al.*, 2023]. However, using a single sensor for capturing human activity data may introduce measurement bias, which limits its ability to distinguish ambiguous activity with similar data sequences [Liu *et al.*, 2024b]. Hence, integrated devices (such as smart wristbands, smartphones, and data gloves) have been increasingly utilized in HAR, a trend made possible by the rapid advancements in sensor technology. Meanwhile, the extensive application of deep learning (such as GRU [Lalwani and Ramasamy, 2024; Pandey and Kumar, 2024], Transformer [Kitaev *et al.*, 2020; Zhou *et al.*, 2021], GNN [Zhang *et al.*, 2023; Wei *et al.*, 2024; Wei *et al.*, 2025], and Mamba[Li *et al.*, 2024; Zhang *et al.*, 2025b]) has substantially improved the performance of WSHAR. In real-world applications, factors such as sensor malfunctions and signal loss can easily lead to the loss of collected human activity data. Therefore, constructing a robust HAR model has become a primary focus of current research.

2.2 Self-supervised pre-training learning for human activity recognition

Common self-supervised pre-training learning for HAR includes contrastive learning and masked modeling [Haresamudram *et al.*, 2022; Jain *et al.*, 2022]. Contrastive learning [Haresamudram *et al.*, 2021] aims to maximize the similarity between positive samples and minimize the distance between negative samples, which can guide the model to learn robust feature representations. However, contrastive learning is not suitable for data-missing scenario. The primary challenge lies in the fact that contrastive learning relies on maintaining consistency in the feature space across different action samples. In contrast, masked modeling [Cheng *et al.*, 2024; Gao *et al.*, 2024; Miao *et al.*, 2024] can reconstruct missing activity data based on the unmasked data. It can dynamically compensate for missing activity data, thereby enhancing data integrity and improving the model's generalization capacity. Nevertheless, since semantic information in sensor-based activity data is often contained in temporal variations [Dong *et al.*, 2024], most existing masked modeling often ignores spatially correlated information among different wearable sensors. Furthermore, the varying spatial-temporal scales of human activities often hinder activity recognizers from capturing fine-grained spatiotemporal semantics. As a result, existing self-supervised frameworks struggle to effectively model complex spatial-temporal dependencies under multi-scale data-missing conditions.

3 Methodology

As mentioned above, RobustHAR aims to capture multi-scale spatial-temporal correlated information among human activ-

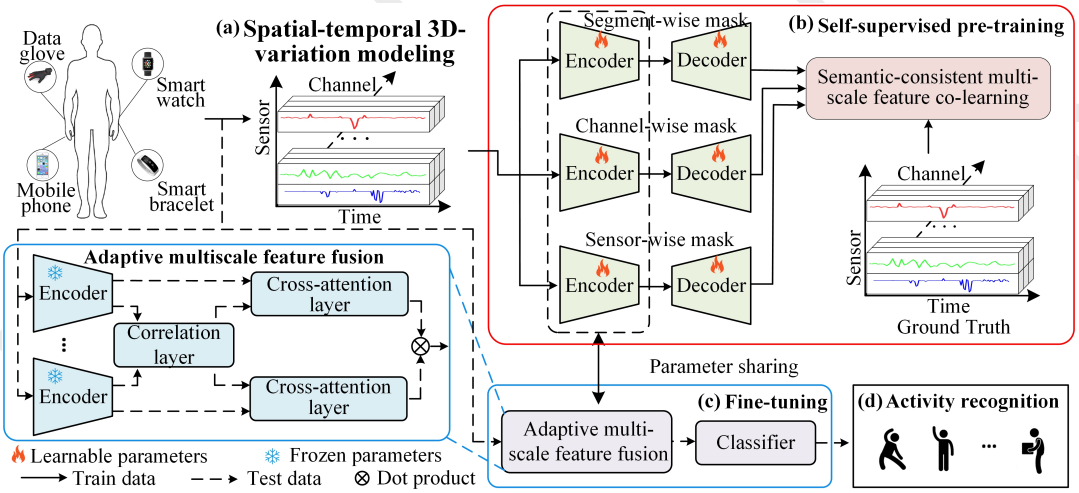

Figure 2: Research Framework. RobustHAR is comprised of three key components: location-inspired spatial-temporal 3D-variation modeling, multi-scale spatial-temporal masked pre-training, and fine-tuning for HAR. The location-inspired spatial-temporal 3D-variation modeling is designed to uncover the intricate spatial-temporal correlated information within human activity data. Based on it, multi-scale spatial-temporal masked pre-training employs multi-scale masking based on positional information and incorporates a semantic-consistent multi-scale feature co-learning to capture contextual semantics across various scales. Finally, fine-tuning for HAR is conducted with adaptive multiscale feature fusion to enhance recognition performance.

ity data in the data-missing scenario, as shown in Fig. 2. The details of RobustHAR are described as follows.

3.1 Location-inspired spatial-temporal 3D-variation modeling

Human activity is temporally dynamic and spatially complex, especially in multi-sensor environments, where the positional relationships and interactions between different sensors can significantly impact the performance of HAR. By considering the spatial distribution of sensors, we can better understand the relative relationships of each body parts during movement, thereby extracting richer spatial-temporal activity features. Therefore, RobustHAR leverages location-inspired spatial-temporal 3D-variation modeling to capture the spatial-temporal correlated information among wearable sensor-based human activity data.

