Cosmo: Contrastive Fusion Learning with Small Data for Multimodal Human Activity Recognition

Xiaomin Ouyang[1], Xian Shuai[1], Jiayu Zhou[2], Ivy Wang Shi[3], Zhiyuan Xie[1], Guoliang Xing[1,*] and Jianwei Huang[4,5,*]
[1]The Chinese University of Hong Kong, [2]Michigan State University, [3]Li Po Chun United World College, Hong Kong, [4]The Chinese University of Hong Kong, Shenzhen, [5]Shenzhen Institute of Artificial Intelligence and Robotics for Society

ABSTRACT

Human activity recognition (HAR) is a key enabling technology for a wide range of emerging applications. Although multimodal sensing systems are essential for capturing complex and dynamic human activities in real-word settings, they bring several new challenges including limited labeled multimodal data. In this paper, we propose Cosmo, a new system for contrastive fusion learning with small data in multimodal HAR applications. Cosmo features a novel two-stage training strategy that leverages both unlabeled data on the cloud and limited labeled data on the edge. By integrating novel fusion-based contrastive learning and quality-guided attention mechanisms, Cosmo can effectively extract both *consistent and complementary information* across different modalities for efficient fusion. Our evaluation on a cloud-edge testbed using two public datasets and a new multimodal HAR dataset shows that Cosmo delivers significant improvement over state-of-the-art baselines in both recognition accuracy and convergence delay.

CCS CONCEPTS

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

KEYWORDS

Human activity recognition, Heterogeneous multimodal fusion, Contrastive learning

ACM Reference Format:
Xiaomin Ouyang[1], Xian Shuai[1], Jiayu Zhou[2], Ivy Wang Shi[3], Zhiyuan Xie[1],, Guoliang Xing[1,*] and Jianwei Huang[4,5,*]. 2022. Cosmo: Contrastive Fusion Learning with Small Data for Multimodal Human Activity Recognition. In *The 28th Annual International Conference On Mobile Computing And Networking (ACM MobiCom '22), October 17–21, 2022, Sydney, NSW, Australia.* ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3495243.3560519

*Guoliang Xing and Jianwei Huang are the corresponding authors.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
ACM MobiCom '22, October 17–21, 2022, Sydney, NSW, Australia
© 2022 Association for Computing Machinery.
ACM ISBN 978-1-4503-9181-8/22/10…$15.00
https://doi.org/10.1145/3495243.3560519

1 INTRODUCTION

Human activity recognition (HAR) has a wide range of applications such as virtual/ augmented reality (VR/AR) [41, 42], smart home [3, 5], and smart health [13, 18]. In real-world scenarios, human activities are usually highly complex and dynamic in nature, most of which are difficult to capture using only a single sensor modality. Moreover, certain sensors such as RGB cameras may not be an option in many applications due to the increasing privacy concerns. To address these issues, several new multimodal sensing systems have been proposed to leverage multiple privacy-preserving sensor modalities recently emerged, e.g., depth camera and radar, for HAR applications [10, 32, 35, 45, 52]. For example, in Alzheimer's Disease monitoring, these sensors can track the elder's daily activities [12, 17, 26], which are important digital biomarkers for early diagnosis.

However, fusing multiple sensor modalities in HAR applications presents several major challenges. First, different types of sensors usually produce highly heterogeneous information about the same events/activities. For example, the inertial measurements and depth images not only have significantly different dimensions and patterns but also may not be synchronized in the time domain, making the fusion challenging. Second, there usually exists a very limited amount of *labeled* data, as it is difficult to label multimodal data in real-world settings [23, 37, 45]. For example, the data of many sensors such as IMU and mmWave radar is not intuitive for human, making the annotation extremely labor-intensive [45]. Third, the sensor data in HAR applications is often privacy-sensitive in nature and cannot be transmitted to the cloud. Lastly, the activity recognition model needs to be customized for individuals whose activities may yield dynamic characteristics over time [39, 46], which requires *on-device training* using continuous multimodal data.

Unfortunately, existing multimodal sensing systems cannot address these challenges collectively. Most previous efforts [6, 32, 34, 35] are focused on a pair of specific sensor modalities, and cannot be extended to the fusion of other heterogeneous sensor modalities. There are also general multimodal fusion frameworks based on deep learning [36, 54, 55]. However, they are based on fully supervised learning approaches, which need to be trained by extensive labeled multimodal data and hence are ill-suited for real-world HAR applications with only limited data labels. Recently, *contrastive learning* [38, 48] has been proposed in the field of self-supervised representation learning to address the challenge of limited labeled data. However, they are either developed for HAR tasks with a single modality [19, 47, 53] or designed for tasks in other domains with two similar modalities (e.g., different channels of an image in vision tasks) [4, 22, 49], which cannot be applied for significantly heterogeneous data modalities such as depth images and inertial measurements in multimodal HAR applications [33, 34]. Lastly,

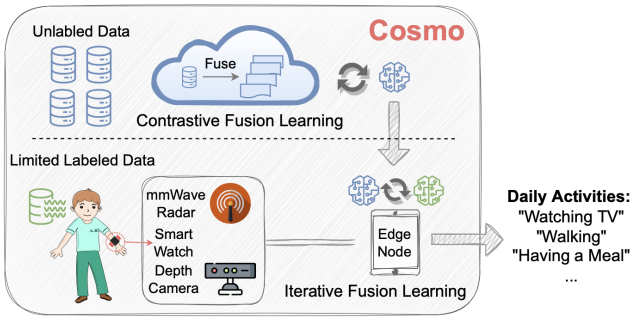


Figure 1: A typical application of Cosmo in multimodal human activity recognition. The first stage (i.e., contrastive fusion learning) is trained with unlabeled multimodal data on the cloud, and the second stage (i.e., iterative fusion learning) is trained with limited labeled data locally.

traditional supervised learning models for multimodal fusion are typically trained on the cloud due to the high compute overhead and the need of a large amount of training data [37, 60].

In this paper, we propose Cosmo, a novel system for contrastive fusion learning with small data in multimodal HAR applications. Cosmo features a novel two-stage training strategy using both unlabeled and (limited) labeled data from multiple heterogeneous sensors. In the first stage, Cosmo employs a novel fusion-based contrastive learning approach to train the feature encoders using unlabeled multimodal data. As a result, Cosmo can extract *consistent information* that represents the common knowledge shared among different modalities. In the second stage, a new quality-guided attention mechanism is designed to allow the classifier to capture the strengths of different modalities based on only limited labeled data, which explores the *complementary information* of different modalities. We then propose a novel iterative fusion learning algorithm, which improves both the accuracy and convergence performance of the system. As shown in Figure 1, Cosmo naturally enables a cloud-edge implementation architecture. The first stage is trained on the cloud with unlabeled multimodal data gathered from multiple users or public datasets. Then the second stage is trained locally with limited labeled data, which incurs low compute overhead and preserves user privacy. Cosmo also adapts to dynamic environments by improving the training accuracy iteratively through a small amount of data labeled by users locally (e.g., marking the time of having lunch would automatically label the multimodal data during lunch).

We evaluate the performance of Cosmo extensively on a testbed of a server and Nvidia Jetson TX2 platforms, using two public datasets and a new multimodal HAR dataset, which, in total, consist of data from five different sensor modalities (i.e., accelerator, gyroscope, skeleton points, depth images, and mmWave radar) and 55 different daily human activities. Our evaluation shows that, Cosmo delivers an accuracy improvement of 51.61%, 26.73% and 20.90% over the single modal learning, supervised fusion learning and contrastive learning baseline, respectively, and converges much faster than conventional supervised fusion learning.

In summary, we make the following key contributions:

- Through a motivational case study based on real-world HAR datasets, we show that multimodal fusion should leverage both *consistent and complementary information* of different modalities simultaneously, when there exists only limited labeled data.
- We design Cosmo, a novel system for fusing multiple heterogeneous sensor modalities in HAR applications with only limited labeled data. Cosmo effectively explores *consistent information* from unlabeled multimodal data through a new fusion-based contrastive learning approach, and integrates *complementary information* from limited labeled data through a novel quality-guided attention mechanism and an iterative learning algorithm.
- Our evaluation on a cloud-edge testbed using two public datasets and a new multimodal HAR dataset shows that Cosmo outperforms state-of-the-art baselines and converges much faster than conventional supervised learning approaches.

2 RELATED WORK

Human Activity Recognition (HAR) has a wide range of applications such as virtual/augmented reality (VR/AR) [41, 42], smart home [3, 5], and smart health [13, 18]. Machine learning algorithms based on handcrafted features [24, 46] and deep neural networks [25, 53] have been applied to HAR applications. Several recent studies are focused on leveraging federated learning systems to improve the performance of HAR while protecting user privacy [39].

Multimodal Sensing for HAR. Multimodal sensing systems have become prevalent in identifying complex and dynamic human activities [29, 31, 32, 58]. For example, PDLens [58] utilizes multiple built-in sensors of a smartphone to monitor a user's daily activity. Liu et al. [32] aim to fuse the RFID and depth camera for recognizing human gestures. However, these approaches are designed for a fixed pair of specific sensor modalities, thus cannot be extended to the fusion of other heterogeneous sensor modalities. There are also general multimodal fusion frameworks based on deep learning [36, 37, 55]. For example, *AttnSense* [36] leverages an attention-based module to dynamically learn the weights for concatenating the features of different modalities. Most of the previous work in this area is based on supervised learning, which may fail to adapt to real-world scenarios where only limited labeled data is available. To address the challenge of collecting and labeling massive training data, previous RF sensing approaches [8] exploit the correlations among online videos and RF signals. Cosmo takes a different approach by leveraging unlabeled multimodal data that is easier to obtain to improve fusion performance for HAR applications.

