MESEN: Exploit Multimodal Data to Design Unimodal Human Activity Recognition with Few Labels

Lilin Xu[1], Chaojie Gu[1,*], Rui Tan[2], Shibo He[1], Jiming Chen[1]
[1]Zhejiang University, [2]Nanyang Technological University
Email:{lilinxu,gucj,s18he,cjm}@zju.edu.cn,tanrui@ntu.edu.sg

ABSTRACT

Human activity recognition (HAR) will be an essential function of various emerging applications. However, HAR typically encounters challenges related to modality limitations and label scarcity, leading to an application gap between current solutions and real-world requirements. In this work, we propose MESEN, a multimodal-empowered unimodal sensing framework, to utilize unlabeled multimodal data available during the HAR model design phase for unimodal HAR enhancement during the deployment phase. From a study on the impact of supervised multimodal fusion on unimodal feature extraction, MESEN is designed to feature a multi-task mechanism during the multimodal-aided pre-training stage. With the proposed mechanism integrating cross-modal feature contrastive learning and multimodal pseudo-classification aligning, MESEN exploits unlabeled multimodal data to extract effective unimodal features for each modality. Subsequently, MESEN can adapt to downstream unimodal HAR with only a few labeled samples. Extensive experiments on eight public multimodal datasets demonstrate that MESEN achieves significant performance improvements over state-of-the-art baselines in enhancing unimodal HAR by exploiting multimodal data.

CCS CONCEPTS

• **Human-centered computing → Ubiquitous and mobile computing systems and tools**; • **Computing methodologies → Machine learning**.

KEYWORDS

Mobile sensing, human activity recognition, self-supervised learning, contrastive learning

ACM Reference Format:
Lilin Xu, Chaojie Gu, Rui Tan, Shibo He, Jiming Chen. 2023. MESEN: Exploit Multimodal Data to Design Unimodal Human Activity Recognition with Few Labels. In *The 21st ACM Conference on Embedded Networked Sensor Systems (SenSys '23), November 12–17, 2023, Istanbul, Turkiye.* ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3625687.3625782

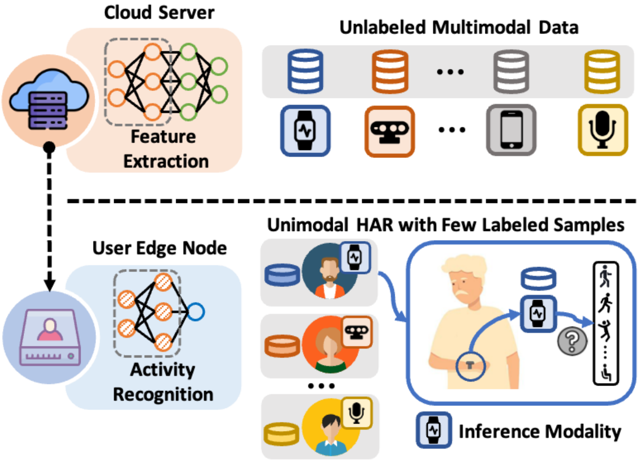


Figure 1: The application scenario of MESEN. Multimodal data are available on the server for HAR model design, while the user at the edge deploys unimodal HAR with few labels.

1 INTRODUCTION

In recent years, human activity recognition (HAR) regains research attention due to the increasing applications and the potential performance leaps enabled by deep learning. In particular, the recent advances in multimodal encoding in general artificial intelligence tasks, such as ImageBind [13], have implied that exploiting multimodal data for building HAR models is promising. The prior studies [9, 21, 22, 26] have shown the HAR performance improvements brought by multimodal data fusion.

However, HAR in real-world scenarios still faces practical challenges and application gaps in exploiting multimodal data. On the one hand, high annotation costs and label scarcity issues are widely present in practical HAR applications, resulting in scenarios where only a few labeled samples are available. Manually annotating data is tedious and time-consuming. This issue becomes more severe in multimodal sensing, since annotating multimodal data requires correlating data across varied modalities and possessing knowledge of multiple modalities. In comparison, unlabeled data are readily available and easier to access. Potential solutions utilizing these easy-to-access unlabeled data can further boost data availability for HAR applications. On the other hand, despite the growing prominence of multimodal research and deployment, unimodal HAR remains the most typical application paradigm. In reality, many HAR applications are still deployed using a single modality. Even in scenarios where multiple sensors are available, there may still be requirements for unimodal HAR. For instance, in smart home

[*] Chaojie Gu is the corresponding author.

SenSys '23, November 12–17, 2023, Istanbul, Turkiye

ACM ISBN 979-8-4007-0414-7/23/11…$15.00
https://doi.org/10.1145/3625687.3625782

scenarios where mmWave radar sensors are deployed and smartwatches' built-in IMU sensors are available, users may not always be within the sensing range of the radar, resulting in situations of modality limitation. In these scenarios, although multimodal data can be collected, unimodal HAR remains accessible and important for users. Thus, to achieve universal performance enhancement for HAR applications, it is important and meaningful to investigate the benefits of increasingly available multimodal data during the HAR model design phase on unimodal HAR during the deployment phase.

To this end, we aim to address a universal application situation as depicted in Figure 1, where the server has multimodal data for HAR model design while the user at the edge obtains a unimodal HAR model with few labeled samples. This scenario raises an essential question that remains to be studied: ***how can we effectively exploit unlabeled multimodal data to improve the performance of unimodal HAR with few labels?***

To answer the above question, we design MESEN, a multimodal-empowered unimodal sensing framework, to exploit multimodal data for designing unimodal HAR with few labels. In this way, increasingly available unlabeled multimodal data can be exploited for effective unimodal feature extraction, thereby achieving universal enhancement for unimodal HAR. From a study on the supervised multimodal fusion's effects on unimodal feature extraction, we observe that the correlations within temporally aligned multimodal samples and the distinct intra-modality spaces across different modalities are both vital for activity recognition. Besides, we investigate the relationships between unimodal predicted probabilities and final fusion results. In light of these observations, MESEN is designed to feature a multi-task mechanism during the multimodal-aided pre-training stage. By integrating cross-modal feature contrastive learning and multimodal pseudo-classification aligning, the mechanism exploits the correlations and relationships within unlabeled multimodal data, not only in the feature extraction stage's representation space but also in the pseudo-classification stage's representation space. Equipped with effective unimodal features extracted during pre-training, MESEN then can adapt to downstream unimodal HAR with only a few labeled samples through fine-tuning. This framework respects the single-modality constraint while effectively utilizing available unlabeled multimodal data.

We evaluate the performance of MESEN on eight multimodal datasets that encompass a range of modalities (accelerometer, gyroscope, magnetometer, skeleton points, depth images, and mmWave radar), user scales, and human activities. Our evaluation indicates that MESEN achieves significant performance improvements on all datasets, yielding an average increase of 30.7% accuracy and 34.5% F1-score over supervised unimodal learning and at least an average increase of 25.2% accuracy and 26.4% F1-score over the contrastive learning baselines.

Our key contributions are summarized as follows:

- By examining the performance of supervised multimodal fusion, we demonstrate the effects of utilizing multimodal data during training on unimodal feature extraction. We further investigate the correlations and relationships within multimodal data in both the representation spaces of the feature extraction stage and the classification stage during multimodal training.
- We propose to utilize the increasing availability of multimodal data to enhance unimodal HAR, given the widespread applicability of unimodal HAR in real-world scenarios. MESEN[1], a multimodal-empowered unimodal sensing framework, is designed to exploit unlabeled multimodal data for effective unimodal feature extraction by integrating cross-modal feature contrastive learning and multimodal pseudo-classification aligning. With effective unimodal features, MESEN can adapt to downstream unimodal HAR with few labels.
- We extensively evaluate the performance of MESEN on eight public multimodal datasets. The results show that MESEN outperforms the state-of-the-art approaches in enhancing unimodal HAR performance by exploiting unlabeled multimodal data.

The rest of this paper is organized as follows. §2 reviews related studies. §3 presents the measurement study and motivation. §4 introduces the detailed design of MESEN. §5 presents evaluation results. §6 discusses some related issues. §7 concludes this work.

2 RELATED WORK

In this section, we overview the research related to our work, demonstrating the research gaps that we aims to address.

HAR is empowered by various modality sources, facilitating a wide range of applications including health monitoring [7, 9, 15], daily routine monitoring [3, 12], and smart gym [28, 29]. However, real-world HAR often encounters the challenge of limited labeled data, primarily due to the high costs and labor intensity associated with the data labeling process, especially for the modalities beyond RGB videos which are in general non-interpretable by people. Consequently, the study of HAR with limited labeled data has garnered significant research interest in recent years.

Generative methods [5, 44] craft data or labels from existing data to mitigate the issue of label scarcity. SenseGAN [44] utilizes limited labeled supervision and abundant unlabeled data. It features a generator producing sensing data with random labels, a classifier producing labels for unlabeled data, and a discriminator discriminating real labeled samples and partially generated samples. HMGAN [5], as a data augmentation technique, employs multiple generators to generate multimodal data from limited labeled data to enlarge the training set.

Besides, different from generative methods, there are studies on employing self-supervised learning techniques to exploit available unlabeled data directly to their advantage. Depending on the modalities involved in the model design and deployment phases, existing studies can be divided into the categories of single-modality [31, 33] and multi-modality [10, 14, 23, 26, 42].

IMU-based methods [14, 23, 31, 42] have been extensively studied in recent years. While TPN [31] is designed to deal with unimodal IMU data by recognizing eight different data transformations applied to the raw accelerometer signal, the approaches in [14, 23, 42] are capable of processing data from multiple IMU sensors by performing the data-level fusion among modalities with a relatively small gap. Multi-task deep cluster [23] employs a framework with three tasks trained iteratively to obtain effective representation. CPCHAR [14] utilizes the Contrastive Predictive Coding (CPC) framework to capture the temporal structure of the accelerometer

[1] https://github.com/initxu/MESEN/

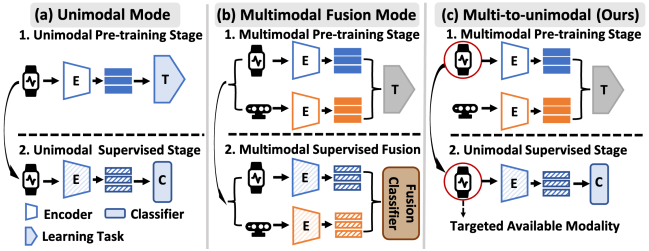

(a) Unimodal Mode
1. Unimodal Pre-training Stage
E
T
2. Unimodal Supervised Stage
E
C
Encoder
Classifier
Learning Task
(b) Multimodal Fusion Mode
1. Multimodal Pre-training Stage
E
E
T
2. Multimodal Supervised Fusion
E
E
Fusion Classifier
(c) Multi-to-unimodal (Ours)
1. Multimodal Pre-training Stage
E
E
T
2. Unimodal Supervised Stage
E
C
Targeted Available Modality

Figure 2: (a) & (b): Prior works [14, 26, 33] designed with label scarcity include the unimodal mode and the multimodal fusion mode. (c): MESEN operates in a multi-to-unimodal mode to improve unimodal HAR performance by exploiting unlabeled multimodal data.

and gyroscope data. LIMUBert [42] proposes an adaptive BERT-like self-supervised task to extract generalizable features from unlabeled IMU sensor data. RadarAE [33] adapts the idea of Masked Autoencoders (MAE) to radar sensing with unlabeled radar data. These approaches are specifically designed for a single modality or modalities with similar natures, which prevents them from being easily ported to other modalities or handling heterogeneous multimodal data. We propose a universal framework that is capable of handling various modalities without any extra changes.