Given that the multi-dimensional data collected from N sensors is represented as $\mathbf{D} = \{d_{i,j}\}_{i=1}^{N} \in \mathbb{R}^{T\times N\times C}, \quad j = 1, 2, \ldots, T.$ where $d_{i,j}$ denotes the measurement value of the i-th sensor at time j, C denotes the dimensionality of the features generated by each sensor, and T is the number of time steps. Meanwhile, each sensor's relative spatial position is represented by position matrix $\mathbf{A}$. Therefore, we can transform the original multisensor human activity data into location-inspired spatial-temporal 3D-variation, and the transformation formula is shown as follows:

$$\mathbf{X} = \mathbf{DA}, \tag{1}$$

where $\mathbf{X} \in \mathbb{R}^{T\times C\times N}$ denotes the spatial-temporal 3D-variation, as shown in Fig.1. The spatial-temporal 3D variation $\mathbf{X}$ represents human activity data that are spatially adjacent and also adjacent in the semantic space. This adjacency ensures the coherence and validity of human activity data. Specifically, we map the sensors to the Z-axis of a 3D coordinate system according to their spatial positions on the human body. Subsequently, the time and feature dimensions of the sensors along the Z-axis are projected onto the X and Y axes of the 3D coordinates.

3.2 Multiscale spatial-temporal masked self-supervised pre-training

Multi-scale spatial-temporal masked: Human activities inherently exhibit multi-scale temporal and spatial characteristics. For instance, actions such as walking and running, though similar in basic motion patterns, differ significantly in their spatial-temporal representations. To capture such fine-grained distinctions, multi-scale spatial-temporal masked is used to extract rich contextual semantics across diverse scales. This approach not only enhances the model's ability to learn comprehensive activity features but also improves robustness in data-missing scenarios by leveraging spatial-temporal correlations at different scales.

Common masks in WSHAR include segment-wise mask, channel-wise mask, and sensor-wise mask. Segment-wise mask, channel-wise mask, and sensor-wise mask are in a similar manner, so we take sensor-wise mask as an example. Given spatial-temporal 3D-variation $\mathbf{X} \in \mathbb{R}^{T\times C\times N}$, we randomly select N_m out of N wearable sensors for masking. Meanwhile, we define the spatial masked matrix $\mathbf{M} \in \mathbb{R}^{T\times C\times N}$ with the same shape as spatial-temporal 3D-variation $\mathbf{F}$, so $\mathbf{M}$ can be represented as follows:

$$\mathbf{M} = \begin{cases} 0 & , \mathrm{z} \in Random(N_m), \\ 1 & , \text{others}, \end{cases} \tag{2}$$

where z represents the Z-axis index of spatial-temporal 3D-variation $\mathbf{X}$. So we obtain sensor-wise masked human activity

data by the following equation:

$$\mathbf{F^s_m} = \mathbf{Encoder}(\mathbf{M} \odot \mathbf{X}), \tag{3}$$

where $\odot$ represents element-wise product, $\mathbf{F^s_m}$ means sensor-wise masked human activity feature, $\mathbf{Encoder}(\cdot)$ denotes feature extractor, we use transformer as feature encoder. Similarly, we can obtain segment-wise masked activity feature $\mathbf{F^{se}_m}$, and channel-wise masked activity feature $\mathbf{F^c_m}$.

After obtaining the masked multi-scale activity data features, we restore them to a form that closely approximates the original data by the decoder layer. This decoding process not only recovers the masked spatial-temporal activity data, but also reinforces the model's comprehension of data integrity by progressively reconstructing human activity features. Such a process enhances the model's robustness in the presence of data-missing interference, while facilitating the extraction of more meaningful features during training, thereby improving the overall performance in activity recognition tasks. We use $\mathbf{X^s_m}$, $\mathbf{X^{se}_m}$, and $\mathbf{X^c_m}$ to represent reconstructed sensor-wise activity data, segment-wise activity data, and channel-wise activity data, respectively. Hence, the decoding process is shown in the following equation:

$$\begin{cases} \mathbf{X^s_m} = \mathbf{Decoder}(\mathbf{F^s_m}), \\ \mathbf{X^{se}_m} = \mathbf{Decoder}(\mathbf{F^{se}_m}), \\ \mathbf{X^c_m} = \mathbf{Decoder}(\mathbf{F^c_m}), \end{cases} \tag{4}$$

where $\mathbf{Decoder}(\cdot)$ denotes decoder layer, we use Transformer as the decoder layer. Given ground truth of the masked activity data $\mathbf{X^s}$, $\mathbf{X^{se}}$, and $\mathbf{X^c}$, we can compute the loss L_{rec} among ground truth and the reconstructed data by mean absolute error (MAE), as shown in the following equation:

$$\mathcal{L}_{rec} = \mathcal{L}^s_{rec} + \mathcal{L}^{se}_{rec} + \mathcal{L}^c_{rec}, \tag{5}$$

where $\mathcal{L}^s_{rec} = \frac{1}{B}\sum_{i=1}^{B} \mid X^s_i - X^s_{m,i} \mid$, $\mathcal{L}^{se}_{rec} = \frac{1}{B}\sum_{i=1}^{B} \mid X^{se}_i - X^{se}_{m,i} \mid$, and $\mathcal{L}^c_{rec} = \frac{1}{B}\sum_{i=1}^{B} \mid X^c_i - X^c_{m,i} \mid$ represent sensor-wise reconstructed loss, segment-wise reconstructed loss, and channel-wise reconstructed loss, respectively. B denotes the number of batch size.