Self-Supervised Multimodal Learning. To address the challenge of limited labeled data, contrastive multimodal learning [38, 48] has been proposed recently in the field of self-supervised representation learning. Most of current approaches are developed with two similar modalities (e.g., different channels of images in vision tasks). They either learn a cross-modal embedding space by contrasting different modalities [48, 49], or perform mutual clustering among multimodal features [4, 9, 22, 40], which only captures the information shared across modalities while failing to fully leverage multimodal synergies [33]. In particular, the sensor modalities become increasingly more heterogeneous (e.g., depth images and IMU data) in real-world HAR applications. In this work, we propose a novel fusion-based contrastive learning framework, which is able to to extract key information from heterogeneous multimodal

data for efficient fusion. We also explore the strengths of different modalities from limited labeled data, and effectively combine it with information learned from unlabeled data. There are also contrastive predictive coding methods designed for single-modality HAR applications that capture the temporal structure of sensor data [19, 47, 53]. However, they cannot be applied to multimodal HAR tasks as they do not explore multimodal data fusion during the unsupervised learning stage.

3 MOTIVATION

In this section, we first compare different supervised learning approaches using real-world multimodal HAR datasets, and then evaluate their performance with limited labeled data. The key insights from these results motivate the design of Cosmo.

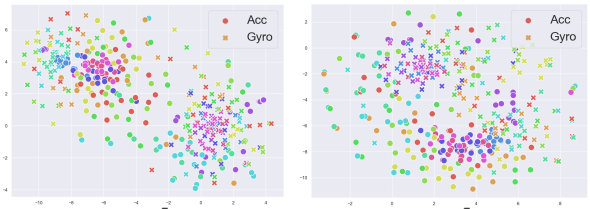
3.1 Understanding the Fusion Performance

In this section, we compare three supervised learning approaches, including single modal learning (referred to as *SingleModal*), as well as two state-of-the-art fusion approaches *Deepsense* [55] and *AttnSense* [36] that are representative fusion approaches in multimodal HAR tasks [11, 37, 54]. In *Deepsense*, features from different sensor modalities are concatenated together for fusion. In *AttnSense*, an attention-based module dynamically learns the weights for concatenating features of different sensors.

Specifically, we evaluate the above approaches using 14 subjects' multimodal data from the public USC dataset [59]. The task is to classify 12 human activities using the accelerometer (Acc) and gyroscope (Gyro) data. We use data from ten subjects for training and the other four subjects for testing. The deep learning model comprises five CNN layers, two GRU layers, and one fully connected layer. Each experiment is repeated five times.

3.1.1 Complementary nature of different modalities. Figure 3 shows the mean testing accuracy of each activity when using Acc-only and Gyro-only in *SingleModal* approach. We observe that the accelerometer performs better for recognizing walking-related activities (e.g., walking forward, left, right, upstairs, downstairs), while the gyroscope outperforms in the remaining activities. It clearly shows the intrinsic complementary nature of the two sensors, which offers different strengths in a given activity recognition task and can be used to improve the fusion performance.

3.1.2 Consistent information across modalities. Figure 2a visualizes the features of Acc and Gyro generated by *Deepsense* using t-Distributed Stochastic Neighbor Embedding (t-SNE) [51]. Here different colors represent different activities, while different shapes of points denote different modalities. We observe that the features of the two modalities are well aligned and almost symmetrical along the diagonal. Moreover, the average cosine distance between Acc and Gyro features generated by *Deepsense* is 0.7288, which is smaller than that learned by *SingleModal* (0.8067). As shown in [16, 60], the features are more consistent if the averaged distance between them is smaller. This means that through concatenating the features from different modalities, *Deepsense* captures more consistent information between two modalities, which makes it more robust to the noisy multimodal data.



(a) *Deepsense* ($\bar{D} = 0.7288$). (b) *Attnsense* ($\bar{D} = 0.7685$).

Figure 2: Visualization of Acc and Gyro features generated by different fusion approaches. Here $\bar{D}$ denotes the mean cosine distance between Acc and Gyro features. *Deepsense* learns more consistent information (the features of two modalities have a smaller mean distance and are more aligned), while *AttnSense* combines more complementary information.

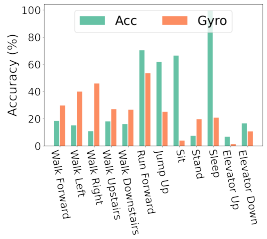


Figure 3: Acc outperforms Gyro in walking-related activities while it is the opposite for other activities.

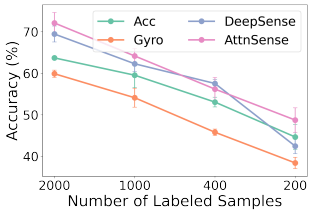


Figure 4: Accuracy comparison with different numbers of labeled samples.

3.1.3 Combining consistent and complementary information. As shown in Figure 2b, the features of different sensors generated by *AttnSense* are less aligned compared with *DeepSense* in Figure 2a. Moreover, the average distance of Acc and Gyro features in *AttnSense* (0.7685) is between those of *SingleModal* and *Deepsense*. This means that, through concatenating the features with different weights, *AttnSense* combines consistent and complementary information from different modalities [16, 60].

3.2 Impact of Small Data

Labeling multimodal data for HAR tasks is extremely difficult in real-world settings [23, 37, 45], which poses significant challenges to supervised learning-based fusion approaches. Here we investigate the performance of the above three supervised learning approaches with different numbers of labeled samples. The training data is reduced randomly in the experiments with a balanced class distribution. The results are shown in Figure 4.

First, the accuracy of all methods drops with less labeled data. Moreover, *Deepsense* improves the activity recognition performance over *SingleModal* only when there exists enough labeled data (e.g., 2,000 samples). In particular, when the number of labeled multimodal samples is small (e.g., 200), the accuracy of *Deepsense* (42.43%) is even lower than that of Acc-only (44.64 %). This means that only capturing consistent information is not enough to improve fusion performance. In this case, *AttnSense* performs better (48.72%) through assigning dynamic fusion weights to the input data. That is, it is beneficial to combine both consistent and complementary

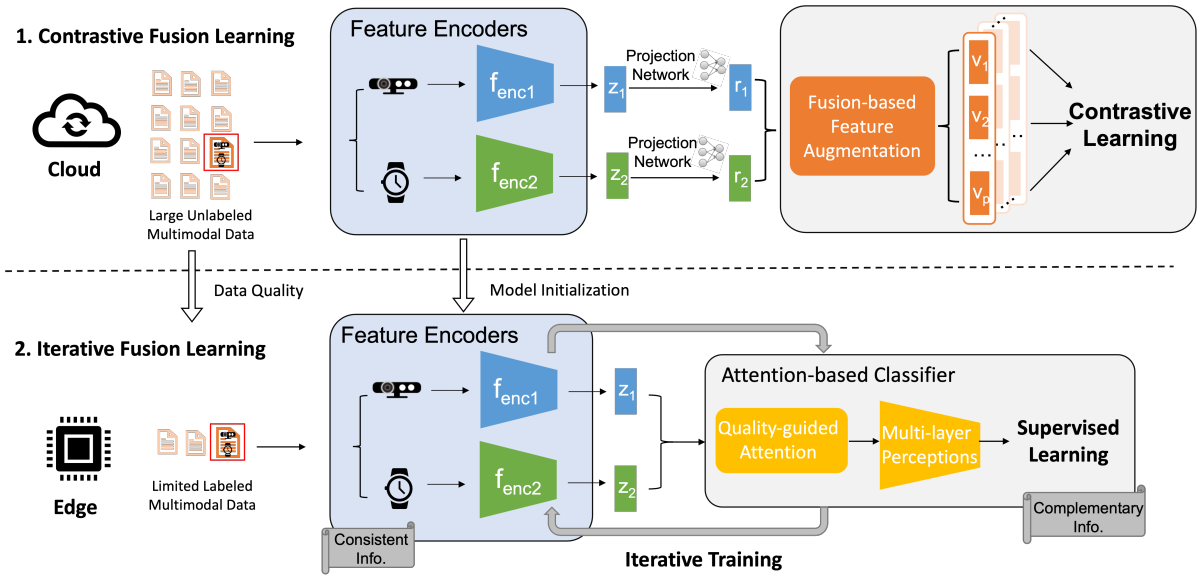


Figure 5: Cosmo consists of two stages, i.e., contrastive learning on the cloud for capturing consistent information from unlabeled multimodal data, and supervised learning on edge for combining complementary information from limited labeled multimodal data, respectively.

information from different modalities. However, the accuracy improvement of *AttnSense* over a single modality is still limited, as the small amount of labeled multimodal data is insufficient to learn a robust fusion strategy.

3.3 Summary

We now summarize the analysis on different fusion approaches.

- The **consistent information** from different data modalities helps align features, making the fusion more robust to noise.
- The **complementary information** from different data modalities, on the other hand, exploits the strength of different sensors and promotes fusion performance.
- When there exists only limited labeled multimodal data, both consistent and complementary information should be learned and leveraged simultaneously to achieve more robust fusion performance in real-world HAR tasks.

4 SYSTEM OVERVIEW

We now introduce Cosmo, a new system for contrastive fusion learning with small data in multimodal HAR applications. Motivated by the insights from Section 3, our key idea is to first capture consistent information from unlabeled multimodal data, and then iteratively learn complementary information of different sensors from limited labeled data. We will first discuss the application scenarios and then describe the system architecture.

4.1 Application Scenarios

Cosmo is designed for a wide range of applications where multiple heterogeneous sensors are deployed to track users' activities in a continuous and longitudinal manner [12, 28, 58]. For example, in Alzheimer's Disease monitoring applications [12], wearables and smartphone sensors can continuously track the elder's daily activities such as sleeping and social interaction, which are important digital biomarkers [1, 26] for early diagnosis.