With the development of contrastive learning and its excellent representation learning performance in multiple fields, such as computer vision [6, 37] and natural language [11, 40], this effective idea is introduced to HAR to deal with unlabeled multimodal data. COCOA [10] performs contrastive learning between features extracted from multisensor data for fusion. Cosmo [26] designs a multimodal feature fusion contrastive method to fully use multimodal synergies within heterogeneous multimodal data. These solutions focus on improving the performance of multimodal HAR for scenarios where multimodal data are available. Thus, they cannot be directly applied to our target unimodal HAR application scenarios where only a single modality is available during the deployment phase.

Different from the above single-modality or multi-modality methods, our method, MESEN, aims to utilize increasingly available unlabeled multimodal data to achieve universal enhancement for unimodal HAR. Thus, the proposed framework operates in a multi-to-unimodal mode as shown in Figure 2 (c). Although similar two-stage training frameworks (Figure 2 (a) & (b)) are used in previous works [14, 26, 33] to address label scarcity issues for unimodal or multimodal HAR, MESEN significantly differs from them due to the specific application scenarios involving modality limitations during the HAR model design and the deployment phases.

3 MOTIVATION

In this section, we conduct a measurement study to understand modality differences and investigate the impact of multimodal fusion on unimodal feature extraction. The observations motivate the design of MESEN, a framework that addresses practical multi-to-unimodal HAR application scenarios with few labels.

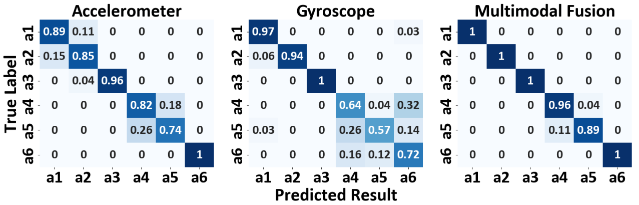


Figure 3: Unimodal and multimodal recognition results on the UCI dataset. Activities from $a1$ to $a3$ are walking-related while the rest are stationary activities.

3.1 Measurement Study

Different modalities capture different aspects of a process. For instance, accelerometer measures linear acceleration; gyroscope measures angular velocity and is thus sensitive to changes in orientation. As a result, they are sensitive to different types of human activities. We conduct experiments on the UCI dataset [30] comprising accelerometer and gyroscope data for six activities, with four users' data for training and the rest for validation and testing. We use 1D-CNN [36] as modality encoders and a single linear layer as classifier heads. Figure 3 shows that when employed individually, each modality yields better performance on certain types of activities. Gyroscope demonstrates better recognition performance for dynamic activities compared with stationary ones, due to its sensitivity to orientation changes and angular velocity. Accelerometer outperforms gyroscope in recognizing stationary activities, due to its ability to measure static forces like gravity.

However, the gyroscope's relatively worse performance in identifying stationary activities does not mean it lacks relevant information for distinguishing these activities. Instead, the relevant valuable information and features are present but not effectively captured during the training. Indeed, the results show that gyroscope data can still recognize a part of these activities effectively.

Actually, the ***free yet accessible*** features, which are effective for recognition, can be unearthed when another modality is present during training, as in multimodal fusion. Multimodal fusion has been considered in HAR [2, 17, 22, 43] and has shown notable performance improvements when the multimodal features or the predicted probabilities are properly fused. We investigate the impact of the assistance provided by an additional modality during training by studying unimodal performance under supervised multimodal fusion training. We utilize a score-level fusion method [17], which applies a weighted fusion on unimodal predicted probabilities obtained from each modality. On the one hand, as Figure 3 shows, multimodal fusion improves the recognition performance. On the other hand, the fusion also aids in extracting effective unimodal features that might otherwise remain undetected during unimodal training. With t-SNE visualization [38], we analyze the gyroscope features extracted by the modality encoder under three conditions: without training, post unimodal training, and during multimodal fusion. Figure 4 shows that when accelerometer data are present during training, the gyroscope features form more distinct clusters compared with other conditions. This suggests that effective unimodal features can be better extracted with the aid of another modality under supervised multimodal fusion.

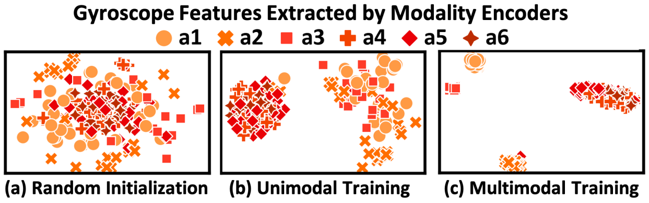


Figure 4: The visualization of extracted gyroscope features under three conditions.

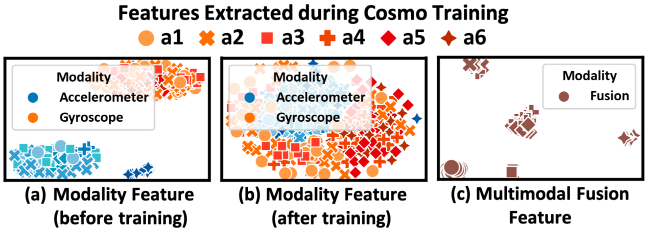


Figure 5: The unimodal features extracted by Cosmo are beneficial to subsequent multimodal fusion instead of unimodal recognition.

3.2 Problem Statement

Based on the experiments, §3.1 gives insights into the effectiveness of unimodal feature extraction achieved during supervised multimodal fusion training. With these insights, we aim to apply such multimodal assistance for unimodal HAR with unlabeled multimodal data.

The scenarios we focus on have two major differences from supervised multimodal HAR, i.e., exploiting unlabeled multimodal data and enhancing unimodal HAR. Firstly, real-world scenarios often face the challenge of lacking annotations, but unlabeled temporally aligned multimodal data are easily accessible. By utilizing unlabeled multimodal data, the availability of data can be significantly increased, thus avoiding the constraints imposed by reliance on labeled data. Secondly, driven by the ubiquity of unimodal HAR application scenarios, we focus on utilizing unimodal data for activity recognition during the deployment phase rather than applying multimodal HAR. Our aim is to improve the performance of unimodal HAR with available multimodal data.

The most related work is Cosmo [26], which is capable of handling unlabeled multimodal data for HAR. However, it cannot be directly applied to our target application scenarios as it is designed to fuse multimodal data for scenarios where multimodal data are available during the deployment phase of HAR applications. Figure 5 shows that the unimodal features extracted during Cosmo's pre-training stage do not exhibit the same clear clustering characteristic as the fusion features. This indicates that Cosmo is designed to utilize unlabeled multimodal data for extracting features that are beneficial to subsequent multimodal fusion. In contrast, our target scenarios require directly extracting effective unimodal features for downstream unimodal HAR. Consequently, we design MESEN to exploit unlabeled multimodal data and achieve effective unimodal feature extraction with multimodal assistance, thereby enhancing unimodal HAR with few labels.

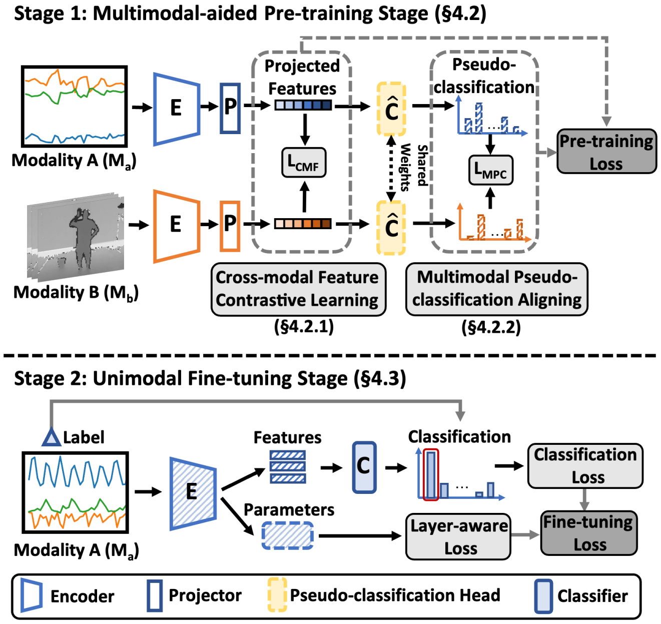


Figure 6: The overview of MESEN.

4 MESEN DESIGN

In this section, we indicate the specific application scenarios that our work focuses on. Subsequently, we demonstrate the design of MESEN in detail.

4.1 Overview

■ **Problem formulation:** The unlabeled multimodal data with N modalities available are defined as $\mathcal{D}_m = \{\mathcal{D}^{M_1}, \ldots, \mathcal{D}^{M_N}\}$, where $\mathcal{D}^{M_i} = \{x_1^{M_i}, \ldots, x_{n_{M_i}}^{M_i}\}$ and n_{M_i} denotes the sample number. $x_k^{M_i}$ and $x_k^{M_j}$ denote temporally aligned paired samples recording the same activity process. The unimodal data during the deployment phase are defined as $\mathcal{D}_u = \{\mathcal{D}^{M_a}\}$, where $\mathcal{D}^{M_a} = \{\hat{x}_1^{M_a}, \ldots, \hat{x}_{m_{M_a}}^{M_a}, x_1^{M_a}, \ldots, x_{n_{M_a}}^{M_a}\}$. It involves labeled data $\hat{x}^{M_a}$ with a small value of m_{M_a} of the available modality $M_a \in \{M_1, \ldots, M_N\}$. MESEN is designed to exploit multimodal data from $\mathcal{D}m$ to enhance the unimodal HAR performance on $\mathcal{D}u$.

■ **Framework overview:** Figure 6 demonstrates the oveview design of MESEN. For clarity, we use two modalities M_a and M_b as the illustration in the figure and the subsequent design description. MESEN comprises the multimodal-aided pre-training stage and the unimodal fine-tuning stage. During pre-training, MESEN aims to train modality encoders for effective unimodal feature extraction, utilizing unlabeled multimodal data and modality relationships. During fine-tuning, the encoder of the available modality is fine-tuned with a few labeled samples and then used for unimodal HAR.