Semantic-consistent multi-scale feature co-learning: When performing masking operations on multi-scale data (such as segment-wise mask, channel-wise mask, and sensor-wise mask), HAR model needs to ensure that the information representation at different scales remains consistent. If the masking operations at each scale lead to inconsistent semantic information, HAR model may lose its overall understanding of human activity, resulting in a decline in recognition performance. Besides, semantic-consistent multi-scale feature co-learning can enhance the robustness of HAR model when wearable sensor data at one scale is incomplete, the others can still provide sufficient semantic information. Hence, we introduce similarity metric to ensure consistent semantic representations in the HAR model under different masking scales, as shown in the following equation:

$$\mathcal{L}_{con} = \mathbf{sim}(\mathbf{X^s_a}, \mathbf{X^{se}_a}) + \mathbf{sim}(\mathbf{X^s_a}, \mathbf{X^c_a}) + \mathbf{sim}(\mathbf{X^c_a}, \mathbf{X^{se}_a}), \tag{6}$$

where $\mathbf{X^i_a}, i \in \{se, s, c\}$ denotes complete human activity data that its masked portions has been replaced by the generated data, $\mathbf{sim}(\cdot)$ represents similarity metric function (such as cosine similarity).

As a result, the loss function $\mathcal{L}_{pre}$ of multi-scale spatial-temporal masked self-supervised pretraining model mainly consists of multi-scale reconstruction loss and semantic consistency multi-scale feature co-learning loss, as shown in the following equation.

$$\mathcal{L}_{pre} = \lambda\mathcal{L}_{rec} + (1-\lambda)\mathcal{L}_{con}, \tag{7}$$

where λ represents balance coefficient in the loss function. The former is used to ensure the accurate reconstruction of the masked data across different scales, helping the model recover the missing information and preserve the original structure of human activity data. The latter enforces semantic consistency across the different scales, ensuring that the feature representations at each scale align in terms of their underlying semantic information.

3.3 Fine-tuning with adaptive multi-scale feature fusion

From multi-scale spatial-temporal masked self-supervised pretraining, we can obtain multi-scale feature encoders $\mathbf{Encoder^s}$, $\mathbf{Encoder^{se}}$, and $\mathbf{Encoder^c}$. Given a small set of labeled human activity data $\mathbf{D^l} = \{\mathbf{d^l_i}, \mathbf{y_i}\}, i = 1, 2, \cdots, n_l$, where n_l denotes the number of labeled activity data, y_i represents the label of activity sample data d^l_i. We then extract multi-scale activity features by encoder after self-supervised pre-training, as shown in the following equation:

$$\mathbf{F^j_i} = \mathbf{Encoder^j_i}(\mathbf{d^l_i}), \tag{8}$$

where $j \in \{s, se, c\}$ means sensor-wise operation, segment-wise operation, and channel-wise operation. Features at different scales can capture distinct aspects of human activity. For example, segment-wise features focus on subtle motion details, while sensor-wise features highlight the spatial correlations between different sensors. Based on it, we construct adaptive multi-scale feature fusion [Wu *et al.*, 2024; Guan *et al.*, 2025; Liu *et al.*, 2025] to integrate features across various scales by cross-attention [Hou *et al.*, 2019], as shown in Fig.2.

We use sensor-wise activity feature $\mathbf{F^s}$ and channel-wise feature $\mathbf{F^c}$ as examples. Firstly, we construct correlation layer to calculate a relationship between sensor-wise and channel-wise human activity features with cosine distance, as calculated in Equation (9):

$$\begin{cases} \mathbf{R^s} = & (\dfrac{\mathbf{F^s}}{\parallel \mathbf{F^c} \parallel_2})^{\mathrm{T}}(\dfrac{\mathbf{F^s}}{\parallel \mathbf{F^c} \parallel_2}), \\ \mathbf{R^c} = & (\dfrac{\mathbf{F^c}}{\parallel \mathbf{F^s} \parallel_2})^{\mathrm{T}}(\dfrac{\mathbf{F^c}}{\parallel \mathbf{F^s} \parallel_2}), \end{cases} \tag{9}$$

where $\mathbf{R^s}$ and $\mathbf{R^c}$ represent the relevance between sensor-wise and channel-wise human activity features. Cross-attention mechanism is then used to generate sensor-wise $\mathbf{A^s}$ and channel-wise $\mathbf{A^c}$ attention maps, respectively. Given the sensor-wise correlation $\mathbf{R^s}$ and channel correlation $\mathbf{R^c}$, so fused feature can be calculated in Equation (10).