We now outline several challenges of such applications. First, the user data (either unlabeled or labeled) in this application is privacy-sensitive and usually cannot be uploaded to the cloud. Second, even when the user data can be shared, labeling such multimodal data is usually challenging in real-world settings [23, 37, 45]. For example, data of many sensors such as IMU and radar is not intuitive for human, making the annotation extremely labor-intensive. Finally, different users usually yield highly diverse activities and behaviors. The characteristics of activities for the same user may also change over time because of a number of reasons, such as the change of daily routine or progression of the disease. These challenges together motivate the design of Cosmo.

4.2 System Architecture

Cosmo features a novel two-stage training strategy that can efficiently learn consistent and complementary information from unlabeled and limited labeled multimodal data. Here we use the fusion of a depth camera and wrist-worn wearables in HAR as a running example to illustrate our design (while Cosmo supports the fusion of more than two modalities). Figure 5 shows the overall system architecture of Cosmo.

In the first stage, we design a fusion-based contrastive learning framework that trains the feature encoder of each modality (e.g., depth images and IMU data) to learn consistent information from unlabeled multimodal data. First, the feature encoders generate the unimodal features, which will then be transformed to the same dimension through the projection networks. Second, a fusion-based feature augmentation module will extensively augment the unimodal features to a group of fused features via weighted sum or

concatenation. Finally, the contrastive learning will bring the fused features from the same multimodal data sample (denoted as *positive samples*) together, while pushing fused features from other data samples (denoted as *negative samples*) apart. In this way, the feature encoders are trained to learn consistent information across modalities by maximizing the mutual information of features from different modalities. For example, when fusing a depth image containing the subjects and the motion sensor data from the wearables, this stage will learn common information that aligns the features of the depth map (e.g., the bounding boxes of subjects) with features of wearable data (e.g., motion vectors).

In the second stage, we design an iterative fusion learning approach to effectively combine the complementary information from limited labeled multimodal data. First, we will initialize the feature encoders using model weights trained in the first stage and then fine-tune them. Second, we design a novel *quality-guided attention fusion* module that allows the classifier to capture the complementary information of different modalities based on only limited labeled data, where the data quality of each modality is measured using unlabeled multimodal data. The complementary information explores the exclusive and unique contents of different modalities. For example, when detecting the activities of "having a lunch", a depth image will exhibit the sitting postures of subjects while the wearables can capture the hand motions during eating. Finally, the feature encoders and classifier will be trained iteratively to gradually learn the complementary information until convergence. A key advantage of such iterative learning design is that it addresses the challenge of exploring complementary information from consistent features and prevents the model from degrading to the one learned by conventional supervised learning approaches.

Most multimodal fusion approaches incur high computation overhead and hence can only be trained on the cloud or powerful platforms. In contrast, our two-stage training strategy of Cosmo naturally enables a could-edge implementation architecture. The first stage can be trained on the cloud with large amounts of unlabeled multimodal data gathered from multiple users or public datasets. When there is connectivity to the cloud, the model weights of feature encoders can be downloaded to edge devices. Then the second stage can be trained locally with limited labeled data. Such a cloud-edge implementation offers several advantages. First, it only incurs low computation overhead and hence is affordable for resource-limited edge devices like Nvidia Jetson TX2 [2]. Second, the labeled data is kept locally to preserve user privacy. Third, the overall training performance can be improved iteratively by allowing users to provide a small number of labels (e.g., marking the time of having lunch would automatically label the multimodal data during lunch).

5 DESIGN OF COSMO

5.1 Contrastive Fusion Learning

In the first stage, we develop a new contrastive fusion learning approach to learn representations that capture consistent information shared between multiple sensory data, only from the unlabeled multimodal data. We first introduce the main components in our representation learning framework, and then present the loss function of contrastive fusion learning.

5.1.1 Representation Learning Framework. In real-world multimodal HAR applications, the modality and dimension of different sensory data can be very diverse. For example, during one time slot, the IMU sensor data usually has two dimensions (i.e., time and 9-dimension vectors) while the depth video has three dimensions (i.e., time and 2D depth images). Therefore, we will first use different feature encoders to extract deep unimodal features from different modalities, and then transform the features to the same dimension through the projection networks. During contrastive learning on the unlabeled multimodal data, we aim to train the feature encoders to generate rich representations that are consistent among different modalities. However, directly contrasting the features from heterogeneous sensor modalities will fail to fully leverage multimodal synergies [33]. Therefore, we propose a novel fusion-based feature augmentation module, which extensively augments the unimodal features to a group of fused features via weighted sum or concatenation. Next, we introduce the three main components of our contrastive learning framework respectively.

Unimodal Feature Encoders. As the data of different modalities can be very heterogeneous (e.g., IMU data and depth videos), we use different feature encoders to extract the unimodal representations. Suppose that we have N unlabeled multimodal data samples $\mathbf{x} = \{\mathbf{x}^i, \forall i = 1, ..., N\}$, where $\mathbf{x}^i = \{\mathbf{x}_1^i, \mathbf{x}_2^i, ..., \mathbf{x}_M^i\}$ contains M ($M \geq 2$) different modalities. Then $\mathbf{x}_j^i$ denotes the data of the j_{th} modality in the i_{th} multimodal sample. Next we use the i_{th} multimodal training sample $\mathbf{x}^i$ as an example to introduce the function of each operation. First, the data of different modalities will be separately fed into M unimodal feature encoders $(f_{enc1}(\cdot), ..., f_{encM}(\cdot))$ to generate M representation vectors:

$$\mathbf{z}_j^i = Flatten(f_{encj}(\mathbf{x}_j^i)), j = 1, ..., M, \tag{1}$$

where $\mathbf{z}_j^i \in \mathbb{R}^{D_j}$ is the flattened one-dimensional vector of the extracted feature from the j_{th} sensor modality with a dimension D_j. The flatten operation will make it easier to fuse heterogeneous features from different modalities (which may have two or three dimensions such as IMU data and depth videos). Moreover, the unimodal feature encoder can be any off-the-shelf deep learning models (e.g., convolutional neural network [37] or recurrent neural network [56]) depending on the sensor modalities. This implies that our framework is general and flexible to various applications.

Feature Projection Networks. As the features of different modalities may have different dimensions, the projection networks $(h_1(\cdot), ..., h_M(\cdot))$ will then map the unimodal features to the same dimension using multi-layer perceptions:

$$\mathbf{r}_j^i = Norm(h_j(\mathbf{z}_j^i)), j = 1, ..., M. \tag{2}$$

This will ensure that the output features have the same dimension D and lie on the unit hypersphere after normalization, i.e., $\{\mathbf{r}_j^i, \forall j = 1, ..., M\} \in \mathbb{R}^D$ (D is usually set to be 128 in our experiments). Therefore, the projection network can unify the heterogeneous inputs from different modalities for the next-step fusion operations (e.g., summation or concatenation). As in other self-supervised learning approaches [38, 48], we will discard the projection networks at the end of contrastive learning and only keep the feature encoders in the second stage of supervised learning.

Fusion-based Feature Augmentation. Based on the projected features $\{\mathbf{r}_j^i, \forall j = 1, ..., M\}$ from sample $\mathbf{x}^i$, we now augment a group of positive features for the subsequent contrast learning. Instead of directly contrasting features of M different modalities [19, 48], in Cosmo, we propose to contrast various fused features of different modalities to extract consistent information for efficient fusion. Specifically, we randomly generate P fusion-based feature augmentations as $\{\mathbf{v}_k^i, \forall k = 1, ..., P\}$ from sample $\mathbf{x}^i$. Each of them represents a different fusion combination of the sensor features and contains some subset of information in the original data sample. Here P is the total number of augmented features, which is independent of the number of sensor modalities M. For example, if there are three modalities to be fused, we can augment ten fused features with different combinations of them. Specifically, the k_{th} fused feature augmentation of the i_{th} data sample is given by:

$$\mathbf{v}_k^i = Aug(\mathbf{r}_1^i, ..., \mathbf{r}_M^i) = \sum_{j=1}^{M} \alpha_{jk} \mathbf{r}_j^i, k = 1, ..., P, \tag{3}$$

where $\alpha_{1k}, ..., \alpha_{Mk} \in [0, 1]$ are randomly sampled and $\sum_{j=1}^{M} \alpha_{jk} = 1$. As shown in Figure 6, the augmented features represent different weighted combinations of the sensor modalities. The weighted combination of features from different modalities is a representative method for dynamic fusion [36]. Through contrastive learning among these features in a unified fusion space, the feature encoders will generate features invariant to different fusion schemes. Moreover, the augmentation of features can be designed according to specific applications and sensor modalities. For example, we may set a larger range of sampling weights (e.g., 0.1-0.9) for more heterogeneous modalities (e.g., depth images and IMU data), while setting a smaller sampling range (e.g., 0.4-0.6) for similar modalities (e.g., acc and gyro data). We then normalize the augmented features to lie on the unit hypersphere, which enables using an inner product to measure distances of features during contrastive learning.

5.1.2 Contrastive Fusion Loss. As mentioned before, we aim to train the feature encoders to generate robust representations that are invariant to different fusion schemes. Therefore, the goal of contrastive fusion learning is to push the augmented features from the same original multimodal sample closer, while separating the augmented features from different original samples. We now introduce the contrastive fusion loss designed to achieve this goal.

Through the above representation learning framework, we will augment a mini-batch of N training samples to $P \times N$ fused features. Let $s \in S \equiv \{1, 2, ..., P \times N\}$ be the index of an arbitrary augmented feature, and let $p \in P(s)$ be the index of the other augmented features originating from the same source sample. Then the feature with index s is called the anchor, and the feature with index p is called positive features. The augmented features from other data samples serve as negative features. Here $P(s)$ is the set of indices of all positive features of s in the minibatch and distinct from s. The contrastive fusion loss can be defined as:

$$\mathcal{L}^{conf} = \sum_{s \in S} \frac{-1}{|P(s)|} \sum_{p \in P(s)} log \frac{exp(\mathbf{v}_s \cdot \mathbf{v}_p / \tau)}{\sum_{a \in S \setminus \{s\}} exp(\mathbf{v}_s \cdot \mathbf{v}_a / \tau)}. \tag{4}$$

Here $\mathbf{v}_s$ is the feature output of the fusion-based augmentation module, and the symbol $\cdot$ denotes the inner product of feature vectors.