4.2 Multimodal-aided Pre-training Stage

As mentioned in §3, supervised multimodal fusion can aid unimodal feature extraction. To achieve this assistance with unlabeled multimodal data, we further investigate the correlations and relationships within multimodal data in both the representation spaces

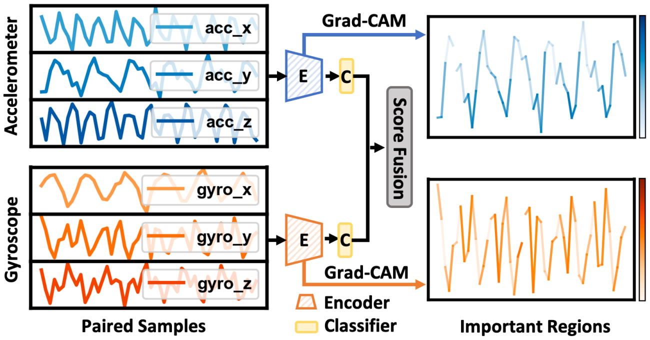


Figure 7: Temporally aligned paired samples of two modalities, recording the same activity (*Walking*), exhibit similar variation patterns and correlated important regions. The color intensity indicates the importance of the region.

of the feature extraction stage and the classification stage. Based on the insights observed from supervised multimodal fusion, we develop a multi-task mechanism, integrating cross-modal feature contrastive learning and multimodal pseudo-classification aligning. With the mechanism, unimodal features can be effectively extracted from unlabeled multimodal data for subsequent unimodal HAR.

4.2.1 Cross-modal Feature Contrastive Learning. Paired multimodal data are records of the same process, suggesting their inherent correlations, while they capture different aspects of the process due to their distinct physical properties. Thus, we investigate the representation space of the feature extraction stage for both the correlations and differences between modalities. We observe there are inter-modality correlations in paired multimodal data (as depicted in Figure 7) and distinct intra-modality spaces reflecting the differences between modalities (as depicted in Figure 8). Based on the observations, we employ a cross-modal feature contrastive learning method rather than applying contrastive learning directly. On the one hand, this method emphasizes the similarity between paired multimodal features to capture inter-modality correlations. On the other hand, it maintains the modality differences by excluding consideration of dissimilarity within intra-modality when maximizing the dissimilarity between non-paired multimodal features.

■ **Inter-modality correlation:** Multiple modalities can acquire crucial correlated information when recording the same activity process. For instance, the left portion of Figure 7 shows temporally aligned paired samples of two modalities (accelerometer and gyroscope), which record the same *Walking* activity process. The paired samples exhibit similar variation patterns due to the rhythmic nature of walking. These correlated patterns play a vital role in recognition. To highlight significant regions in the input sensor data that contribute to the final prediction in multimodal fusion, we apply Gradient-weighted Class Activation Mapping (Grad-CAM) [32] to the last convolutional layer of each modality encoder individually after supervised multimodal fusion training. As shown in the right portion of Figure 7, there is a correlation between the important regions across modalities. Actually, such correlated information is ubiquitously present in paired multimodal data. On the UCI dataset, we conduct canonical correlation analysis on data from the same participant. The results show an average correlation of 0.600 between paired multimodal samples and an average correlation of 0.313 between non-paired samples, indicating that paired samples hold more correlated information.

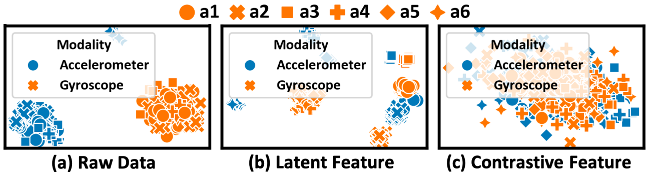


Figure 8: (a) & (b): Distinct modality spaces of different modality data are clear both at the raw data level and the latent feature representation level. (c): modality spaces are distorted in contrastive learning.

We aim to utilize the correlated information in multimodal data to enhance unimodal feature extraction. An effective strategy to capture the correlations involves maximizing the similarity of paired multimodal features while maximizing the dissimilarity of non-paired features.

■ **Intra-modality space:** Different modalities have distinct physical properties. While non-paired features in multimodal scenarios encompass both inter-modality and intra-modality types, it is inappropriate to maximize the dissimilarity of both types in our task due to the modality differences. Figure 8 (a) demonstrates distinct intra-modality spaces reflecting the physical differences between modalities. Furthermore, Figure 8 (b) shows that at the level of latent feature representation in supervised multimodal fusion, features from the same modality tend to cluster together, forming distinct modality-specific spaces for different modalities. This indicates that the intra-modality spaces related to modality properties are crucial for activity recognition. Therefore, it is beneficial to maintain distinct intra-modality spaces during pre-training. However, as shown in Figure 8 (c), maximizing the dissimilarity of both inter-modality and intra-modality non-paired features during contrastive learning can lead to distortion of the distinct intra-modality feature spaces. This is due to the neglect of the distinct properties of different modalities and treating them equally in contrastive learning.

To prevent this kind of distortion and maintain modality differences, we only maximize the dissimilarity of inter-modality non-paired features without maximizing the dissimilarity between intra-modality features.

■ **Design for correlation capturing & difference maintaining:** Given the observations above, we design cross-modal feature contrastive learning to capture inter-modality correlations and maintain distinct intra-modality spaces during training.

As shown in Figure 6, $\mathbf{x^a}$ and $\mathbf{x^b}$ denote temporally aligned paired samples from modalities M_a and M_b. These samples are processed through respective modality encoders $\boldsymbol{f}(\cdot)$ and modality projectors $\boldsymbol{g}(\cdot)$, yielding features $\mathbf{z^a} \in \mathbb{R}^{N_{fc}}$ and $\mathbf{z^b} \in \mathbb{R}^{N_{fc}}$, which can be expressed by

$$\begin{aligned} \mathbf{h^a} &= \boldsymbol{f}^{M_a}(\mathbf{x^a}), \hat{\mathbf{h}}^\mathbf{a} = \boldsymbol{g}^{M_a}(\mathbf{h^a}), \mathbf{z^a} = \boldsymbol{M_{Norm}}(\hat{\mathbf{h}}^\mathbf{a}); \\ \mathbf{h^b} &= \boldsymbol{f}^{M_b}(\mathbf{x^b}), \hat{\mathbf{h}}^\mathbf{b} = \boldsymbol{g}^{M_b}(\mathbf{h^b}), \mathbf{z^b} = \boldsymbol{M_{Norm}}(\hat{\mathbf{h}}^\mathbf{b}), \end{aligned} \tag{1}$$

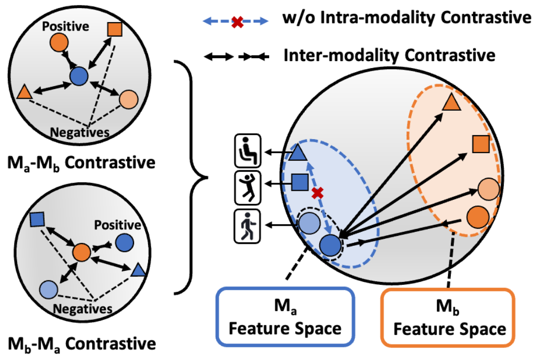


Figure 9: The design of cross-modal feature contrastive learning.

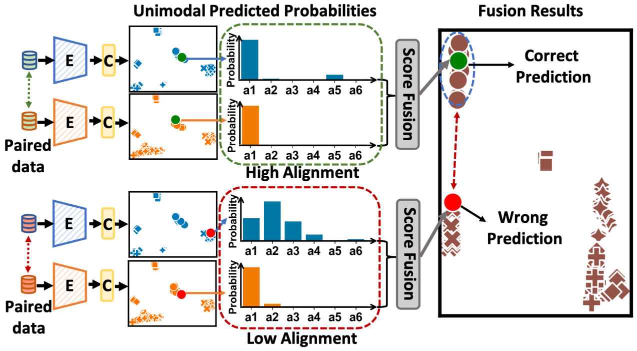


Figure 10: The effects of alignment between paired unimodal predicted probabilities on final fusion results.

where the modality features $\hat{\mathbf{h}}$ are mapped and normalized within the batch by $\boldsymbol{M_{Norm}}(\cdot)$ to $\mathbf{z}$ for contrastive learning. During multimodal pre-training, every input pairs $(\mathbf{x^a}_i, \mathbf{x^b}_i)$ produces a pair of feature vectors $(\mathbf{z^a}_i, \mathbf{z^b}_i)$. Across N input pairs in the mini-batch $\mathcal{B}$, we get $\mathcal{B}_z^a = \{\mathbf{z^a}_1, \ldots, \mathbf{z^a}_N\}$ and $\mathcal{B}_z^b = \{\mathbf{z^b}_1, \ldots, \mathbf{z^b}_N\}$ for the computation of the contrastive loss.

Unlike the single-modality contrastive loss [6], which constructs one positive pair and $2N-2$ negative pairs for each sample within a mini-batch, we obtain one inter-modality positive pair, $N-1$ inter-modality negative pairs and $N-1$ intra-modality negative pairs. Specifically, for the sample $\mathbf{x^a}_i$, as shown in Figure 9, we compute the contrastive loss based on the positive set $\mathbf{P_i^a} = \{\mathbf{z^b}_i\}$ and the inter-modality negative set $\mathbf{N_i^{(a\to b)}} = \{\mathbf{z^b}_j | \mathbf{z^b}_j \in \mathcal{B}_z^b, j \neq i\}$, while ignoring the intra-modality negative set $\mathbf{N_i^{(a\to a)}} = \{\mathbf{z^a}_j | \mathbf{z^a}_j \in \mathcal{B}_z^a, j \neq i\}$. Consequently, the M_a-to-M_b contrastive loss $L_i^{(a\to b)}$ for the sample $\mathbf{x^a}_i$ can be expressed by

$$\mathrm{L}_i^{(a\to b)} = -\log \frac{\exp(\mathbf{z^a}_i \cdot \mathbf{z^b}_i)/\tau}{\exp(\mathbf{z^a}_i \cdot \mathbf{z^b}_i)/\tau + \sum_{\mathbf{z^b}_j \in \mathrm{N_i^{(a\to b)}}} \exp(\mathbf{z^a}_i \cdot \mathbf{z^b}_j)/\tau}, \tag{2}$$

where $\tau \in \mathbb{R}^+$ denotes the temperature parameter. Indeed, Eq. 2 can be interpreted as the loss of seeking to correctly identify $(\mathbf{z^a}_i, \mathbf{z^b}_i)$ as the temporally aligned pair, while treating the rest as negatives. Additionally, we can obtain the M_b-to-M_a contrastive loss $L_i^{(b\to a)}$ for the sample $\mathbf{x^b}_i$ in the same way, it can be expressed by

$$\mathrm{L}_i^{(b\to a)} = -\log \frac{\exp(\mathbf{z^b}_i \cdot \mathbf{z^a}_i)/\tau}{\exp(\mathbf{z^b}_i \cdot \mathbf{z^a}_i)/\tau + \sum_{\mathbf{z^a}_j \in \mathrm{N_i^{(b\to a)}}} \exp(\mathbf{z^b}_i \cdot \mathbf{z^a}_j)/\tau}. \tag{3}$$

The cross-modal feature contrastive loss L_{CMF} for the mini-batch $\mathcal{B}$ of N paired samples can be expressed as

$$L_{\text{CMF}} = \frac{1}{N} \sum_{i=1}^{N} (\alpha L_i^{(a\to b)} + \beta L_i^{(b\to a)}), \tag{4}$$

where $\alpha > 0$ and $\beta > 0$ are weights measuring the importance of different modalities contributing to L_{CMF}. To adapt to various modality combinations, we assign equal weights to different modalities by setting $\alpha = \beta = 0.5$.