$$\begin{cases} \mathbf{A^s} = & \mathbf{Crossatt}(\mathbf{R^s}, \mathbf{F^s}), \\ \mathbf{A^c} = & \mathbf{Crossatt}(\mathbf{R^c}, \mathbf{F^c}), \end{cases} \tag{10}$$

Dataset	Devices	Activities	Subjects	Size
Opportunity	5	17	4	8165
RealWorld	7	8	13	39281
CZU-MHAD	10	22	7	880

Table 1: Statistical description of datasets.

where $\mathbf{Crossatt}(\cdot)$ denotes cross-attention. Subsequently, we integrate sensor-wise information and channel-wise information to more accurately capture the dynamic features of human activity, as shown in Equation (11).

$$\mathbf{F^{sc}} = (\mathbf{A^s} \otimes \mathbf{A^c}), \tag{11}$$

where $\mathbf{F^{sc}}$ represents fused human activity features, $\otimes$ denotes dot product. Finally, the fusion of human activity features is used to predict the label of human activity data ($\mathbf{D^l}$). Meanwhile, the cross-entropy loss function is introduced to train HAR model.

3.4 Complexity analysis

The overall computing complexity of RobustHAR is $O(LTNC^2)$, where L is the number of Transformer layers, T is the time step, N is the number of wearable sensors, and C is the dimensionality of the features generated by wearable sensors. It is slightly higher than other models (such as ST-MAE: $O(LTNC)$). However, its robustness and accuracy in data-missing scenarios far exceed those of the others.

4 Experiments

4.1 Datasets and Evaluation Metrics

Datasets. To validate the effectiveness of the proposed method, we conduct experiments on the following three multi-sensor public datasets: Opportunity [Roggen *et al.*, 2010], RealWorld [Sztyler and Stuckenschmidt, 2016], and CZU-MHAD [Chao *et al.*, 2022], the detailed description of datasets is shown in the Table.1.

Evaluation Metrics. The evaluation metrics are essential for assessing the robustness and generalization of the human activity recognition models. We employ accuracy (denoted as Acc), and F1 score as the evaluation metrics for the model.

4.2 Baselines

To evaluate whether RobustHAR can achieve performance comparable to supervised learning with limited labeled data, we have chosen four supervised HAR models (Reformer [Kitaev *et al.*, 2020], Informer [Zhou *et al.*, 2021], TimeNet [Wu *et al.*, 2023], and MSGNet [Cai *et al.*, 2024]). Meanwhile, to verify the superiority of RobustHAR in the spatial-temporal feature extraction, we have selected four self-supervised HAR models (SimMTM [Dong *et al.*, 2024], MaskCAE [Cheng *et al.*, 2024], STD-MAE [Gao *et al.*, 2024], ST-MAE [Miao *et al.*, 2024]).

4.3 Implementation Details

To ensure fair comparisons, we carefully tune all models, including both baselines and our proposed RobustHAR. Specifically, we first initialize hyperparameters in each baseline according to guidelines provided in the original papers, and then fine-tune them on our used datasets for ensuring their fair performance. The experiments are conducted with two NVIDIA GeForce RTX 4090 GPUs with 24GB memory. All models are implemented with Python 3.9 and PyTorch 2.4. In the pre-training phase, we set the batch size to 64 and the number of training epochs to 200, with a learning rate of 0.001. Encoder layer and decoder layer both use a single-layer transformer with 8 attention heads. During the fine-tuning RobustHAR, the batch size is set to 32 and the number of training epochs is set to 100, with the learning rate of 0.001. Meanwhile, we divide datasets into training, validation, and test sets in a ratio of 8:1:1. When fine-tuning the pre-training model with a small set of labeled data, the labeled data accounts for 15% of the training dataset. Besides, balance coefficient λ is chosen from $\{0.1, 0.3, 0.5, 0.7, 0.9\}$, and masking rate is selected from $\{0.1, 0.3, 0.5, 0.7, 0.9\}$. All experimental results are reported as mean values obtained from five independent trials.

4.4 Comparison with baselines

The experimental results demonstrate that RobustHAR achieves superior performance compared to other baselines on the Opportunity, RealWorld, and CZU-MHAD datasets, as shown in the table 2. Specifically, Transformer-based action recognition models (such as Reformer and Informer) can leverage the modeling capacity of a Transformer with an architecture that can capture long-sequence semantic information of activity data. However, these methods overlook the varying inter-series correlations across different time scales. Hence, TimeNet and MSGNet can improve the accuracy of activity recognition. Compared to supervised learning models, self-supervised learning-based human activity recognition models, particularly in masked modeling (such as SimMTM, MaskMAE, and ST-MAE), these models can enable the activity recognition model to learn global and contextual semantic features by masking portions of the input data. Meanwhile, STD-MAE can model the heterogeneity of spatial-temporal activity data by spatial-temporal-decoupled masked pre-training. While these masked modeling methods can effectively capture the spatial-temporal features of human activities, they often fail to account for the influence of spatial-temporal features across different scales on HAR. Therefore, RobustHAR achieves excellent recognition per-

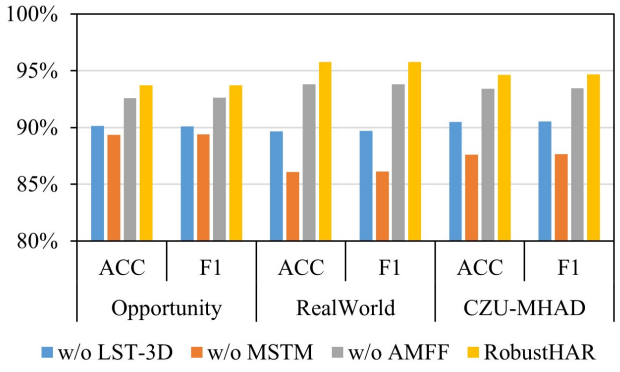

Figure 3: Ablation studies.