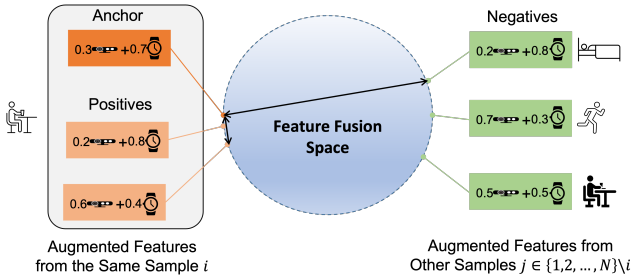


Figure 6: Illustration of fusion-based contrastive learning on the normalized feature space. The positive features are generated by sampling different weighted combinations of modalities from the same multimodal sample, while the negatives are augmented from the remaining multimodal samples in the batch. The contrastive fusion loss contrasts the positives to be closer to each other and pushes away the negative features.

Parameter $\tau \in \mathbb{R}^+$ represents the temperature used to adjust the impact of different samples [49], where an appropriate temperature can help to improve the performance of representation learning (we use $\tau = 0.07$ as in previous contrastive learning studies). Note that for each anchor s, there are $|P(s)| = P - 1$ positive pairs and $P \times (N - 1)$ negative pairs. The denominator has $P \times (N - 1)$ terms (including the positive and negatives). Therefore, minimizing the contrastive fusion loss in Eqn. 4 will bring the positive features together and push the negatives apart.

Learning Consistent Information. As shown in Figure 6, the designed contrastive fusion loss will contrast the set of all fused features from the same multimodal sample as positives, while against the negatives from the remaining samples of the batch. This will result in an embedding space, where features with different fusion schemes from the same multimodal samples will be more closely aligned. Moreover, minimizing the contrastive loss will maximize the lower bound of mutual information among positive features [48]. Therefore, as introduced in Section 3, the feature encoders will learn more consistent information that extracts the common features from the unlabeled multimodal data.

The advantages of our contrastive fusion learning design are as follows. First, the feature encoders will be trained to generate unimodal features invariant to different fusion schemes, which will enable good adaptation performance for the attention-based fusion in Section 5.2. For example, in the fusion of depth videos and motion data, the attention module may give a higher weight to depth videos for the activity of "falling" while a lower one for "running". In this case, the features are robust to be fused with different attention schemes. Second, our experiments show that based on the pre-trained feature encoders, the fusion learning on limited labeled multimodal data will converge much faster than conventional supervised fusion learning. Third, the learned representations will contain more consistent information across different modalities that is robust to noisy multimodal data. For example, the data of the activity "having a lunch" from a single modality like the IMU or the depth can be very noisy, while it will be easier to recognize through combining the common features of the two modalities.

5.2 Iterative Fusion Learning

In the second stage, we aim to explore the complementary information of different sensor modalities from the limited labeled data, and carefully combine it with the consistent information learned on the unlabeled multimodal data. With such an approach, the trained model will be not only robust to noisy multimodal data but also can leverage the strength of different sensors to improve the fusion performance. We first introduce a novel quality-guided attention module for feature fusion with limited labeled data. Then we propose an iterative fusion learning approach to effectively combine the consistent and complementary information.

5.2.1 Quality-guided Attention Fusion. In multimodal human activity recognition, different sensors may carry different amounts of information to a specific task, due to various reasons such as the type of the sensing signal, the quality of hardware, the device placement, as well as the ambient noise and settings [11, 54]. An ideal activity recognition approach should be able to capture the variance in both data quality and contributions among different sensor modalities, and rely on more informative ones to promote the fusion performance. To this end, we design a quality guided attention fusion module to dynamically give different attentions to the sensor features generated by the feature encoder networks.

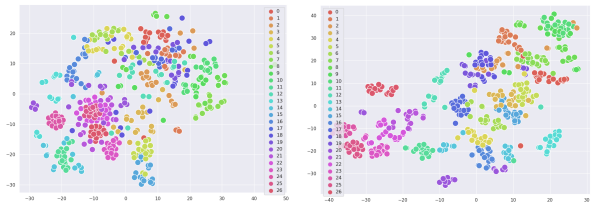
Specifically, the unimodal feature encoders will be initialized using the model weights trained in the first stage, which will generate deep features $(\mathbf{z}_1^i, ..., \mathbf{z}_M^i)$ from different modalities. Then we will use a soft-fusion-based attention module [11] to capture the complementary information of different modalities. The fusion-based attention structure can be formalized as follows:

$$\mu_\mathbf{j} = tanh(\mathbf{W} \cdot \mathbf{z_j} + \mathbf{b}), \tag{5}$$

$$\beta(Attn)_j = \frac{exp(\mu_\mathbf{j} \cdot \mathbf{z_j})}{\sum_j exp(\mu_\mathbf{j} \cdot \mathbf{z_j})}, j = 1, ..., M. \tag{6}$$

Here we use multi-layer perceptions to get $\mu_\mathbf{j}$ from $\mathbf{z_j}$. And $\beta(Attn)_1$, ..., $\beta(Attn)_M$ are weights of different modalities generated by the attention module. Although the attention module can capture the strength of different modalities, it may not be effective when the labeled multimodal data is limited. In particular, when the attention module is trained with only a very small amount of labeled data, the generated weights may include interfering noises and dynamics, which will significantly affect the fusion performance. For example, the depth images generally contain more useful information than IMU data in HAR tasks, while the attention module may give a lower weight to depth if trained from a very small amount of labeled data. Therefore, besides the weights learned from limited labeled data, we also incorporate another set of fusion weights by evaluating the data quality from large amounts of unlabeled data. Specifically, we exploit the clusterability of unlabeled data to assess the quality of a single modality among all modalities. The rationale is that the clusterability of latent spaces is strongly correlated with their resulting classification accuracy [7, 21]. Therefore, when the data of one modality has a larger tendency to be clustered, it will contribute more useful information to a classification task.

We quantify the clustering tendency of unlabeled unimodal data using the Hopkins statistic [15], which is a statistical metric between 0 and 1. A higher Hopkins statistic means stronger data clusterability. Moreover, to further reduce the impact of other dynamic factors (e.g., subject variance, environments) on the clusterability measurement, we also measure the absolute difference between the number of clusters and the ground truth. Specifically, the number of clusters is obtained by the one that has the highest Silhouette score (a metric indicating the goodness of a clustering result) [44] after Kmeans clustering [20]. Suppose that the Hopkins statistic of the j^{th} modality is $H(\mathbf{x}_j)$, and the absolute value of difference between the optimal number of clusters and the number of classes is c_j, then the quality of the j_{th} modality is measured by $q_j = H(\mathbf{x}_j)/c_j$. Then the quality weight of the j^{th} modality is given by normalizing the quality index among all modalities as $\beta(QoM)_j = q_j / \sum_{j=1}^{M} q_j$.



(a) IMU ($q_1 = 0.26$). **(b) Depth ($q_2 = 0.74$).**

Figure 7: Measuring data quality from unlabeled data. $q_j (j = 1, 2)$ is the calculated quality weight. Compared with IMU, the depth data has a higher clusterability (0.8513), and the optimal number of clusters (24) is closer to the number of total classes (27).

Figure 7 shows an example of data quality measurement on the unlabeled multimodal data from the UTD dataset (see Section 6), where the data is visualized using t-SNE. The task here is to recognize 27 activities using IMU and depth data. Compared with IMU, the depth data has a higher clusterability (0.8513), and the optimal number of clusters (24) is closer to the number of total classes (27). This means the depth data has better quality than IMU in the HAR task. As a result, the calculated quality weight for depth is 0.74, which is larger than that of IMU (0.26).

We then combine the weights generated by the attention module and the quality weights calculated using the unlabeled data as:

$$\beta_j = (1 - \lambda)\beta(Attn)_j + \lambda\beta(QoM)_j. \tag{7}$$

Here λ is the hyper-parameter to adjust the impact of quality-based weights, which can be tuned according to different datasets and settings. For example, when the labeled data is very limited and noisy, we may give more confidence to the large amount of unlabeled multimodal data and thus choose a larger λ. Next we normalize the combined weights β_j among the M modalities using $\beta_j = \beta_j / \sum_{j=1}^{M} \beta_j$. Then based on the quality-guided attention weights $(\beta_1, ..., \beta_M)$, we have the following two types of fusion mechanisms for applications with different sensor modalities.

- For sensors of similar modalities (e.g., motion sensors like accelerator and gyroscope) and when the unimodal features have the same dimension, we use weighted sum to estimate the contribution of sensors:

$$\mathbf{v}^i = SumAttn(\mathbf{z}_1^i, ..., \mathbf{z}_M^i) = \sum_{j=1}^{M} \beta_j \mathbf{z}_j^i \tag{8}$$

- For sensors of extremely diverse modalities (e.g., the depth camera and IMU) or when the unimodal features have different dimensions, we use weighted concatenation to selectively leverage the extracted features:

$$\mathbf{v}^i = ConcateAttn(\mathbf{z}_1^i, ..., \mathbf{z}_M^i) = [\beta_1 \mathbf{z}_1^i, ..., \beta_M \mathbf{z}_M^i] \quad (9)$$

Through combining the weight of quality calculated from unlabeled multimodal data, the attention-based fusion module can yield more robust weighted fusion. As a result, features from each modality can offer different strengths for the HAR task. For example, in the fusion of depth cameras and wearables, the attention module would give higher weights to depth features for the activity of "falling" while giving more attention to IMU features for "running". In this way, the model can leverage the complementary nature of sensor modalities to promote the fusion performance even trained with limited labeled multimodal data. In the implementation, the data quality $\beta(QoM)_j$ is measured using a large amount of unlabeled multimodal data on the cloud and then sent to the edge, while the attention module is trained using limited labeled multimodal data on the edge to generate the attention weights $\beta(Attn)_j$. Then the two sets of weights are combined using Equ.(7). Therefore, the quality-guided attention mechanism will not introduce high overhead in the supervised training on the edge, which is also demonstrated in our experimental results in Section 6.6.