4.2.2 *Multimodal pseudo-classification aligning.*

After utilizing the multimodal correlations in the representation space of the feature extraction stage, there is still a lack of a pre-training task that is directly related to the downstream recognition task in the design. Besides, the multimodal correlations existing in the representation space of the classification stage remain unexploited. Thus, we employ multimodal pseudo-classification aligning. On the one hand, the pseudo-classification task can be a prompt for the downstream recognition task. On the other hand, it can utilize relationships within multimodal predicted probabilities in the classification stage's representation space.

■ **Task prompt:** Introducing a pseudo-classification task into the pre-training stage provides a beneficial prompt for downstream recognition. Given that activity recognition is essentially a classification task, employing a pseudo-classification task during pre-training can benefit feature extraction, and enhance the effectiveness of the fine-tuning process even if there are only a few labeled samples available.

■ **Multimodal predicted probabilities alignment:** The alignment measures the degree of similarity between the unimodal prediction probabilities of paired samples. As discussed in §3, the fusion of unimodal predicted probabilities can guide unimodal feature extraction. To utilize pseudo-classification results without the fusion step, we investigate the relationships between the multimodal predicted probabilities alignment and the final fusion prediction results under supervised multimodal fusion.

During supervised multimodal fusion, unimodal predicted probabilities ($\mathbf{y^a}$ and $\mathbf{y^b}$) are individually mapped from modality features through the classifier head. The fusion result $\mathbf{y}$ is obtained by combining $\mathbf{y^a}$ and $\mathbf{y^b}$. Our observations show that for $\mathbf{y}$ that is correctly classified (as depicted at the middle top portion in Figure 10), $\mathbf{y^a}$ and $\mathbf{y^b}$ exhibit a high degree of alignment. Conversely, for the misclassified fusion result, $\mathbf{y^a}$ and $\mathbf{y^b}$ vary significantly from each other (as depicted at the middle bottom portion in Figure 10). We measure the Euler distance between $\mathbf{y^a}$ and $\mathbf{y^b}$ on the UCI dataset, revealing that the average distance associated with misclassified fusion results is 0.842, which is 2.38 times the average distance of 0.354 related to correctly classified results. If the unimodal predicted probabilities of paired samples are significantly different, it suggests that at least one of the results provides a less reliable classification prediction, which adversely affects the final fusion result.

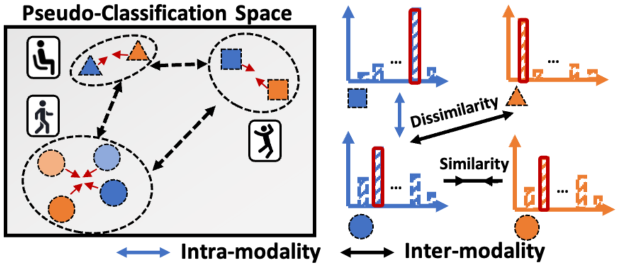
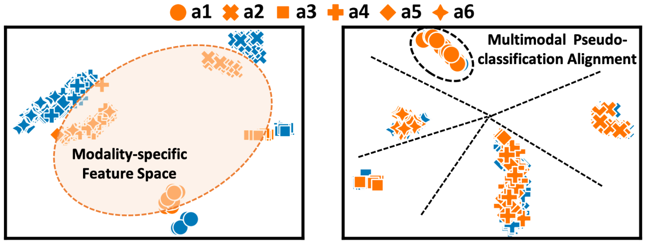


Figure 11: The design of multimodal pseudo-classification aligning.



(a) Contrastive Feature Space **(b) Pseudo-classification Space**

Figure 12: The performance of MESEN design.

Based on the above observations, as shown in Figure 11, our objective is to ensure that the pseudo-predicted probabilities ($\hat{\mathbf{y}}_\mathbf{i}^\mathbf{a}$, $\hat{\mathbf{y}}_\mathbf{i}^\mathbf{b}$) obtained from paired multimodal samples ($\mathbf{x}_\mathbf{i}^\mathbf{a}$, $\mathbf{x}_\mathbf{i}^\mathbf{b}$) are classified into the same pseudo-class, while those derived from non-paired samples should be classified into different pseudo-classes.

■ **Design for pseudo-classification aligning:** Given the insights above, we employ multimodal pseudo-classification aligning. It first implements pseudo-classification on modality features and then utilizes the relationships within multimodal pseudo-class probability probabilities.

Firstly, we apply a N_{cls}-way classification by using the pseudo-classification head $\hat{\mathbf{c}}(\cdot)$, where N_{cls} is the number of activity categories which can be easily obtained with no extra effort. The pseudo-predicted probabilities $\hat{\mathbf{y}}^\mathbf{a} = \hat{\mathbf{c}}(\hat{\mathbf{h}}^\mathbf{a})$ and $\hat{\mathbf{y}}^\mathbf{b} = \hat{\mathbf{c}}(\hat{\mathbf{h}}^\mathbf{b})$ are obtained from the modality features $\hat{\mathbf{h}}^\mathbf{a}$ and $\hat{\mathbf{h}}^\mathbf{b}$ individually. Then, we directly utilize $\hat{\mathbf{y}}^\mathbf{a}$ and $\hat{\mathbf{y}}^\mathbf{b}$ instead of combining them as in supervised multimodal fusion.

Specifically, within the mini-batch $\mathcal{B}$, we obtain results of pseudo-classification $\hat{\mathbf{Y}}^\mathbf{a}$ and $\hat{\mathbf{Y}}^\mathbf{b}$, both of which belong to the space $\mathbb{R}^{N \times N_{cls}}$ and can be expressed as

$$\hat{\mathbf{Y}}^\mathbf{a} = \begin{bmatrix} \hat{\mathbf{y}}_\mathbf{1}^\mathbf{a} \\ \cdots \\ \hat{\mathbf{y}}_\mathbf{N}^\mathbf{a} \end{bmatrix} = \begin{bmatrix} \hat{\mathbf{q}}_\mathbf{1}^\mathbf{a} \cdots \hat{\mathbf{q}}_{\mathbf{N_{cls}}}^\mathbf{a} \end{bmatrix}, \hat{\mathbf{Y}}^\mathbf{b} = \begin{bmatrix} \hat{\mathbf{y}}_\mathbf{1}^\mathbf{b} \\ \cdots \\ \hat{\mathbf{y}}_\mathbf{N}^\mathbf{b} \end{bmatrix} = \begin{bmatrix} \hat{\mathbf{q}}_\mathbf{1}^\mathbf{b} \cdots \hat{\mathbf{q}}_{\mathbf{N_{cls}}}^\mathbf{b} \end{bmatrix}. \tag{5}$$

The i-th element in the vector $\hat{\mathbf{y}}$ represents the probability that $\mathbf{x}$ is the i-th pseudo-class. The i-th column $\hat{\mathbf{q}}_\mathbf{i}$ in the matrix $\hat{\mathbf{Y}}$ represents the i-th pseudo-class, mapping that which samples in $\mathcal{B}$ are classified into the i-th pseudo-class. Ideally, paired multimodal samples $\mathbf{x}^\mathbf{a}$ and $\mathbf{x}^\mathbf{b}$ should be classified into the same pseudo-class. As a result, $\hat{\mathbf{q}}_\mathbf{i}^\mathbf{a}$ and $\hat{\mathbf{q}}_\mathbf{i}^\mathbf{b}$, both denoting the i-th pseudo-class results, should be as similar as possible. In contrast, the dissimilarity between the i-th pseudo-class $\hat{\mathbf{q}}_\mathbf{i}$ and any other j-th pseudo-class $\hat{\mathbf{q}}_\mathbf{j}$ (including $\hat{\mathbf{q}}_\mathbf{j}^\mathbf{a}$ and $\hat{\mathbf{q}}_\mathbf{j}^\mathbf{b}$) should be maximized. Consequently, different from single-modality clustering in [20], for the i-th pseudo-class result $\hat{\mathbf{q}}_\mathbf{i}^\mathbf{a}$, we obtain the positive pseudo-class set $\hat{\mathbf{P}}_\mathbf{i}^\mathbf{a} = \{\hat{\mathbf{q}}_\mathbf{i}^\mathbf{b}\}$, and the negative pseudo-class set $\hat{\mathbf{N}}_\mathbf{i}^\mathbf{a} = \{\hat{\mathbf{q}}_\mathbf{j}^\mathbf{a}, \hat{\mathbf{q}}_\mathbf{j}^\mathbf{b} | \hat{\mathbf{q}}_j^\mathbf{a}, \hat{\mathbf{q}}_j^\mathbf{b} \in \mathcal{B}_{\hat{q}}, j \neq i\}$ containing $2N_{cls} - 2$ different classes from both intra-modality and inter-modality. Thus, the pseudo-classification aligning loss for $\hat{\mathbf{q}}_\mathbf{i}^\mathbf{a}$ can be expressed as

$$\hat{L}_i^{(a)} = -\log \frac{\exp(\hat{\mathbf{q}}_\mathbf{i}^\mathbf{a} \cdot \hat{\mathbf{q}}_\mathbf{i}^\mathbf{b})/\hat{\tau}}{\exp(\hat{\mathbf{q}}_\mathbf{i}^\mathbf{a} \cdot \hat{\mathbf{q}}_\mathbf{i}^\mathbf{b})/\hat{\tau} + \sum_{\hat{\mathbf{q}}_\mathbf{j} \in \hat{\mathbf{N}}_\mathbf{i}^\mathbf{a}} \exp(\hat{\mathbf{q}}_\mathbf{i}^\mathbf{a} \cdot \hat{\mathbf{q}}_\mathbf{j})/\hat{\tau}}. \tag{6}$$

Classification labels are applied across all modalities in an equal way. Thus, different from the cross-modal feature contrastive loss, $\hat{L}_i^{(a)}$ and $\hat{L}_i^{(b)}$ are symmetric. The multimodal pseudo-classification aligning loss L_{MPC} for the mini-batch $\mathcal{B}$ of N input paired samples with N_{cls}-way pseudo-classification can be expressed as

$$L_{\mathrm{MPC}} = \frac{1}{2N_{cls}} \sum_{i=1}^{N_{cls}} (\hat{L}_i^{(a)} + \hat{L}_i^{(b)}) + \lambda_{\mathrm{PR}} L_{\mathrm{PR}}, \tag{7}$$

where $L_{\mathrm{PR}} = -\sum_{i=1}^{N_{cls}} P(\hat{\mathbf{q}}\mathbf{i}) \log P(\hat{\mathbf{q}}\mathbf{i})$ is a Shannon entropy-based regularization loss [20] with $\lambda_{\mathrm{PR}} \in \mathbb{R}^-$ acting as a scale weight. It is employed to prevent the situation where most samples are classified into the same pseudo-class.