Method	Venue	Opportunity		RealWorld		CZU-MHAD	
		ACC(%)↑	F1(%)↑	ACC(%) ↑	F1(%)↑	ACC(%)↑	F1(%)↑
Reformer	ICLR'20	86.58	86.62	89.79	89.83	88.35	88.43
Informer	AAAI'21	89.64	89.65	90.37	90.35	90.46	90.48
TimeNet	ICLR'23	91.05	91.02	91.74	91.77	92.19	92.33
MSGNet	AAAI'24	92.23	92.24	92.43	92.45	91.86	91.83
SimMTM	NeurIPS'23	92.51	92.49	92.04	92.06	91.46	91.48
MaskCAE	IJBHI'24	90.34	90.36	92.65	92.67	89.72	89.73
STD-MAE	IJCAI'24	92.43	92.42	93.63	93.66	91.39	91.41
ST-MAE	Ubicomp'24	92.65	92.58	93.07	93.11	92.73	92.74
RobustHAR	-	**93.72**	**93.74**	**95.76**	**95.78**	**94.63**	**94.69**

Table 2: Comparison with baselines.

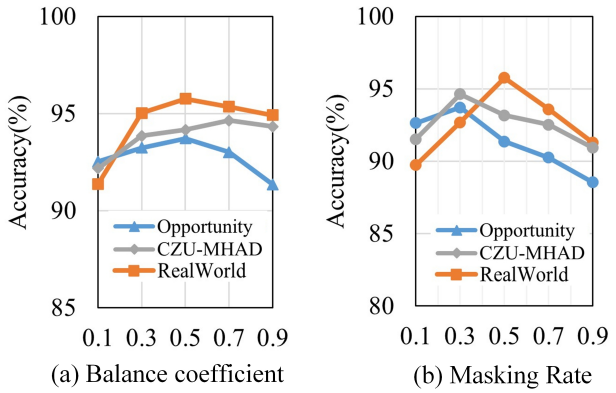

Figure 4: Parameter Analysis.

formance on the Opportunity (ACC:93.72%, F1: 93.74%), RealWorld (ACC: 95.76%, F1: 95.78%), and CZU-MHAD datasets (ACC:94.63%, F1: 94.69%).

4.5 Ablation Studies

RobustHAR primarily consists of location-inspired spatial-temporal 3D-variation modeling module (LST-3D), multi-scale spatial-temporal masked self-supervised pre-training module (MSTM), and fine-tuning the pre-training model with adaptive multiscale feature fusion (AMFF). In order to evaluate the contributions of each individual module, we conduct a series of ablation experiments. These experiments systematically remove (denoted as w/o) specific components to assess their impact on the overall performance, as shown in Fig.3. Specifically, the experimental results indicate that the performance of RobustHAR is most significantly impacted when multi-scale spatial-temporal masked pre-training is not employed. On the RealWorld dataset, this results in a 10% reduction in recognition accuracy of RobustHAR. Furthermore, it is apparent that both the location-inspired spatial-temporal 3D-variation modeling and fine-tune pre-training model with adaptive multi-scale feature fusion play crucial roles in enhancing the final recognition performance. Overall, RobustHAR can capture the spatial-temporal relationships among inter-sensor and intra-sensor at different scales, thereby ensuring the semantic consistency of features across various scales.

4.6 Parameters Analysis

Impact of balance coefficient in the loss function. During the pre-training phase, loss function $\mathcal{L}_{pre}$ in RobustHAR is primarily composed of reconstruction loss $\mathcal{L}_{rec}$ and semantic-consistency loss $\mathcal{L}_{con}$. The former ensures that RobustHAR can recover the masked human activity data with the unmasked part, thereby capturing intricate spatial-temporal correlated information among inter-sensor and intra-sensor channels. The latter enforces semantic-consistent in activity features across different scales. Hence, how to balance the reconstruction loss and semantic-consistent loss is crucial to achieving optimal performance in RobustHAR. As shown in Fig.4(a), RobustHAR attains the highest recognition accuracy on the Opportunity and RealWorld dataset when balance coefficient is set to 0.5. Meanwhile, RobustHAR achieves the best performance on the CZU-MHAD dataset when balance coefficient is set to 0.7.