5.2.2 Iterative Training between Encoders and Classifiers. As discussed in Section 3, the consistent and complementary information should be utilized simultaneously during the fusion process. Through contrasting the fused features, the feature encoders are trained to capture consistent information across the sensor modalities. Thus the output representations will actually filter out valuable complementary information, which impedes us from exploring strengths of different sensors. Moreover, if we simply train the encoders and classifier together, the whole model may forget the consistent knowledge inherited in the pre-trained feature encoders and degrades to conventional supervised fusion learning. Therefore, given the feature encoders pre-trained in the first stage, the challenge is how to incorporate it with the attention-based classifier in the supervised learning stage.

To tackle this challenge, Cosmo features a novel iterative learning approach to effectively combine the complementary and consistent nature of different sensor modalities based on limited labeled data. As shown in Figure 8, the objective of this iterative training is to explore complementary information from labeled multimodal data while avoiding overfitting on sensor-specific features. Specifically, we first initialize the feature encoders using the model weights trained in the first stage (discarding the projection network and augmentation modules), and randomly initialize the classifier. Second, the feature encoders will be fine-tuned for T_{iter} epochs with the classifier fixed, where the hyper-parameter T_{iter} denotes the epochs of iterative training. Third, it rolls over to train the classifier for T_{iter} epochs with the encoders fixed to balance the consistent and complementary information. This procedure will run until the preset epoch number is reached.

Figure 9 shows the performance comparison with or without iterative fusion learning during the supervised learning process. We can see that with the iterative training, the training loss curve

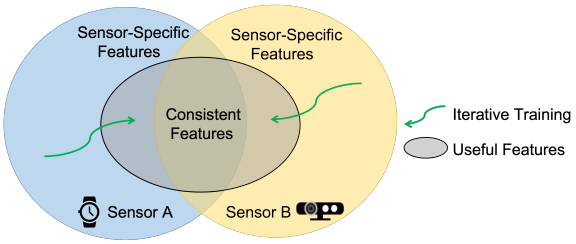
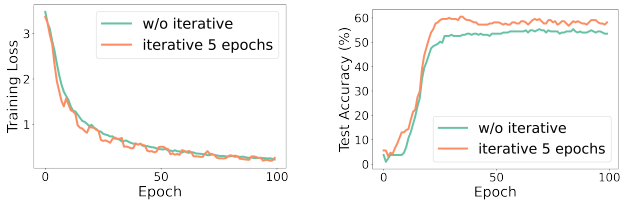


Figure 8: Illustration of iterative fusion learning. Through iterative training between the feature encoders and attention-based classifier, the complementary information from labeled data will be gradually added without forgetting the consistent features.



(a) Training Loss. **(b) Testing Accuracy**

Figure 9: Training loss and testing accuracy during supervised training with or without iterative training.

exhibits a square-wave behavior in every T_{iter} epochs. This is because the model is trained to gradually add the complementary information, which leads to a better test accuracy, as shown in Figure 9b. Moreover, this mechanism will not incur more training delays than conventional supervised training, as the total number of training epochs is the same. In our experiments, the choice of different iteration epochs T_{iter} will slightly affect the performance of learned models, which varies among different datasets and settings. Therefore in real-world implementations of Cosmo, we can use a small validation dataset to determine the optimal iterative epochs.

5.3 System Implementation

5.3.1 Cloud-edge implementation. The two-stage training strategy of Cosmo naturally enables a could-edge implementation architecture. The first stage can be trained on the cloud with unlabeled multimodal data gathered from multiple users or public datasets. The second stage can be trained locally with a small amount of labeled data, which only incurs low compute overhead. We implement Cosmo with Python and PyTorch[1]. The first stage of Cosmo is run on a server with 8 NVIDIA GEFORE TITAN Xp GPUs, 256 GB RAM, and two 16-core Intel Xeon E5-2620 (2.10GHz) CPUs. The second stage is run on NVIDIA Jetson TX2 [2].

5.3.2 Baselines. In our experiments, we compare the performance of Cosmo with the following baselines.

SingleModal, which predicts human activities using labeled data of a single modality.

***DeepSense*[55],** which is one of the state-of-the-art supervised learning approaches for multimodal HAR, where features from different sensor modalities are concatenated together for fusion.

[1]The code is available at https://github.com/xmouyang/Cosmo.

***AttnSense* [36],** which is an attention-based supervised multimodal fusion approach that dynamically assigns weights to features of different modalities before fusion.

Contrastive Predictive Coding (CPC) [19], which is a state-of-the-art contrastive learning approach designed for HAR with IMU data. CPC performs unsupervised learning by capturing the temporal structure of sensor data.

Contrastive Multi-view Learning (CMC) [48], which is a state-of-the-art contrastive learning approach designed for multi-view (e.g., RGB, depth) computer vision tasks. This approach trains the feature encoders by directly contrasting features of different modalities from the unlabeled data.

5.3.3 Configurations. For the feature encoders, we use CNN layers to extract deep features and plus RNN layers to capture the time-series properties. Here 2D-CNN is used for inertial data and 3D-CNN is used for skeleton, depth and radar data, and multi-layer perceptions are used for the classifier. The learning rate and batch size are set as 0.01 and 64 for contrastive learning and as 0.001 and 16 for supervised learning. Each experiment is repeated five times. To ensure a fair evaluation, all baseline models are trained with the same hyper-parameters and labeled set. Cosmo, CPC and CMC are trained with the same unlabeled set.

6 EVALUATION

6.1 Datasets

We evaluate the performance of Cosmo using two public datasets and a new multimodal HAR dataset collected by ourselves in real-world settings[2]. Table 1 shows the summary of the three datasets. The reasons for using the datasets are as follows. The USC dataset [59] has a large number of samples, which can be used to evaluate performance with various settings of labeled and unlabeled data. The UTD dataset [10] is a commonly used multimodal dataset that collects data of 27 various activities. Compared with the USC dataset, our dataset provides a richer combination of modalities, including RGB, depth, IMU, and mmWave radar. In addition, our dataset involves higher user-level feature diversity (e.g., 30 subjects) than the UTD dataset (only 864 samples from 8 subjects).

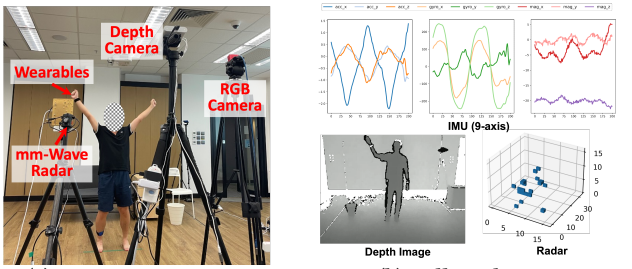
USC dataset [59]. This dataset comprises data of a 3-axis accelerator and 3-axis gyroscope from 14 users performing 12 activities (e.g., sitting, walking). The sampling rate of the two sensors is 100 Hz. We choose a 2-second time window that generates a 600-dimensional vector for data of each modality. We use data from 10 subjects for training (split as labeled and unlabeled data) and data from the other four subjects for testing. Although the two sensors have a relatively small modality gap, they still have complementary properties in HAR tasks, as shown in Section 3.

UTD dataset [10]. This dataset contains data of 27 actions (e.g., waving, arms crossing) collected from 8 subjects, and four modalities (RGB, depth, skeleton, and inertial measurements). Due to the small amount of data (864 samples in total), in this paper, we use the skeleton and IMU data with a relatively low dimension compared with RGB and depth to avoid overfitting on the training set. The skeleton contains 3D positions of 20 joints, and the inertial measurements include 3-axis acceleration and 3-axis rotation signals.

[2]All the data collection was approved by IRB of the authors' institution.

Dataset	Modality	Activity	Subject	Samples
USC	Acc, Gyro	12	14	38312
UTD	IMU, Skeleton	27	8	864
Self-Collected	IMU, Depth, Radar	14	30	3434

Table 1: Summary of the three multimodal datasets.



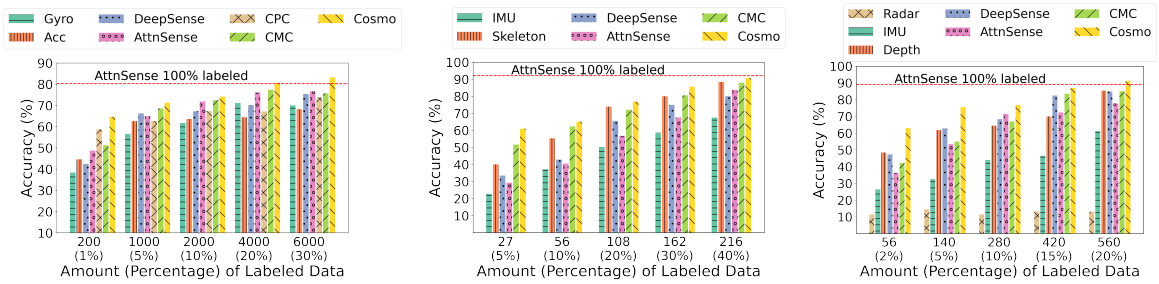
(a) Experiment Setting. **(b) Collected Data.**

Figure 10: The setting and data of the multimodal HAR dataset collected by ourselves.