4.2.3 Multi-task combination. After getting L_{CMF} from cross-modal feature contrastive learning and L_{MPC} from multimodal pseudo-classification aligning, we combine them to obtain the pre-training loss L_{PT} as

$$L_{\mathrm{PT}} = L_{\mathrm{CMF}} + \delta L_{\mathrm{MPC}}, \tag{8}$$

where $\delta \in \mathbb{R}^+$ is computed by the values of L_{CMF} and L_{MPC} within each mini-batch for loss balancing.

Figure 12 demonstrates the contrastive feature space and the pseudo-classification space after MESEN's multimodal-aided pre-training stage, demonstrating its ability to maintain distinct modality spaces and ensure multimodal pseudo-classification alignment.

4.3 Unimodal Fine-tuning Stage

During fine-tuning, we obtain the pre-trained unimodal encoder according to the specific modality used in subsequent unimodal HAR, and refine it with a classifier head using labeled data. However, this process involves two potential issues, i.e., loss of knowledge from pre-training and overfitting due to label scarcity, both of which will degrade recognition performance.

To mitigate these issues, we employ a layer-aware fine-tuning mechanism with the regularization loss L_{FR}. During fine-tuning, the model consists of two parts: the pre-trained encoder which is to be fine-tuned, and the classifier head which needs to be trained from scratch. Their parameters are denoted by θ_e and θ_c, respectively. The regularization loss L_{FR} can be expressed as

$$L_{\mathrm{FR}} = \sum_{i=1}^{n_e} (\gamma_i ||\theta_{e\,i}||^2) + ||\theta_c||^2, \tag{9}$$

where i denotes different layers of the encoder and n_e is the total layer number of the encoder, γ_i controls the degree to which the learned knowledge from the pre-training stage is retained. If the equal weight $\gamma_i = 1$ is assigned to all the layers, the loss becomes $L_{\mathrm{FR}} = ||\theta_e||^2 + ||\theta_c||^2$. The unimodal fine-tuning loss L_{FT} can be expressed as

$$L_{\mathrm{FT}} = L_{\mathrm{CLS}} + \lambda_{\mathrm{FR}} L_{\mathrm{FR}}, \tag{10}$$

where L_{CLS} represents the classification loss, and λ_{FR} acts as a scale weight for the regularization term.

5 PERFORMANCE EVALUATION

5.1 Datasets and Methodology

5.1.1 Multimodal HAR Datasets. To evaluate the effectiveness of MESEN, we use eight multimodal datasets that span diverse modalities (accelerometer, gyroscope, magnetometer, skeleton points, depth images, and mmWave radar), activities, user scales, and collection environments. Table 1 provides a summary of these datasets.

UCI dataset [30]. It contains accelerometer and gyroscope data from 30 users performing 6 activities with a smartphone (Samsung Galaxy S II) on the user's waist. The sampling rate is 50 Hz.

MotionSense dataset [24]. It comprises data collected by accelerometer and gyroscope sensors with a sampling rate of 50 Hz. During data collection, a total of 24 users performed 6 activities with an iPhone 6s placed in the user's front pocket in the same environment and conditions.

HHAR dataset [35]. It contains accelerometer and gyroscope readings collected by a variety of smartphones from 9 users performing 6 daily activities. During data collection, the devices were carried by the users around their waists with the sampling rate ranging from 100 Hz to 200 Hz.

USC dataset [45]. It consists of accelerometer and gyroscope data from 14 users performing 12 activities based on their own style. During data collection, the device was placed at the user's front right hip with a reachable sampling rate of 100 Hz.

Shoaib dataset [34]. It contains magnetometer readings along with accelerometer and gyroscope data collected by Samsung Galaxy SII smartphones from 10 users performing 7 activities. The sampling rate is 50 Hz. During data collection, devices were placed in five different positions (*right pocket*, *left pocket*, *belt*, *upper arm*, and *wrist*) on users.

Cosmo-MHAD dataset [26]. It includes multimodal snippets from 30 users freely performing 14 activities. The dataset contains data of three modalities (depth, IMU, and mmWave radar) collected at the sampling rate of 20 Hz, 100 Hz, and 15 Hz, respectively. Due to the poor performance of radar when used alone as reported in [26], we use only IMU and depth data from the dataset.

mRI dataset [1]. It is a multimodal 3D human pose dataset. The released mRI includes over 5 million frames of mmWave and IMU data from 20 users with a sampling rate of 10 Hz and 50 Hz, respectively. The IMU data are from 6 sensors placed on different positions (*left wrist*, *right wrist*, *left knee*, *right knee*, *head*, and *pelvis*) of the user. Applying a 1-second window on the frame sequences, we obtained 4,105 samples of 11 movements from the dataset.

UTD dataset [4]. It contains data collected by a Microsoft Kinect sensor and a wearable inertial sensor with a sampling rate of 30 Hz and 50 Hz, respectively. We use IMU, skeleton, and depth modalities from it. During data collection, 8 subjects performed 27 different actions in an indoor environment. The inertial sensor was placed on the wrist or the thigh depending on the type of action.

Table 1: Dataset summary.

Dataset	Modality	Activity	User (train/valtest)	Sample
UCI [30]	Acc, Gyro	6	30 (24/6)	2088
MotionSense [24]	Acc, Gyro	6	24 (19/5)	4534
HHAR [35]	Acc, Gyro	6	9 (7/2)	9166
USC [45]	Acc, Gyro	12	14 (10/4)	38312
Shoaib [34]	Acc, Gyro, Mag	7	10 (8/2)	10500
Cosmo-MHAD [26]	IMU, Depth	14	30 (25/5)	3434
mRI[1]	IMU, Radar	11	20 (16/4)	4105
UTD [4]	IMU, Skeleton, Depth	27	8 (6/2)	861

For evaluation, we split users into two different subsets for training and inference (including validation and testing), respectively. The specific details of the user subsets are demonstrated in Table 1. Moreover, the data in the inference subset is evenly divided into validation and testing sets, maintaining a 1:1 ratio.

5.1.2 Baselines. MESEN is evaluated and compared with baselines covering three aspects: supervised unimodal performance, self-supervised unimodal performance, and multi-to-unimodal (*m2u*) performance which aligns with our training mode. The baselines used for comparison are as follows.

Labeltrain. It is the supervised learning method that uses labeled data of every single modality in the datasets to predict activities, serving as the supervised unimodal baseline.

SimCLR [6] and CC [20]. These two methods are state-of-the-art contrastive-based approaches in computer vision tasks, creating two distinct views of the same sample through data augmentation for conducting single-modality contrastive learning. We implement them for each modality in the multimodal datasets, serving as the self-supervised unimodal baselines.

CPCHAR [14]. It is a state-of-the-art self-supervised learning method designed for IMU data. The method utilizes the temporal structure of IMU data. We implement it for the IMU modality.

CMC (m2u) [37]. It is a leading multi-view contrastive learning method in computer vision tasks, training the modality encoders by directly contrasting multimodal features. In our experiments, considering the limitation of only unimodal data available during the deployment phase in our target scenarios, we implement CMC under the *m2u* mode, pre-training the modality encoders with multimodal data and fine-tuning each encoder using unimodal data individually as shown in Figure 2 (c).

Cosmo (m2u) [26]. It is a state-of-the-art contrastive-based multimodal HAR method. It features a feature fusion contrastive learning approach for extracting effective information from unlabeled multimodal data. We implement it under the *m2u* mode, maintaining all other parts as in [26], except for the multimodal feature fusion during fine-tuning.

These baselines can be categorized as the supervised unimodal learning baseline and the contrastive learning baselines.

5.1.3 Configurations. MESEN and other baseline models are implemented by using Python and Pytorch [27]. They are implemented

Table 2: Performance comparison. MESEN outperforms baselines on all datasets, with the relative improvements over the best-performing baselines (marked in underline) highlighted in blue.

Dataset	UCI				MotionSense				HHAR				Shoaib					
Labeling rate	0.35%				0.17%				0.08%				0.08%					
Modality	Acc		Gyro		Acc		Gyro		Acc		Gyro		Acc		Gyro		Mag	
Metrics	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1
Labeltrain	0.442	0.365	0.353	0.251	0.490	0.391	0.439	0.369	0.446	0.340	0.370	0.278	0.316	0.297	0.287	0.190	0.227	0.186
SimCLR	0.516	0.491	0.460	0.435	0.520	0.458	0.496	0.447	0.488	0.441	0.488	0.411	0.339	0.315	0.333	0.317	0.254	0.253
CC	0.415	0.425	0.444	0.427	0.301	0.283	0.463	0.370	0.449	0.368	0.435	0.351	0.263	0.250	0.360	0.340	0.214	0.212
CMC (m2u)	0.505	0.495	0.346	0.286	0.507	0.455	0.506	0.418	0.451	0.436	0.394	0.353	0.322	0.310	0.294	0.291	0.296	0.291
Cosmo (m2u)	0.513	0.494	0.440	0.379	0.506	0.467	0.493	0.411	0.489	0.436	0.493	0.411	0.315	0.313	0.298	0.297	0.318	0.309
MESEN (Ours)	**0.888** **+0.372**	**0.890** **+0.395**	**0.695** **+0.235**	**0.636** **+0.201**	**0.790** **+0.270**	**0.807** **+0.340**	**0.682** **+0.176**	**0.694** **+0.247**	**0.659** **+0.170**	**0.676** **+0.235**	**0.660** **+0.167**	**0.651** **+0.240**	**0.487** **+0.148**	**0.445** **+0.130**	**0.462** **+0.102**	**0.434** **+0.094**	**0.458** **+0.140**	**0.429** **+0.120**
Dataset	USC				mRI				Cosmo-MHAD				UTD					
Labeling rate	0.06%				0.36%				0.51%				4.18%					
Modality	Acc		Gyro		IMU		Radar		IMU		Depth		IMU		Skeleton		Depth	
Metrics	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1
Labeltrain	0.229	0.205	0.239	0.175	0.443	0.403	0.233	0.206	0.288	0.213	0.272	0.187	0.411	0.358	0.643	0.587	0.176	0.130
SimCLR	0.363	0.331	0.317	0.294	0.492	0.469	0.272	0.258	0.368	0.288	0.242	0.184	0.452	0.400	0.670	0.624	0.222	0.171
CC	0.326	0.252	0.314	0.302	0.508	0.460	0.373	0.361	0.405	0.306	0.389	0.276	0.456	0.410	0.657	0.616	0.204	0.163
CMC (m2u)	0.517	0.485	0.436	0.428	0.388	0.366	0.280	0.259	0.328	0.243	0.301	0.261	0.370	0.335	0.639	0.600	0.300	0.268
Cosmo (m2u)	0.333	0.271	0.274	0.224	0.202	0.111	0.214	0.197	0.274	0.181	0.355	0.292	0.341	0.287	0.680	0.635	0.285	0.233
MESEN (Ours)	**0.723** **+0.206**	**0.700** **+0.215**	**0.695** **+0.259**	**0.684** **+0.256**	**0.869** **+0.361**	**0.866** **+0.397**	**0.825** **+0.452**	**0.810** **+0.449**	**0.506** **+0.101**	**0.392** **+0.086**	**0.511** **+0.122**	**0.429** **+0.137**	**0.628** **+0.172**	**0.583** **+0.173**	**0.735** **+0.055**	**0.702** **+0.067**	**0.550** **+0.250**	**0.517** **+0.249**

in a server with 4 NVIDIA GeForce RTX 3090 GPUs, 96 GB memory, and an Intel(R) Xeon(R) Gold 6326 (2.90GHz) CPU. Besides, we also run the unimodal fine-tuning stage on Jetson Nano[25] to show the performance of MESEN on edge nodes.