Impact of masking rate during the pre-training phase. The masking rate in the pre-training phase plays a critical role in determining both the learning effectiveness and the ultimate performance of HAR model. An appropriately chosen masking rate strikes a balance between the complexity of HAR task and the capacity of HAR to learn meaningful and generalized representations from human activity data. When the masking rate is too low, HAR model has access to an excessive amount of information, facilitating rapid convergence but potentially leading to overfitting. It primarily stems from HAR model's excessive reliance on the available human activity data. Conversely, if the masking rate is too high, HAR model may struggle to learn sufficient meaningful patterns of human activity. This can result in hindering its ability to capture the underlying structure of human movements. As shown in Fig.4(b), when masking rate is set to 0.3, RobustHAR achieves the highest recognition accuracy on the Opportunity and CZU-MHAD datasets; when the masking rate is set to 0.5, RobustHAR performs best on the RealWorld dataset.

4.7 Robustness for HAR in the data-missing scenario

Human activity recognition (HAR) models often face challenges when data is missing, either due to sensor failures or noise in the environment. To evaluate the robustness of the RobustHAR model under such conditions, we simulate data-missing scenarios by randomly removing data points at

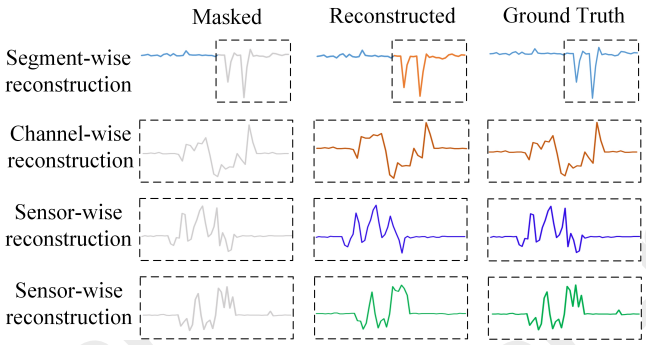

Figure 5: Reconstruction performance in self-supervised pre-training. The dashed box indicates the masked and reconstructed parts of human activity data.

different scales. Specifically, we analyze the robustness of different methods in data-missing scenarios by randomly setting different masking rates among different scales. For the data-missing rates, we select three conditions: 5%, 10%, and 15%. The experimental results are shown in Table.3. As can be seen from the table, the robustness of the self-supervised pre-trained model outperforms that of the supervised model. More importantly, RobustHAR achieves the best robust performance by capturing spatial-temporal dependencies at different scales.

4.8 Case study

The reconstruction performance during the self-supervised pre-training phase plays a crucial role in understanding the intricate internal structure of human activity data. We examine how well RobustHAR can recover the missing information, as well as the quality of the reconstructed features, by comparing the reconstructed data to the original input. As illustrated in Fig.5, RobustHAR exhibits varying degrees of reconstruction accuracy depending on the scale at which human activity data is masked. At the segment-wise mask, RobustHAR can capture long-range dependencies between different activities. Meanwhile, the reconstruction performance applied at the channel-wise mask remains robust. This suggests that RobustHAR can retain the channel-specific features for accurately recognizing human activities. Besides, at the sensor-wise level, where individual sensor data is masked, RobustHAR can effectively exploit cross-sensor correlations to reconstruct the missing sensor values, demonstrating its ability to capture spatial dependencies between different sensor locations. Although there is a discrepancy between the reconstructed human activity data (sensor-wise mask) and the original ground-truth data, sensor-wise features can ensure semantic coherence across different feature scales by semantic-consistent multi-scale feature co-learning. This consistency enables the model to better leverage a small amount of labeled data during the fine-tuning stage.

5 Conclusion

This paper proposes RobustHAR, a novel multi-scale spatial-temporal masked self-supervised pre-training framework.

Method	Opportunity	RealWorld	CZU-MHAD
Reformer	84.58%	86.59%	84.42%
Informer	87.82%	87.63%	88.64%
Timenet	90.26%	89.94%	88.57%
MSGNet	90.37%	90.88%	89.39%
SimMTM	90.87%	91.35%	90.12%
MaskCAE	88.24%	89.17%	85.93%
STD-MAE	90.61%	91.32%	88.29%
ST-MAE	90.04%	90.45%	89.32%
RobustHAR	**92.96**%	**94.85**%	**93.58**%
Reformer	81.92%	82.75%	79.94%
Informer	83.28%	84.68%	84.21%
Timenet	85.69%	85.37%	84.62%
MSGNet	85.21%	85.72%	86.77%
SimMTM	88.56%	87.61%	88.54%
MaskCAE	84.17%	86.44%	84.65%
STD-MAE	87.73%	88.83%	83.82%
ST-MAE	85.75%	87.27%	85.73%
RobustHAR	**91.37**%	**94.04**%	**92.17**%
Reformer	75.85%	74.68%	70.78%
Informer	77.64%	75.94%	78.66%
Timenet	78.46%	77.28%	80.36%
MSGNet	76.91%	79.63%	80.25%
SimMTM	82.48%	81.85%	83.72%
MaskCAE	79.59%	80.46%	82.57%
STD-MAE	83.46%	82.67%	79.06%
ST-MAE	80.62%	83.11%	81.57%
RobustHAR	**88.67**%	**86.73**%	**89.52**%

Table 3: Robustness for HAR in the data-missing scenario. The data-missing rate in the green section is 5%, in the red section is 10%, and in the blue section is 15%. The values in the table are recognition accuracy.