The sampling rate of the skeleton and inertial data is 30 Hz and 50 Hz. The data from six subjects is used for training, and data from the other two subjects is used for testing.

Self-collected multimodal HAR dataset. We collect a real-world multimodal HAR dataset in a sitting room setting as shown in Figure 10a. Our new dataset includes data from four different modalities (RGB, depth, IMU, and mmWave radar), 30 subjects and 14 activities. The activities include: sitting, pacing, lying, jumping, throwing, picking up, rummaging, stomping, hand-waving, falling, squatting, kicking, hand shaking, and thumping the cabinet. Some activities like "hand shaking" involve two people. During the recording of the dataset, we instructed the subjects to freely perform these activities, where we did not fix the order or frequency of performing such activities. We note that it is extremely challenging to collect multimodal HAR datasets under fully uncontrolled settings. One of the major challenges in collecting such data is that sensors (e.g., depth/radar) have very limited coverage [10, 45]. For example, in order to capture the subjects in the range of the sensors, we can only collect short-term HAR data with limited types of events in typical living environments (e.g., a bedroom), which is consistent with our settings. The total period of collected raw data from 30 subjects is about 20 hours, where the sampling rate of RGB-D camera, IMU, and radar is 20Hz, 100Hz, and 15 Hz, respectively. We synchronize the data of four modalities through a uniform global timestamp and label the data using the RGB images[3]. Then we removed incomplete modalities or invalid activity data, and split it into 2-second recordings, which follows the input settings of existing deep learning-based HAR approaches [19, 47, 53]. Moreover, we preprocess the data to a fixed dimension [40, 480, 640], [200, 9] and [30, 2, 16, 32, 16] for IMU, depth and radar, respectively. We finally obtained 3,434 snippets multimodal samples after data cleaning and pre-processing. Figure 10b visualizes an example of the collected multimodal data, where the data from different modalities is very heterogeneous, making the fusion very challenging in HAR tasks. We use data from 25 subjects for training and data from the remaining five subjects for testing.

[3]Note that we use the RGB images only for labeling but do not in the evaluation, as it is privacy-sensitive for HAR applications.



(a) USC **(b) UTD** **(c) Self-collected**

Figure 11: Accuracy comparison with different amounts of labeled data. Cosmo consistently outperforms other baselines, and can achieve comparable accuracy of AttnSense (with 100% labeled data) with only a small portion of labeled data.

Dataset	USC	UTD	Self-collected
Labeling rate	1%	5%	2%
Labeled training data	200	27	56
Testing data	2000	216	560
SingleModal	0.4464 / 0.3839	0.2279 / 0.4008	0.2643 / 0.4849 / 0.1142
DeepSense	0.4243	0.3348	0.4726
AttnSense	0.4872	0.2930	0.3631
CMC	0.5125	0.5062	0.4214
Cosmo	**0.6450**	**0.6093**	**0.6304**

Table 2: Accuracy comparison on three real-world multimodal datasets with limited labeled data.

6.2 Accuracy on Different Datasets

6.2.1 Overall performance on limited labeled data. We first evaluate the accuracy performance of different approaches on the three datasets with limited labeled data[4]. Table 2 shows the mean accuracy with 1%, 5%, and 2% labeling rate for the USC, UTD, and self-collected dataset, respectively. The labeling rate denotes the percentage of labeled multimodal samples in all training samples. In our experiments, the number of labeled training samples is very small (i.e., 200, 27, and 56 for the three datasets, respectively). This setting reflects the practical challenge that the labels of activity data are hard to obtain in real-world HAR applications. First, *SingleModal* on all HAR tasks has extremely low accuracy due to the limited labeled data. *Deepsense* and *AttnSense* provide only slightly accuracy improvement or even perform worse than *SingleModal* (e.g., in UTD dataset, 0.2930 mean accuracy for *AttnSense* while 0.4008 for Skeleton-only). Moreover, the accuracy improvement of CMC is limited as it does not fully leverage multimodal synergies. Cosmo shows significant accuracy improvement with limited labeled multimodal data, e.g., outperforms by 38.14%, 20.93 %, 27.45%, 31.63 %, and 10.31% over IMU-only, Skelton only, *DeepSense*, *AttnSense*, and CMC on the UTD dataset, respectively.

6.2.2 Different amounts of labeled data. We then compare the accuracy performance with different amounts of labeled data, where we fix the amount of total training data. The results are shown in Figure 11. First, all approaches tend to have a higher recognition accuracy with more labeled data, and Cosmo consistently outperforms other baselines in various settings. Second, the performance improvement of Cosmo is more significant when with a very small amount of labeled data. For example, for the self-collected dataset

[4]As shown in Section 6.1, the training and testing data comes from different subjects.

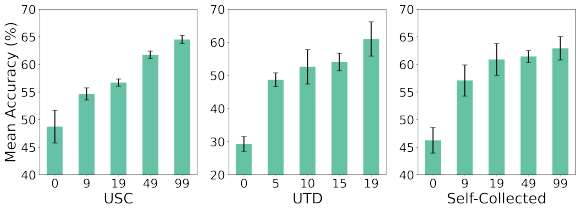


Figure 12: Accuracy of Cosmo with different amounts of unlabeled data. Values in X-axis are $\frac{\#Unlabeled\ data}{\#Labeled\ data}$.

Figure 13: Accuracy with labeled data from target subjects.

with only 56 labeled training samples, Cosmo outperforms *SingleModal*, *DeepSense*, *AttnSense*, and CMC by 51.61%, 15.77%, 26.73%, and 20.90%, respectively. Third, compared with CPC on the USC dataset, Cosmo is more efficient in supervised learning when adding more labeled samples. For example, the accuracy improvement of Cosmo over CPC increases from 5.94% to 13.45% when the labeling rate changes from 1% to 20%. Moreover, Cosmo can achieve comparable accuracy of AttnSense with all labeled data (the red line in Figure 11), when trained with unlabeled and only a small portion of labeled data, e.g., 20%, 40% and 20% labeled data for the USC, UTD and self-collected dataset, respectively.

6.2.3 Different amounts of unlabeled data. We then fix the number of labeled samples (the same numbers in Table 2) and evaluate the performance of Cosmo with different amounts of unlabeled training data. The results are shown in Figure 12. We observe that the accuracy of Cosmo increases with a larger amount of unlabeled data, which means that Cosmo is able to effectively explore useful knowledge for fusion from unlabeled multimodal data.

6.2.4 Performance with labeled data from target subjects. In Section 6.2.1 to 6.2.3, the labeled training data and testing data come from different subjects, which may contain large domain gaps. Figure 13 compares the accuracy of Cosmo when the second stage is trained with labeled data from testing subjects (w. target) or from other

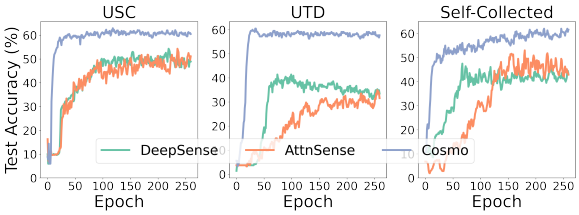
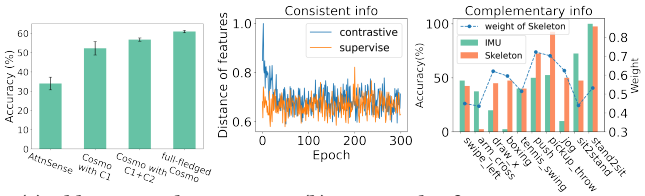


Figure 14: Convergence comparison on limited labeled data. Cosmo converges faster than supervised learning baselines.



(a) Ablation Study. **(b) Extracted Information.**

Figure 15: Understanding Cosmo's Performance.

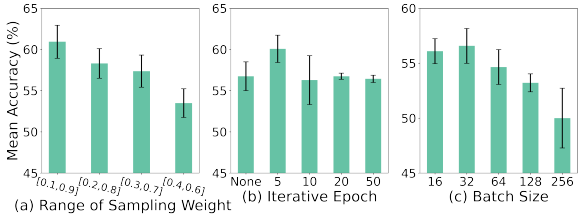


Figure 16: Accuracy of Cosmo under different settings.

6.3 Convergence Performance

In the cloud-edge architecture of Cosmo, the second stage is trained on edge using labeled multimodal data. As the edge devices usually own limited computation resources, the supervised training should be more efficient. Here we investigate the convergence performance of Cosmo on the three multimodal datasets when training with limited labeled data (the same setting as 6.2.1).

The results are shown in Figure 14. Compared with the supervised learning-based fusion approaches *DeepSense* and *AttnSense*, Cosmo converges much faster and only needs about 30 to 50 epochs of training to achieve the highest testing accuracy. The reason is that Cosmo is trained based on the feature encoders pre-trained on the large amounts of unlabeled multimodal data and features an iterative training strategy in the second stage, which enables fast adaptations on the limited labeled data.

6.4 Understanding Cosmo's Performance

In this section, we perform ablation study and show the intermediate results to undertand the effectiveness of Cosmo. We show the testing results of the UTD dataset with 5% labeling rate, since the results on the other two datasets are similar.

6.4.1 Ablation study. Figure 15a compares the mean accuracy with different design components of Cosmo. Here we evaluate the impact of three components, including contrastive fusion learning (denoted as C1), iterative fusion learning (denoted as C2) and quality-guided attention modules (denoted as C3). Compared with AttnSense, the results with different components of Cosmo all show significant accuracy improvement (i.e., over 20%). Moreover, The mean accuracy increases with more components added, which validates the effectiveness of different components in Cosmo's design.