To perform a fair comparison, we conduct baselines and MESEN with the same modality encoders and classifier heads under the same experiment settings, including the same hyperparameters and the same dataset split. The modality encoders consisting of convolutional layers and transformer encoder layers [39] are utilized for all modalities except for the skeleton modality. We use co-occurrence [19] as the modality encoder for skeleton data. The modality projectors used during pre-training are two-layer convolutional layers. The classifier heads are single linear layer classifiers with softmax activation. The learning rate is set at 0.001 for both pre-training and supervised training. The batch size of the pre-training stage is set at 128. For supervised training with labeled data, the batch size is generally set at 64. If the number of labels utilized is smaller than 64, the batch size corresponds to the exact number of labels. Each experiment is conducted independently five times, with different model initialization for each repetition, to mitigate the effect of model initialization on performance.

5.1.4 Metrics. We employ both accuracy and F1-score to measure the performance of baselines and MESEN. Accuracy measures the proportion of correctly predicted samples among all samples. F1-score considers both false positives and false negatives for each activity category.

5.2 Evaluation Results

5.2.1 Overall performance. Labeled data are usually scarce in real-world HAR applications. To evaluate the effectiveness of MESEN, we utilize only a few labeled samples during training. Table 2 provides the performance comparison of MESEN and other baselines under the situations, where only one labeled sample per activity is utilized during training. The labeling rate is defined as the ratio of labeled samples within the whole training set. For example, we employ six labeled samples in total for six activities on the UCI dataset, amounting to 0.35% of the training set.

According to the results, *Labeltrain* exhibits poor performance on all datasets due to label scarcity, while other methods achieve comparatively better performance by utilizing both labeled data and available unlabeled unimodal or multimodal data. To be specific, SimCLR and CC perform better than *Labeltrain* by utilizing unlabeled unimodal data. However, their performance still falls short of MESEN as they do not fully exploit the available multimodal data. CMC (m2u) and Cosmo (m2u) achieve performance improvements for datasets with minor modality gaps, such as the USC dataset. However, they struggle with datasets containing significant heterogeneous modalities like the UTD dataset, where they fail to enhance the recognition performance for all modalities and may even have a negative impact on the performance.

MESEN achieves state-of-the-art results across all datasets, demonstrating its effectiveness in enhancing unimodal HAR performance by exploiting available unlabeled multimodal during training. Relative to the top-performing baselines for each modality, MESEN obtains **5.5% - 45.2%** accuracy improvements, with most gains exceeding 10%. On average, the recognition accuracy of MESEN on all datasets is 65.7%, notably outperforming other methods (**30.7%**, **25.2%**, **27.0%**, **25.8%**, and **27.8%** higher than *Labeltrain*, SimCLR, CC, CMC (m2u), and Cosmo (m2u), respectively). Moreover, MESEN notably boosts the average F1-score to 63.0% (**34.5%**, **26.4%**, **28.7%**, **26.5%**, and **29.9%** higher than *Labeltrain*, SimCLR, CC, CMC (m2u), and Cosmo (m2u), respectively).

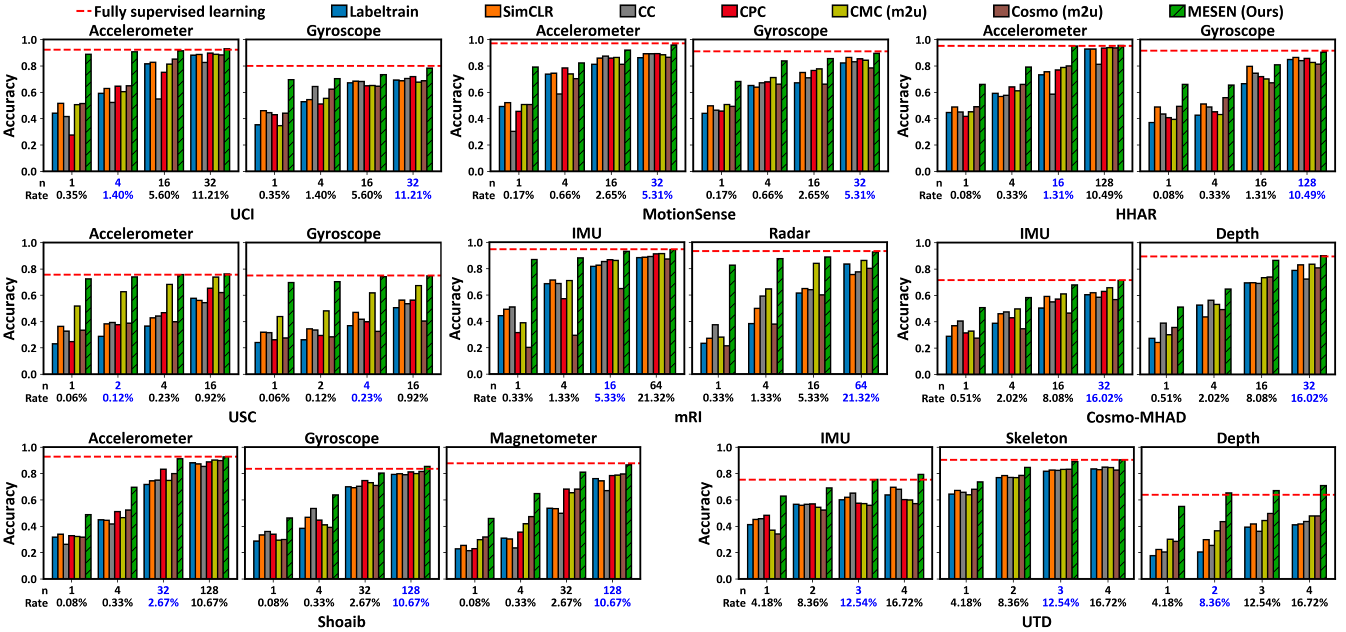


Figure 13: The impact of n labeled samples per activity with different n on HAR performance. 'Rate' represents the labeling rate. The results of MESEN that are comparable to fully supervised learning with 100% labeled data are highlighted in blue.

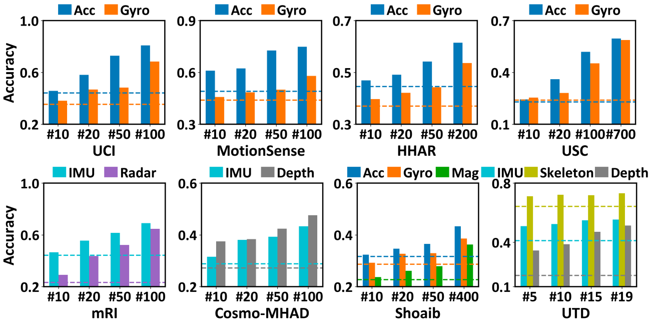


Figure 14: The impact of the available unlabeled multimodal data scale on the performance of MESEN. The x-axis represents the scale of unlabeled data ($\#N = \frac{\text{Unlabeled data}}{\text{Labeled data}}$), and the dashed lines represent *Labeltrain*'s results.

5.2.2 Impact of labeled sample size. We evaluate MESEN and the baselines with different numbers of labeled samples to understand the impact of the labeled sample scale. Specifically, we adopt the settings of utilizing n labeled samples per activity during training. With different n, we have labeling rates that range from 0.06% to 21.32% on all the datasets. As shown in Figure 13, all methods exhibit performance improvements as the number of labeled samples increases. On one hand, MESEN consistently outperforms the baselines under all settings of n across all datasets. On the other hand, MESEN demonstrates a more significant performance improvement when a very small amount of labeled samples is used, indicating its effectiveness in practical scenarios with only a few available labeled samples. Furthermore, even with limited labeled data (ranging from a labeling rate of 0.12% to a labeling rate of 21.32% across different datasets), MESEN achieves performance comparable to or even better than supervised learning with 100% labeled data. This indicates that MESEN effectively extracts effective unimodal features from unlabeled data, reducing the need for extensive labeling.

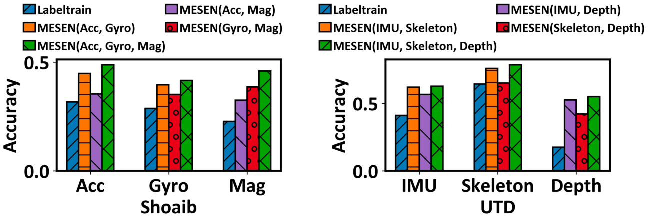
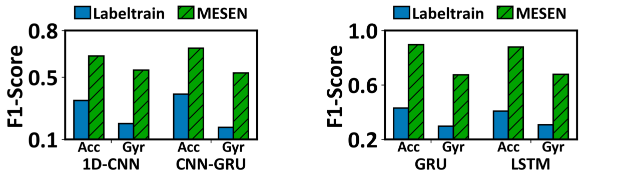


Figure 15: The impact of modality number and type on MESEN's performance. $(\cdot)$ denotes the modalities used during the multimodal-aided pre-training stage.

5.2.3 Impact of unlabeled sample size. MESEN's key principle is to utilize the increasing availability of unlabeled multimodal data for unimodal HAR enhancement. Therefore, the scale of unlabeled data utilized during multimodal-aided pre-training is crucial for MESEN's performance. We conduct experiments to show the impact of different amounts of unlabeled multimodal data. As demonstrated in Figure 14, we fix the number of labeled samples (one labeled sample per activity), and gradually increase the volume of unlabeled multimodal data (#N refers to N times the amount of labeled data). The performance of MESEN increases with a larger unlabeled data amount, suggesting its ability to extract effective information from available unlabeled multimodal data.