It not only captures temporal-spatial correlations information among intra-sensors and inter-sensors by location-inspired spatial-temporal 3D modeling, but also effectively mines multi-scale robust activity features with multi-scale spatial-temporal masked pretraining model. Meanwhile, RobustHAR ensures semantic consistency across activity features at different scales. Besides, experimental results demonstrate that RobustHAR surpasses state-of-the-art baselines, and maintains robust performance for HAR in the data-missing scenario.

Acknowledgments

This work was supported in part by the Nature Science Foundation of China under Grant No.62277046, the Key Research and Development Program of Xuzhou under Grant No.KC23296, the Science and Technology Program of Xuzhou under Grant No.KC22047, the Graduate Innovation Program of China University of Mining and Technology under Grant No.2023WLKXJ179, the Fundamental Research Funds for the Central Universities under Grant No.2023XSCX048, the Postgraduate Research & Practice Innovation Program of Jiangsu Province under Grant No.KYCX23_2728.

References

[Cai *et al.*, 2024] Wanlin Cai, Yuxuan Liang, Xianggen Liu, Jianshuai Feng, and Yuankai Wu. Msgnet: Learning multi-scale inter-series correlations for multivariate time series forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 11141–11149, 2024.

[Chao *et al.*, 2022] Xin Chao, Zhenjie Hou, and Yujian Mo. Czu-mhad: a multimodal dataset for human action recognition utilizing a depth camera and 10 wearable inertial sensors. *IEEE Sensors Journal*, 22(7):7034–7042, 2022.

[Chen *et al.*, 2023] Ling Chen, Yi Zhang, Shenghuan Miao, Sirou Zhu, Rong Hu, Liangying Peng, and Mingqi Lv. Salience: An unsupervised user adaptation model for multiple wearable sensors based human activity recognition. *IEEE Transactions on Mobile Computing*, 22(9):5492–5503, 2023.

[Cheng *et al.*, 2024] Dongzhou Cheng, Lei Zhang, Lutong Qin, Shuoyuan Wang, Hao Wu, and Aiguo Song. Maskcae: Masked convolutional autoencoder via sensor data reconstruction for self-supervised human activity recognition. *IEEE Journal of Biomedical and Health Informatics*, 28(5), 2024.

[Dong *et al.*, 2024] Jiaxiang Dong, Haixu Wu, Haoran Zhang, Li Zhang, Jianmin Wang, and Mingsheng Long. Simmtm: A simple pre-training framework for masked time-series modeling. In *Proceedings of Advances in Neural Information Processing Systems*, volume 36, 2024.

[Gao *et al.*, 2024] Haotian Gao, Renhe Jiang, Zheng Dong, Jinliang Deng, and Xuan Song. Spatio-temporal-decoupled masked pre-training for traffic forecasting. In *Proceedings of International Joint Conference on Artificial Intelligence*, 2024.

[Guan *et al.*, 2025] Renxiang Guan, Wenxuan Tu, Siwei Wang, Jiyuan Liu, Dayu Hu, Chang Tang, Yu Feng, Junhong Li, Baili Xiao, and Xinwang Liu. Structure-adaptive multi-view graph clustering for remote sensing data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 16933–16941, 2025.

[Haresamudram *et al.*, 2021] Harish Haresamudram, Irfan Essa, and Thomas Plötz. Contrastive predictive coding for human activity recognition. In *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, pages 1–26. ACM New York, NY, USA, 2021.

[Haresamudram *et al.*, 2022] Harish Haresamudram, Irfan Essa, and Thomas Plötz. Assessing the state of self-supervised human activity recognition using wearables. In *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, pages 1–47. ACM New York, NY, USA, 2022.

[Hong *et al.*, 2024] Zhiqing Hong, Zelong Li, Shuxin Zhong, Wenjun Lyu, Haotian Wang, Yi Ding, Tian He, and Desheng Zhang. Crosshar: Generalizing cross-dataset human activity recognition via hierarchical self-supervised pretraining. In *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, pages 1–26. ACM New York, NY, USA, 2024.

[Hou *et al.*, 2019] Ruibing Hou, Hong Chang, Bingpeng Ma, Shiguang Shan, and Xilin Chen. Cross attention network for few-shot classification. In *Proceedings of Advances in Neural Information Processing Systems*, volume 32, 2019.

[Jain *et al.*, 2022] Yash Jain, Chi Ian Tang, Chulhong Min, Fahim Kawsar, and Akhil Mathur. Collossl: Collaborative self-supervised learning for human activity recognition. In *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, pages 1–28. ACM New York, NY, USA, 2022.

[Kitaev *et al.*, 2020] Nikita Kitaev, Lukasz Kaiser, and Anselm Levskaya. Reformer: The efficient transformer. In *Proceedings of International Conference on Learning Representations*, pages 1–12, 2020.

[Lalwani and Ramasamy, 2024] Pooja Lalwani and Ganeshan Ramasamy. Human activity recognition using a multi-branched cnn-bilstm-bigru model. *Applied Soft Computing*, 154(3):111344, 2024.