6.4.2 Extracted information. We then show the intermediate results generated by Cosmo in Figure 15b. In the left figure, we plot the normalized mean distance between IMU and skeleton features generated by Cosmo, which decreases during contrastive fusion learning and remains almost unchanged during supervised learning. Moreover, the mean distance between features of two modalities generated by Cosmo (0.7028) is smaller than that learned by SingleModal (0.7981), which means Cosmo can learn more consistent information among different modalities [16, 60]. In the right figure, we plot the mean fusion weights of skeleton across different activities generated by the quality-guided attention module of Cosmo. We observe that when the accuracy of skeleton for recognizing one activity is larger than IMU, the weight of skeleton will be higher (e.g., 0.722 for the activity "push") and vice versa. This shows that Cosmo can capture the complementary information of different modalities through leveraging different strengths of modalities.

6.5 Micro-benchmark Performance

In this section, we perform sensitivity analysis of Cosmo under various system settings. The data settings are the same as Sec. 6.4.

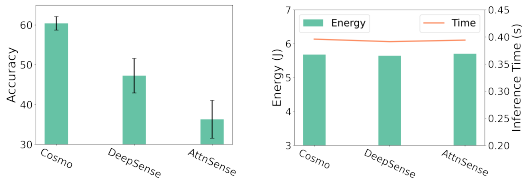
6.5.1 Different weight sampling schemes. We first evaluate the accuracy with different sampling ranges of IMU weight in fusion-based feature augmentation. The weight of skeleton is one minus the sampled IMU weight. The results are shown in Figure 16(a). We observe that generally, the accuracy decreases with a smaller range of sampling weights (e.g., 60.93% for sampling in [0.1-0.9] v.s. 53.48% for sampling in [0.4-0.6]). This means that by pushing the positive features with more biased fusion together, the feature encoder is able to learn representations that contain more consistent information, which finally improves fusion performance.

6.5.2 Different iterative epochs. In this experiment, we study the impact of the iterative epoch T_{iter} (in iterative fusion learning) on the accuracy performance. The results are shown in Figure 16(b). We observe that the setting of $T_{iter} = 5$ will result in the best accuracy performance, compared with other values as well as that of directly supervised training (i.e., 0 epoch). Furthermore, we found that the optimal iterative epoch varies with different datasets. Therefore in real-world implementations of Cosmo, we can use a small validation dataset to determine the optimal iterative epochs.

6.5.3 Different batch sizes in contrastive learning. We further investigate the performance of Cosmo with different batch sizes in contrastive fusion learning. Most of the previous contrastive learning studies benefit from a larger batch size during training [48, 49]. However, as shown in Figure 16(c), Cosmo achieves good performance even with a small batch size (i.e., 16 or 32). The reason is that

Approach	Label rate	Cosmo Stage 1	Cosmo Stage 2	*DeepSense*	*AttnSense*
Time (min)	1%	101.19	**25.93**	38.38	62.52
	2%	90.13	49.87	74.42	120.98
Energy (KJ)	1%	7286.07	**21.34**	30.99	51.71
	2%	6489.70	40.35	60.90	98.58

Table 3: Training overhead. The first stage of Cosmo is trained on the server. The others are trained on Jetson TX2.



(a) Accuracy. **(b) Energy and time.**

Figure 17: Inference performance on Jetson TX2.

Cosmo augments multiple positive features (e.g., nine positives for the range of [0.1, 0.9]) from the same multimodal sample. Therefore, Cosmo can learn richer representations from the fusion-based augmented features. Moreover, the accuracy slightly decreases with larger batch sizes, which may be due to that more diverse negative samples increases the difficulty of convergence.

6.6 System Overhead

In this section, we evaluate the training and inference overhead of Cosmo using the multimodal dataset collected by ourselves. We run contrastive fusion learning of Cosmo on a server, and run the supervised learning of Cosmo and baselines on NVIDIA Jetson TX2.

6.6.1 Training Overhead. Table 3 shows the training performance of different approaches. For the supervised training on edge, Cosmo (stage 2) needs less time and power consumption (e.g., about 26 minutes for training with 1% labeled multimodal data) compared with *Deepsense* and *AttnSense*. The reason is that based on the pre-trained feature encoders in contrastive learning, Cosmo converges much faster (i.e., only 50 epochs) on the labeled multimodal data. Therefore, Cosmo on edge only incurs low computation overhead and hence is affordable for resource-limited edge devices.

6.6.2 Inference Performance. Figure 17 shows the results of inference performance with 2% labeled data, where the time and power is the mean value for inferring one sample (2-second multimodal readings) on Nvidia Jetson TX2. We observe that Cosmo has a similar inference overhead as *Deepsense* and *AttnSense*. Specifically, the inference time of Cosmo for predicting with a 2-second sample including Depth, IMU, and Radar data is less than 0.4s, which can achieve multimodal HAR in real-time. Moreover, Cosmo has a higher accuracy performance (i.e., 60.42%) than *Deepsense* (i.e., 47.26%) and *AttnSense* (i.e., 36.31%) on the noisy multimodal data.

7 DISCUSSION

Impact of practical factors. Here we discuss the impact of several practical factors on the performance of Cosmo. First, in real-world applications where the sensor data is continuously collected, we can simply remove the irrelevant data through some filtering techniques [14, 43]. Then the data can be split using sliding windows (e.g., one or two seconds). In this case, Cosmo can be adopted by feeding the split multimodal data into the deep neural networks. Second, for recognizing the activities with different durations, Cosmo can incorporate exiting techniques [30] such as adaptive time windows according to the characteristics of different activities. Third, to handle the misalignment of multimodal data introduced by time synchronized errors of sensors, Cosmo can leverage the consistent information across the features of different modalities to align the multimodal data stream. For example, we can maximize the correlation of unimodal features to find the optimal alignment. Lastly, to adapt to different environmental conditions, the quality-guided attention module in Cosmo can dynamically assign weights to different modalities in the presence of environmental variations. For example, when the subjects are not in the scene of the depth camera, a larger weight will be given to the IMU data from the smartwatch.

Future work. Here we discuss several extensions of this work. First, we will investigate the efficient end-to-end implementation of Cosmo. For example, we will study how to cache data from different sensors to enable faster convergence for the training of the second stage in Cosmo. Moreover, we will study how to reduce the inference overhead of Cosmo through dynamic sensor selection [27]. For example, if the fusion weight of one sensor is smaller than a threshold (e.g., 0.1) for a long period, it means this sensor does not contribute much to the specific HAR task. We will then discard its sensor data to reduce the inference latency while maintaining the accuracy performance. Second, we will study how to incorporate the federated learning paradigm [39, 50] to further improve the performance of Cosmo while protecting user privacy. For example, each node in federated learning will run the first stage of Cosmo using unlabeled multimodal data. Then the trained feature encoders can be transferred to the server for aggregation using federated unsupervised learning techniques [57]. After federated training of the feature encoders, the node can run the second stage of Cosmo using limited labeled data to train the classifier.

8 CONCLUSION

This paper proposes Cosmo, a new system for contrastive fusion learning with small data in multimodal HAR applications. Cosmo features a novel two-stage training strategy that can efficiently extract consistent and complementary information of different modalities from both unlabeled and limited labeled multimodal data. Extensive experiments show that, Cosmo delivers significant improvement over state-of-the-art baselines in both recognition accuracy and convergence delay.

ACKNOWLEDGMENTS

This work is supported by the Research Grants Council (RGC) of Hong Kong, China, under GRF Grants No. 14203420, the Alzheimer's Drug Discovery Foundation, under Grant RDADB-201906-2019049, the Shenzhen Science and Technology Program (Project JCYJ2021032 4120011032), Guangdong Basic and Applied Basic Research Foundation (Project 2021B1515120008), and the Shenzhen Institute of Artificial Intelligence and Robotics for Society.

REFERENCES

[1] 2017. ALZHEIMER'S DIGITAL BIOMARKERS. https://www.alzdiscovery.org/research-and-grants/funding-opportunities/diagnostics-accelerator-digital-biomarkers-program.

[2] 2022. NVDIA JETSON TX2. https://www.nvidia.com/en-us/autonomous-machines/embedded-systems/jetson-tx2/.

[3] Rebecca Adaimi, Howard Yong, and Edison Thomaz. 2021. Ok Google, What Am I Doing? Acoustic Activity Recognition Bounded by Conversational Assistant Interactions. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5, 1 (2021), 1–24.

[4] Humam Alwassel, Dhruv Mahajan, Bruno Korbar, Lorenzo Torresani, Bernard Ghanem, and Du Tran. 2020. Self-supervised learning by cross-modal audio-video clustering. *Advances in Neural Information Processing Systems* 33 (2020).

[5] 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.

[6] Tara Boroushaki, Isaac Perper, Mergen Nachin, Alberto Rodriguez, and Fadel Adib. 2021. RFusion: Robotic Grasping via RF-Visual Sensing and Learning. In *Proceedings of the 19th ACM Conference on Embedded Networked Sensor Systems*. 192–205.

[7] Deng Cai, Chiyuan Zhang, and Xiaofei He. 2010. Unsupervised feature selection for multi-cluster data. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*. 333–342.

[8] Hong Cai, Belal Korany, Chitra R Karanam, and Yasamin Mostofi. 2020. Teaching rf to sense without rf training measurements. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, 4 (2020), 1–22.

[9] Brian Chen, Andrew Rouditchenko, Kevin Duarte, Hilde Kuehne, Samuel Thomas, Angie Boggust, Rameswar Panda, Brian Kingsbury, Rogerio Feris, David Harwath, et al. 2021. Multimodal clustering networks for self-supervised learning from unlabeled videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 8012–8021.

[10] 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.

[11] Changhao Chen, Stefano Rosa, Yishu Miao, Chris Xiaoxuan Lu, Wei Wu, Andrew Markham, and Niki Trigoni. 2019. Selective sensor fusion for neural visual-inertial odometry. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10542–10551.

[12] Richard Chen, Filip Jankovic, Nikki Marinsek, Luca Foschini, Lampros Kourtis, Alessio Signorini, Melissa Pugh, Jie Shen, Roy Yaari, Vera Maljkovic, et al. 2019. Developing measures of cognitive impairment in the real world from consumer-grade multimodal sensor streams. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2145–2155.