5.2.4 Impact of modality number and type. As shown in Figure 15, the performance of MESEN is affected by the number and type of



(a) Performance with Varying Encoders (b) Performance with Varying Classifiers

Figure 16: The effectiveness of MESEN with varying encoders and classifiers.

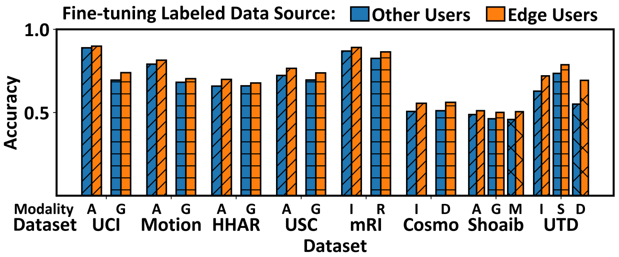
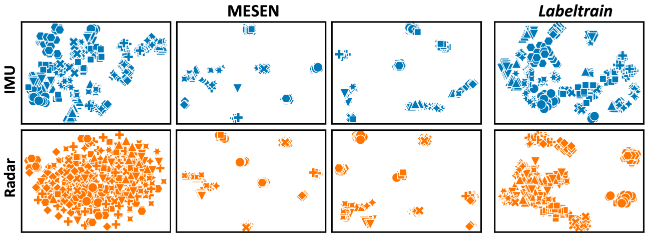


Figure 17: The impact of different labeled data sources on the performance of MESEN.

modalities utilized during the multimodal-aided pre-training stage. On the one hand, MESEN achieves better performance when more modalities are available during pre-training, as multiple modalities provide useful information and guidance for unimodal feature extraction. On the other hand, the effectiveness of MESEN depends on the individual performance of each modality in the recognition task. For example, in the UTD dataset, IMU outperforms other modalities when it is used alone. Therefore, the performance improvement of the Skeleton modality is more significant when IMU and Skeleton are used for pre-training, compared with the improvement achieved when Depth and Skeleton are used.

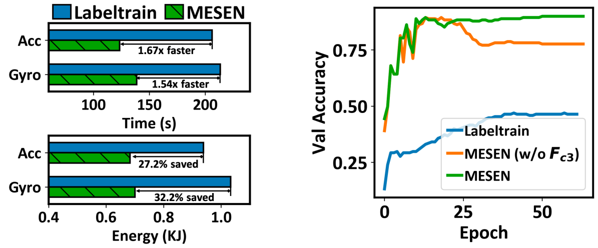
5.2.5 *Impact of different model architectures.* MESEN is designed as a universal framework, which is adaptable to various modality encoders and classifiers. We evaluate the effectiveness of MESEN with varying encoders and classifiers on the UCI dataset. We replace the default encoder with 1D-CNN [36] and CNN-GRU [18] individually to conduct experiments with varying encoders. Similarly, we replace the default classifier with GRU [8] and LSTM [16] individually. As shown in Figure 16, the results demonstrate that MESEN achieves performance improvements compared with *Labeltrain*, regardless of the model architecture used during the pre-training and fine-tuning stages.

5.2.6 *Impact of labeled sample source.* The above experiments are conducted with the user setting as depicted in §5.1. In this setting, the data utilized for inference (validation and testing) comes from users distinct from the user subset involved in the training process. The user subset used for inference refers to 'edge users' and the user subset involved in training is denoted as 'other users'. Furthermore, we evaluate the performance of MESEN when it is fine-tuned with labeled samples from edge users. Figure 17 shows that MESEN can achieve better performance with an average accuracy increase of 4.44% when the labeled samples used during fine-tuning are from edge users.



(a) Before Training (b) Multimodal-aided Pre-training (c) Unimodal Fine-tuning Supervised Unimodal Training

Figure 18: Unimodal features extracted by MESEN and *Labeltrain*. Different shapes denote different activities.



(a) Overhead on Jestson Nano (b) Validation Accuracy

Figure 19: The fine-tuning performance of MESEN.

5.3 System Effectiveness

5.3.1 *Feature visualization.* To demonstrate the benefits of MESEN's multimodal-aided pre-training on unimodal feature extraction, we use t-SNE to visualize the unimodal features extracted by MESEN and *Labeltrain* on the mRI dataset's testing set. The dataset comprises data from two heterogeneous modalities, IMU and mmWave radar. As shown in Figure 18, the unimodal features extracted by MESEN after the multimodal-aided pre-training stage demonstrate clear clustering properties, even before the fine-tuning stage with labeled data. This indicates MESEN's effectiveness in utilizing unlabeled multimodal data during pre-training. Consequently, compared with *Labeltrain*, MESEN acquires more effective features for unimodal HAR with few labels, owing to the information learned during pre-training.

5.3.2 *Performance on the edge node.* We implement the unimodal fine-tuning stage of MESEN on Jetson Nano [25] to evaluate its performance on the user edge node. As shown in Figure 19 (a), when fine-tuned with few labels (one labeled sample per activity) on Jetson Nano, MESEN is more time and energy-efficient than the supervised baseline *Labeltrain*. Specifically, without incurring any additional memory overhead compared with *Labeltrain*, the fine-tuning stage of MESEN is **1.67**× and **1.54**× faster than *Labeltrain* on the UCI dataset for the two modalities (accelerometer and gyroscope), saving **27.2%** and **32.2%** energy usage, respectively.

5.3.3 *Ablation study.* We evaluate the contributions of three components in MESEN: cross-modal feature contrastive learning (P_{c1}), multimodal pseudo-classification aligning (P_{c2}), and the layer-aware fine-tuning mechanism (F_{c3}). Figure 20 shows that combining P_{c1} and P_{c2} significantly improves unimodal HAR performance across various multimodal combinations. Moreover, Figure 19 (b) shows

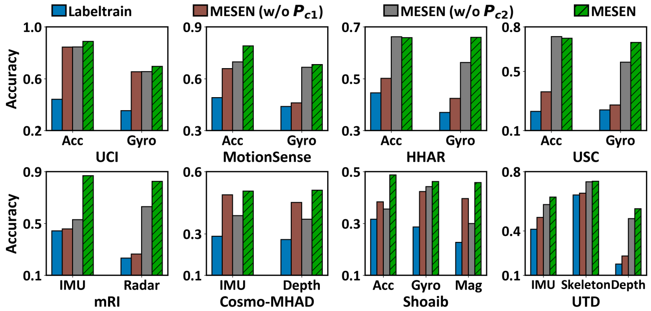


Figure 20: Ablation study for P_{c1} and P_{c2} of MESEN.

the effectiveness of F_{c3} in mitigating overfitting when MESEN adapts to unimodal HAR with only a few labeled samples.

5.3.4 *Impact of parameters settings.* We evaluate MESEN's sensitivity to various system settings on the MotionSense dataset. Figure 21 shows the impact of pre-training batch size, contrastive feature dimension, and the number of pseudo-class on MESEN's performance. Compared with the other two factors, MESEN is particularly sensitive to the pseudo-class number utilized in multimodal pseudo-classification aligning, which is designed as an effective prompt for recognition as described in §4.2.2. The difference between the number of pseudo-classes and the actual activity category number can impede feature extraction, thus affecting recognition performance. However, as the number of activity categories is typically readily accessible and requires no extra effort, the performance decrease can be effectively avoided.

6 DISCUSSION

Scalability. The increasing number of modality pairs will introduce extra training costs during the multimodal-aided pre-training stage. An additional modality encoder and a projector are needed for computing features of each new modality during pre-training. Adopting the assumed atomic operation of the contrastive objective function in COCOA [10], i.e., the dot-product of sample pairs, the time complexity of MESEN is $O(M^2N^2 + M^2{N_{cls}}^2)$, where M is the number of modalities, N is the number of input pairs in the mini-batch $\mathcal{B}$, and N_{cls} is the number of activity categories. When the activity categories and the number of input pairs are fixed during implementation, the complexity of MESEN will be $O(M^2)$. In practical implementation, the extra training costs vary based on the types of additional modalities. For example, compared with the IMU modality, depth images introduce more requirements for computational resources.

Cross-dataset performance. To evaluate the performance of MESEN in cross-dataset scenarios, we conduct experiments on the UCI and MotionSense datasets. Following UniHAR [41], we select four activities (*still*, *walk*, *walk upstairs*, and *walk downstairs*) contained in both datasets. For the cross-dataset setting, we train models on the UCI dataset while validating and testing models on the MotionSense dataset. Other experimental settings align with those in §5.2.1. Under the cross-dataset setting, MESEN achieves accuracy of 0.536 and 0.787 with accelerometer and gyroscope, respectively, while *Labeltrain* exhibits 0.409 and 0.690. The accuracy gains of 0.127 and 0.097 by MESEN over *Labeltrain* demonstrate its adaptability in cross-dataset scenarios. However, MESEN is affected by the domain discrepancies between these two datasets, evidenced by its accuracy of 0.970 and 0.994 on the UCI dataset for the four activities. The *Physics-Informed Data Augmentation* approach proposed by UniHAR [41] to address data heterogeneity provides a solution to improve MESEN in further study.

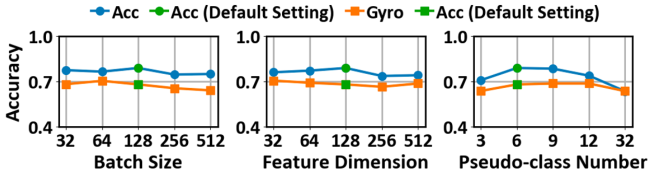


Figure 21: The impact of different parameter settings on the performance of MESEN.

Multimodal inference. MESEN is designed to operate in a multi-to-unimodal mode, which may present limitations in some HAR application scenarios. When multiple modalities are readily available during the deployment phase, MESEN can utilize these multimodal data streams through slight adaptation. This is because each modality encoder has been effectively trained during MESEN's multimodal-aided pre-training. For example, with the experimental settings in §5.2.1, MESEN achieves 0.899 accuracy on the UCI dataset by applying multimodal fusion through concatenating multimodal features during fine-tuning and inference. However, since multimodal fusion is not the primary design focus of MESEN, it might not capture the full potential of such multimodal streams as effectively as Cosmo [26]. Further study is needed to extend MESEN to address broader application scenarios.

7 CONCLUSION

This paper proposes MESEN, a universal framework utilizing increasingly available unlabeled multimodal data to enhance unimodal HAR with few labels. MESEN achieves effective unimodal feature extraction during the multimodal-aided pre-training stage by featuring a multi-task mechanism. The proposed mechanism combines cross-modal feature contrastive learning and multimodal pseudo-classification aligning to exploit the correlations and relationships within multimodal data. With the extracted effective unimodal features, MESEN then can adapt to downstream unimodal HAR with only a few labeled samples. Our evaluation demonstrates that MESEN can significantly improve unimodal HAR performance by exploiting multimodal data.