[Li *et al.*, 2024] Shuangjian Li, Tao Zhu, Furong Duan, Liming Chen, Huansheng Ning, Christopher Nugent, and Yaping Wan. Harmamba: efficient and lightweight wearable sensor human activity recognition based on bidirectional mamba. *IEEE Internet of Things Journal*, 12(3):2373–2384, 2024.

[Liu *et al.*, 2024a] Jinfu Liu, Chen Chen, and Mengyuan Liu. Multi-modality co-learning for efficient skeleton-based action recognition. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 4909–4918, 2024.

[Liu *et al.*, 2024b] Xiao Liu, Guan Yuan, Rui Bing, Zhuo Cai, Shengshen Fu, and Yonghao Yu. When skeleton meets motion: adaptive multimodal graph representation fusion for action recognition. In *Proceedings of 2024 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2024.

[Liu *et al.*, 2024c] Yang Liu, Fang Liu, Licheng Jiao, Qianyue Bao, Lingling Li, Yuwei Guo, and Puhua Chen. A knowledge-based hierarchical causal inference network for video action recognition. *IEEE Transactions on Multimedia*, 26(4):9135–9149, 2024.

[Liu *et al.*, 2025] Yang Liu, Lei Si, Zhongbin Wang, Dong Wei, Xin Li, and Jinheng Gu. Dual discriminator and adaptive multisource feature fusion wgan-gp for coal-rock properties recognition under limited infrared thermal images. *IEEE Transactions on Industrial Informatics*, pages 1–12, 2025.

[Miao *et al.*, 2024] Shenghuan Miao, Ling Chen, and Rong Hu. Spatial-temporal masked autoencoder for multi-device wearable human activity recognition. In *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, pages 1–25. ACM New York, NY, USA, 2024.

[Pandey and Kumar, 2024] Ajeet Pandey and Piyush Kumar. Residual deep gated recurrent unit-based attention framework for human activity recognition by exploiting dilated features. *The Visual Computer*, 40(2):1–20, 2024.

[Roggen *et al.*, 2010] Daniel Roggen, Alberto Calatroni, Mirco Rossi, Thomas Holleczek, Kilian Förster, Gerhard Tröster, Paul Lukowicz, David Bannach, Gerald Pirkl, Alois Ferscha, et al. Collecting complex activity datasets in highly rich networked sensor environments. In *Proceedings of International Conference on Networked Sensing Systems*, pages 233–240. IEEE, 2010.

[Sztyler and Stuckenschmidt, 2016] Timo Sztyler and Heiner Stuckenschmidt. On-body localization of wearable devices: An investigation of position-aware activity recognition. In *Proceedings of IEEE International Conference on Pervasive Computing and Communications*, pages 1–9. IEEE, 2016.

[Wei *et al.*, 2024] Yuecen Wei, Haonan Yuan, Xingcheng Fu, Qingyun Sun, Hao Peng, Xianxian Li, and Chunming Hu. Poincaré differential privacy for hierarchy-aware graph embedding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 9160–9168, 2024.

[Wei *et al.*, 2025] Yuecen Wei, Xingcheng Fu, Lingyun Liu, Qingyun Sun, Hao Peng, and Chunming Hu. Prompt-based unifying inference attack on graph neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 12836–12844, 2025.

[Wu *et al.*, 2023] Haixu Wu, Tengge Hu, Yong Liu, Hang Zhou, Jianmin Wang, and Mingsheng Long. Timesnet: Temporal 2d-variation modeling for general time series analysis. In *Proceedings of International Conference on Learning Representations*, 2023.

[Wu *et al.*, 2024] Zongqian Wu, Yujing Liu, Mengmeng Zhan, Ping Hu, and Xiaofeng Zhu. Adaptive multi-modality prompt learning. In *Proceedings of the ACM International Conference on Multimedia*, pages 8672–8680, 2024.

[Zhang *et al.*, 2023] Guixian Zhang, Debo Cheng, and Shichao Zhang. Fpgnn: Fair path graph neural network for mitigating discrimination. *World Wide Web*, 26(5):3119–3136, 2023.

[Zhang *et al.*, 2025a] Guixian Zhang, Guan Yuan, Debo Cheng, Lin Liu, Jiuyong Li, and Shichao Zhang. Disentangled contrastive learning for fair graph representations. *Neural Networks*, 181(1):106781, 2025.

[Zhang *et al.*, 2025b] Xuebin Zhang, Qicheng Xu, Fuyuan Feng, Xiaochen Lu, and Longting Xu. Fall-mamba: A multimodal fusion and masked mamba-based approach for fall detection. *IEEE Internet of Things Journal*, 12(8):10493–10505, 2025.

[Zhou *et al.*, 2021] Haoyi Zhou, Shanghang Zhang, Jieqi Peng, Shuai Zhang, Jianxin Li, Hui Xiong, and Wancai Zhang. Informer: Beyond efficient transformer for long sequence time-series forecasting. In *Proceedings of the AAAI conference on Artificial Intelligence*, pages 11106–11115, 2021.

[Zhou *et al.*, 2024] Yuxuan Zhou, Xudong Yan, Zhi-Qi Cheng, Yan Yan, Qi Dai, and Xian-Sheng Hua. Blockgcn: Redefine topology awareness for skeleton-based action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2049–2058, 2024.