[13] Xiaoran Fan, Longfei Shangguan, Siddharth Rupavatharam, Yanyong Zhang, Jie Xiong, Yunfei Ma, and Richard Howard. 2021. HeadFi: bringing intelligence to all headphones. In *Proceedings of the 27th Annual International Conference on Mobile Computing and Networking*. 147–159.

[14] Rohan Ghosh, Anupam Gupta, Andrei Nakagawa, Alcimar Soares, and Nitish Thakor. 2019. Spatiotemporal filtering for event-based action recognition. *arXiv preprint arXiv:1903.07067* (2019).

[15] Gene Glass and Kenneth Hopkins. 1996. Statistical methods in education and psychology. *Psyccritiques* 41, 12 (1996).

[16] Weili Guan, Haokun Wen, Xuemeng Song, Chung-Hsing Yeh, Xiaojun Chang, and Liqiang Nie. 2021. Multimodal Compatibility Modeling via Exploring the Consistent and Complementary Correlations. In *Proceedings of the 29th ACM International Conference on Multimedia*. 2299–2307.

[17] Guoliang Xing. 2022. Machine Learning Technologies for Advancing Digital Biomarkers for Alzheimer's Disease, Alzheimer's Drug Discovery Foundation. https://www.alzdiscovery.org/research-and-grants/portfolio-details/21130887.

[18] Tian Hao, Guoliang Xing, and Gang Zhou. 2013. iSleep: unobtrusive sleep quality monitoring using smartphones. In *Proceedings of the 11th ACM Conference on Embedded Networked Sensor Systems*. 1–14.

[19] 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.

[20] John A Hartigan and Manchek A Wong. 1979. Algorithm AS 136: A k-means clustering algorithm. *Journal of the royal statistical society. series c (applied statistics)* 28, 1 (1979), 100–108.

[21] Kalun Ho, Franz-Josef Pfreundt, Janis Keuper, and Margret Keuper. 2021. Estimating the Robustness of Classification Models by the Structure of the Learned Feature-Space. *arXiv preprint arXiv:2106.12303* (2021).

[22] Di Hu, Feiping Nie, and Xuelong Li. 2019. Deep multimodal clustering for unsupervised audiovisual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 9248–9257.

[23] Zhizhang Hu, Tong Yu, Yue Zhang, and Shijia Pan. 2020. Fine-grained activities recognition with coarse-grained labeled multi-modal data. In *Adjunct Proceedings of the 2020 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2020 ACM International Symposium on Wearable Computers*. 644–649.

[24] Sinh Huynh, Rajesh Krishna Balan, JeongGil Ko, and Youngki Lee. 2019. VitaMon: measuring heart rate variability using smartphone front camera. In *Proceedings of the 17th Conference on Embedded Networked Sensor Systems*. 1–14.

[25] Wenjun Jiang, Chenglin Miao, Fenglong Ma, Shuochao Yao, Yaqing Wang, Ye Yuan, Hongfei Xue, Chen Song, Xin Ma, Dimitrios Koutsonikolas, et al. 2018. Towards environment independent device free human activity recognition. In *Proceedings of the 24th Annual International Conference on Mobile Computing and Networking*. 289–304.

[26] Lampros C Kourtis, Oliver B Regele, Justin M Wright, and Graham B Jones. 2019. Digital biomarkers for Alzheimer's disease: the mobile/wearable devices opportunity. *NPJ digital medicine* 2, 1 (2019), 1–9.

[27] Clayton Frederick Souza Leite and Yu Xiao. 2021. Optimal sensor channel selection for resource-efficient deep activity recognition. In *Proceedings of the 20th International Conference on Information Processing in Sensor Networks (co-located with CPS-IoT Week 2021)*. 371–383.

[28] Jia Li, Yu Rong, Helen Meng, Zhihui Lu, Timothy Kwok, and Hong Cheng. 2018. Tatc: Predicting alzheimer's disease with actigraphy data. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 509–518.

[29] Tianxing Li, Jin Huang, Erik Risinger, and Deepak Ganesan. 2021. Low-latency speculative inference on distributed multi-modal data streams. In *Proceedings of the 19th Annual International Conference on Mobile Systems, Applications, and Services*. 67–80.

[30] Jonathan Liono, A Kai Qin, and Flora D Salim. 2016. Optimal time window for temporal segmentation of sensor streams in multi-activity recognition. In *Proceedings of the 13th International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services*. 10–19.

[31] Tiantian Liu, Ming Gao, Feng Lin, Chao Wang, Zhongjie Ba, Jinsong Han, Wenyao Xu, and Kui Ren. 2021. Wavoice: A Noise-resistant Multi-modal Speech Recognition System Fusing mmWave and Audio Signals. In *Proceedings of the 19th ACM Conference on Embedded Networked Sensor Systems*. 97–110.

[32] Xiulong Liu, Dongdong Liu, Jiuwu Zhang, Tao Gu, and Keqiu Li. 2021. RFID and camera fusion for recognition of human-object interactions. In *Proceedings of the 27th Annual International Conference on Mobile Computing and Networking*. 296–308.

[33] Yunze Liu, Qingnan Fan, Shanghang Zhang, Hao Dong, Thomas Funkhouser, and Li Yi. 2021. Contrastive multimodal fusion with tupleinfonce. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 754–763.

[34] Yilin Liu, Shijia Zhang, and Mahanth Gowda. 2021. When video meets inertial sensors: zero-shot domain adaptation for finger motion analytics with inertial sensors. In *Proceedings of the International Conference on Internet-of-Things Design and Implementation*. 182–194.

[35] 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.

[36] 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.

[37] Sebastian Münzner, Philip Schmidt, Attila Reiss, Michael Hanselmann, Rainer Stiefelhagen, and Robert Dürichen. 2017. CNN-based sensor fusion techniques for multimodal human activity recognition. In *Proceedings of the 2017 ACM International Symposium on Wearable Computers*. 158–165.

[38] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748* (2018).

[39] Xiaomin Ouyang, Zhiyuan Xie, Jiayu Zhou, Jianwei Huang, and Guoliang Xing. 2021. ClusterFL: a similarity-aware federated learning system for human activity recognition. In *Proceedings of the 19th Annual International Conference on Mobile Systems, Applications, and Services*. 54–66.

[40] Andrew Owens and Alexei A Efros. 2018. Audio-visual scene analysis with self-supervised multisensory features. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 631–648.

[41] Rui Miguel Pascoal, Ana de Almeida, and Rute C Sofia. 2019. Activity recognition in outdoor sports environments: smart data for end-users involving mobile pervasive augmented reality systems. 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*. 446–453.

[42] 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.

[43] 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.

[44] Peter J Rousseeuw. 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics* 20 (1987), 53–65.

[45] Xian Shuai, Yulin Shen, Yi Tang, Shuyao Shi, Luping Ji, and Guoliang Xing. 2021. millieye: A lightweight mmwave radar and camera fusion system for robust object detection. In *Proceedings of the International Conference on Internet-of-Things Design and Implementation*. 145–157.

[46] 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.

[47] Chi Ian Tang, Ignacio Perez-Pozuelo, Dimitris Spathis, Soren Brage, Nick Wareham, and Cecilia Mascolo. 2021. Selfhar: Improving human activity recognition through self-training with unlabeled data. *arXiv preprint arXiv:2102.06073* (2021).

[48] Yonglong Tian, Dilip Krishnan, and Phillip Isola. 2020. Contrastive multiview coding. In *European conference on computer vision*. Springer, 776–794.

[49] Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. 2020. What makes for good views for contrastive learning. *arXiv preprint arXiv:2005.10243* (2020).

[50] Linlin Tu, Xiaomin Ouyang, Jiayu Zhou, Yuze He, and Guoliang Xing. 2021. Feddl: Federated learning via dynamic layer sharing for human activity recognition. In *Proceedings of the 19th ACM Conference on Embedded Networked Sensor Systems*. 15–28.

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

[52] Zhiyuan Xie, Xiaomin Ouyang, Xiaoming Liu, and Guoliang Xing. 2021. UltraDepth: Exposing High-Resolution Texture from Depth Cameras. In *Proceedings of the 19th ACM Conference on Embedded Networked Sensor Systems*. 302–315.

[53] 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.

[54] Hongfei Xue, Wenjun Jiang, Chenglin Miao, Ye Yuan, Fenglong Ma, Xin Ma, Yijiang Wang, Shuochao Yao, Wenyao Xu, Aidong Zhang, et al. 2019. Deepfusion: A deep learning framework for the fusion of heterogeneous sensory data. In *Proceedings of the Twentieth ACM International Symposium on Mobile Ad Hoc Networking and Computing*. 151–160.

[55] 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.

[56] Wojciech Zaremba, Ilya Sutskever, and Oriol Vinyals. 2014. Recurrent neural network regularization. *arXiv preprint arXiv:1409.2329* (2014).

[57] Fengda Zhang, Kun Kuang, Zhaoyang You, Tao Shen, Jun Xiao, Yin Zhang, Chao Wu, Yueting Zhuang, and Xiaolin Li. 2020. Federated unsupervised representation learning. *arXiv preprint arXiv:2010.08982* (2020).

[58] Hanbin Zhang, Gabriel Guo, Chen Song, Chenhan Xu, Kevin Cheung, Jasleen Alexis, Huining Li, Dongmei Li, Kun Wang, and Wenyao Xu. 2020. PDLens: smartphone knows drug effectiveness among Parkinson's via daily-life activity fusion. In *Proceedings of the 26th Annual International Conference on Mobile Computing and Networking*. 1–14.

[59] 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.

[60] Hongyuan Zhu, Jean-Baptiste Weibel, and Shijian Lu. 2016. Discriminative multimodal feature fusion for rgbd indoor scene recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2969–2976.