ACKNOWLEDGMENTS

Thank Prof. Zhenyu Yan and Dr. Xiaomin Ouyang for granting access to the Cosmo-MHAD dataset and for their insightful discussions. Thank the anonymous reviewers and Shepherd for the valuable comments. This research is supported by the National Natural Science Foundation of China under Grant No. U1909207, Singapore Ministry of Education Academic Research Fund Tier 1 (RG88/22), Key Research and Development Program of Zhejiang Province under Grant No. 2021C03037, and Fundamental Research Funds for the Central Universities under Grants 226-2022-00107/226-2023-00111.

REFERENCES

[1] Sizhe An, Yin Li, and Umit Ogras. 2022. mri: Multi-modal 3d human pose estimation dataset using mmwave, rgb-d, and inertial sensors. *Advances in Neural Information Processing Systems* 35 (2022), 27414–27426.

[2] Olasimbo Ayodeji Arigbabu. 2020. Entropy decision fusion for smartphone sensor based human activity recognition. *arXiv preprint arXiv:2006.00367* (2020).

[3] Chongguang Bi, Guoliang Xing, Tian Hao, Jina Huh, Wei Peng, and Mengyan Ma. 2017. Familylog: A mobile system for monitoring family mealtime activities. In *2017 ieee international conference on pervasive computing and communications (percom)*. IEEE, 21–30.

[4] Chen Chen, Roozbeh Jafari, and Nasser Kehtarnavaz. 2015. UTD-MHAD: A multimodal dataset for human action recognition utilizing a depth camera and a wearable inertial sensor. In *2015 IEEE International conference on image processing (ICIP)*. IEEE, 168–172.

[5] Ling Chen, Rong Hu, Menghan Wu, and Xin Zhou. 2023. HMGAN: A Hierarchical Multi-Modal Generative Adversarial Network Model for Wearable Human Activity Recognition. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 7, 3 (2023), 1–27.

[6] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*. PMLR, 1597–1607.

[7] Zhenyu Chen, Mu Lin, Fanglin Chen, Nicholas D Lane, Giuseppe Cardone, Rui Wang, Tianxing Li, Yiqiang Chen, Tanzeem Choudhury, and Andrew T Campbell. 2013. Unobtrusive sleep monitoring using smartphones. In *2013 7th International Conference on Pervasive Computing Technologies for Healthcare and Workshops*. IEEE, 145–152.

[8] Kyunghyun Cho, Bart Van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the properties of neural machine translation: Encoder-decoder approaches. *arXiv preprint arXiv:1409.1259* (2014).

[9] Neha Dawar and Nasser Kehtarnavaz. 2018. A convolutional neural network-based sensor fusion system for monitoring transition movements in healthcare applications. In *2018 IEEE 14th International Conference on Control and Automation (ICCA)*. IEEE, 482–485.

[10] Shohreh Deldari, Hao Xue, Aaqib Saeed, Daniel V Smith, and Flora D Salim. 2022. COCOA: Cross Modality Contrastive Learning for Sensor Data. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6, 3 (2022), 1–28.

[11] Hongchao Fang, Sicheng Wang, Meng Zhou, Jiayuan Ding, and Pengtao Xie. 2020. Cert: Contrastive self-supervised learning for language understanding. *arXiv preprint arXiv:2005.12766* (2020).

[12] Daniel Gatica-Perez, Joan-Isaac Biel, David Labbe, and Nathalie Martin. 2019. Discovering eating routines in context with a smartphone app. In *Adjunct Proceedings of the 2019 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2019 ACM International Symposium on Wearable Computers*. 422–429.

[13] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. 2023. Imagebind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 15180–15190.

[14] Harish Haresamudram, Irfan Essa, and Thomas Plötz. 2021. Contrastive predictive coding for human activity recognition. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5, 2 (2021), 1–26.

[15] Shibo He, Kun Shi, Chen Liu, Bicheng Guo, Jiming Chen, and Zhiguo Shi. 2022. Collaborative sensing in Internet of Things: A comprehensive survey. *IEEE Communications Surveys & Tutorials* (2022).

[16] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.

[17] Ankita Jain and Vivek Kanhangad. 2017. Human activity classification in smartphones using accelerometer and gyroscope sensors. *IEEE Sensors Journal* 18, 3 (2017), 1169–1177.

[18] Yeon-Wook Kim, Kyung-Lim Joa, Han-Young Jeong, and Sangmin Lee. 2021. Wearable IMU-based human activity recognition algorithm for clinical balance assessment using 1D-CNN and GRU ensemble model. *Sensors* 21, 22 (2021), 7628.

[19] Chao Li, Qiaoyong Zhong, Di Xie, and Shiliang Pu. 2018. Co-occurrence feature learning from skeleton data for action recognition and detection with hierarchical aggregation. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*. 786–792.

[20] Yunfan Li, Peng Hu, Zitao Liu, Dezhong Peng, Joey Tianyi Zhou, and Xi Peng. 2021. Contrastive clustering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 8547–8555.

[21] Chris Xiaoxuan Lu, Muhamad Risqi U Saputra, Peijun Zhao, Yasin Almalioglu, Pedro PB De Gusmao, Changhao Chen, Ke Sun, Niki Trigoni, and Andrew Markham. 2020. milliEgo: single-chip mmWave radar aided egomotion estimation via deep sensor fusion. In *Proceedings of the 18th Conference on Embedded Networked Sensor Systems*. 109–122.

[22] Haojie Ma, Wenzhong Li, Xiao Zhang, Songcheng Gao, and Sanglu Lu. 2019. AttnSense: Multi-level attention mechanism for multimodal human activity recognition.. In *IJCAI*. 3109–3115.

[23] Haojie Ma, Zhijie Zhang, Wenzhong Li, and Sanglu Lu. 2021. Unsupervised human activity representation learning with multi-task deep clustering. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5, 1 (2021), 1–25.

[24] Mohammad Malekzadeh, Richard G Clegg, Andrea Cavallaro, and Hamed Haddadi. 2019. Mobile sensor data anonymization. In *Proceedings of the international conference on internet of things design and implementation*. 49–58.

[25] NVIDIA Corporation. [n. d.]. Jetson Nano Developer Kit. https://developer.nvidia.com/embedded/jetson-nano-developer-kit

[26] Xiaomin Ouyang, Xian Shuai, Jiayu Zhou, Ivy Wang Shi, Zhiyuan Xie, Guoliang Xing, and Jianwei Huang. 2022. Cosmo: contrastive fusion learning with small data for multimodal human activity recognition. In *Proceedings of the 28th Annual International Conference on Mobile Computing And Networking*. 324–337.

[27] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems* 32 (2019).

[28] Fazlay Rabbi, Taiwoo Park, Biyi Fang, Mi Zhang, and Youngki Lee. 2018. When virtual reality meets internet of things in the gym: Enabling immersive interactive machine exercises. *Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies* 2, 2 (2018), 1–21.

[29] Meera Radhakrishnan, Darshana Rathnayake, Ong Koon Han, Inseok Hwang, and Archan Misra. 2020. ERICA: enabling real-time mistake detection & corrective feedback for free-weights exercises. In *Proceedings of the 18th Conference on Embedded Networked Sensor Systems*. 558–571.

[30] Jorge-L Reyes-Ortiz, Luca Oneto, Albert Samà, Xavier Parra, and Davide Anguita. 2016. Transition-aware human activity recognition using smartphones. *Neurocomputing* 171 (2016), 754–767.

[31] Aaqib Saeed, Tanir Ozcelebi, and Johan Lukkien. 2019. Multi-task self-supervised learning for human activity detection. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 3, 2 (2019), 1–30.

[32] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*. 618–626.

[33] Zhiyao Sheng, Huatao Xu, Qian Zhang, and Dong Wang. 2022. Facilitating Radar-Based Gesture Recognition With Self-Supervised Learning. In *2022 19th Annual IEEE International Conference on Sensing, Communication, and Networking (SECON)*. IEEE, 154–162.

[34] Muhammad Shoaib, Stephan Bosch, Ozlem Durmaz Incel, Hans Scholten, and Paul JM Havinga. 2014. Fusion of smartphone motion sensors for physical activity recognition. *Sensors* 14, 6 (2014), 10146–10176.

[35] Allan Stisen, Henrik Blunck, Sourav Bhattacharya, Thor Siiger Prentow, Mikkel Baun Kjærgaard, Anind Dey, Tobias Sonne, and Mads Møller Jensen. 2015. Smart devices are different: Assessing and mitigatingmobile sensing heterogeneities for activity recognition. In *Proceedings of the 13th ACM conference on embedded networked sensor systems*. 127–140.

[36] Wensi Tang, Guodong Long, Lu Liu, Tianyi Zhou, Jing Jiang, and Michael Blumenstein. 2020. Rethinking 1d-cnn for time series classification: A stronger baseline. *arXiv preprint arXiv:2002.10061* (2020), 1–7.

[37] Yonglong Tian, Dilip Krishnan, and Phillip Isola. 2020. Contrastive multiview coding. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*. Springer, 776–794.

[38] Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of machine learning research* 9, 11 (2008).

[39] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).

[40] Dong Wang, Ning Ding, Piji Li, and Haitao Zheng. 2021. CLINE: Contrastive Learning with Semantic Negative Examples for Natural Language Understanding. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. 2332–2342.

[41] Huatao Xu, Pengfei Zhou, Rui Tan, and Mo Li. 2023. Practically Adopting Human Activity Recognition. In *Proceedings of the 29th Annual International Conference on Mobile Computing and Networking*. 1–15.

[42] Huatao Xu, Pengfei Zhou, Rui Tan, Mo Li, and Guobin Shen. 2021. Limu-bert: Unleashing the potential of unlabeled data for imu sensing applications. In *Proceedings of the 19th ACM Conference on Embedded Networked Sensor Systems*. 220–233.

[43] Shuochao Yao, Shaohan Hu, Yiran Zhao, Aston Zhang, and Tarek Abdelzaher. 2017. Deepsense: A unified deep learning framework for time-series mobile sensing data processing. In *Proceedings of the 26th international conference on world wide web*. 351–360.

[44] Shuochao Yao, Yiran Zhao, Huajie Shao, Chao Zhang, Aston Zhang, Shaohan Hu, Dongxin Liu, Shengzhong Liu, Lu Su, and Tarek Abdelzaher. 2018. Sensegan: Enabling deep learning for internet of things with a semi-supervised framework. *Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies* 2, 3 (2018), 1–21.

[45] Mi Zhang and Alexander A Sawchuk. 2012. USC-HAD: A daily activity dataset for ubiquitous activity recognition using wearable sensors. In *Proceedings of the 2012 ACM conference on ubiquitous computing*. 1036–1043.