

Contrastive Learning with Cross-Modal Knowledge Mining for Multimodal Human Activity Recognition

Razvan Brinzea
Maastricht University
Maastricht, Netherlands

Bulat Khaertdinov
Maastricht University
Maastricht, Netherlands

Stylios Asteriadis
Maastricht University
Maastricht, Netherlands

r.brinzea@student.maastrichtuniversity.nl b.khaertdinov@maastrichtuniversity.nl stelios.asteriadis@maastrichtuniversity.nl

Abstract—Human Activity Recognition is a field of research where input data can take many forms. Each of the possible input modalities describes human behaviour in a different way, and each has its own strengths and weaknesses. We explore the hypothesis that leveraging multiple modalities can lead to better recognition. Since manual annotation of input data is expensive and time-consuming, the emphasis is made on self-supervised methods which can learn useful feature representations without any ground truth labels. We extend a number of recent contrastive self-supervised approaches for the task of Human Activity Recognition, leveraging inertial and skeleton data. Furthermore, we propose a flexible, general-purpose framework for performing multimodal self-supervised learning, named Contrastive Multiview Coding with Cross-Modal Knowledge Mining (CMC-CMKM). This framework exploits modality-specific knowledge in order to mitigate the limitations of typical self-supervised frameworks. The extensive experiments on two widely-used datasets demonstrate that the suggested framework significantly outperforms contrastive unimodal and multimodal baselines on different scenarios, including fully-supervised fine-tuning, activity retrieval and semi-supervised learning. Furthermore, it shows performance competitive even compared to supervised methods.

Index Terms—Human Activity Recognition, self-supervised learning, multimodal fusion

I. INTRODUCTION

Human Activity Recognition (HAR) is a joint area of research in the fields of Human-Centered Computing and Human-Computer Interaction, with practical applications in many areas, such as smart homes [1], [2], health monitoring [3], manufacturing automation [4] and sport analytics [5].

The modalities which can be used for HAR include but are not limited to RGB-D streams, skeleton data, wearable sensor data (or inertial data). Different techniques can be employed for HAR depending on the type of the input data, but each modality comes with its own challenges and limitations [6]. Multimodal HAR methods aim to mitigate the shortcomings of unimodal approaches by fusing information extracted from different sources of data [7]. With the breakthrough success of deep learning in the past years, various architectures of deep neural networks have shown impressive performance in multimodal HAR. Nevertheless, they have a common significant

This work has been partially funded by the European Union’s Horizon2020 project: PeRsOnalized Integrated CARE Solution for Elderly facing several short or long term conditions and enabling a better quality of LIFE (Pro-care4Life), under Grant Agreement N.875221.

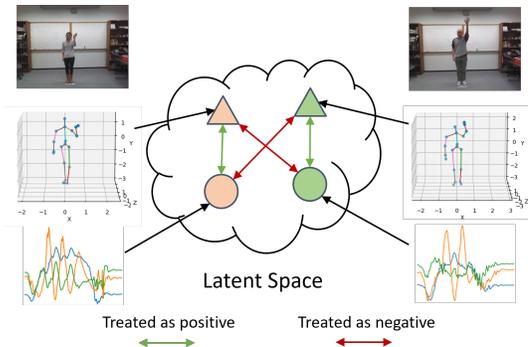


Fig. 1: Illustrated example of a false negative pair in contrastive learning for multimodal HAR. Even though both examples belong to the waving activity class, they are treated as a negative pair, since annotations are not available.

drawback, namely, they require vast amounts of labeled data for training deep models.

Given that labeled data is limited and hard to generate, being able to train a HAR model on unlabeled data could have real practical applications. This is the idea behind the self-supervised learning (SSL) paradigm. Specifically, SSL models aim to learn robust feature representations by solving an auxiliary task which can be defined entirely in the unlabeled setting. This process is also known as self-supervised pre-training, and the auxiliary tasks are commonly named pre-text tasks. Then, during the fine-tuning stage, the obtained representations are used to train a shallow classification model for the original (downstream) task using limited amounts of annotated data.

SSL methods have been successfully applied to HAR using individual modalities, but the multimodal setting is still insufficiently explored. Most recent self-supervised models in visual or sensor domains rely on a contrastive learning objective that aims to project the raw inputs into a feature space, such that similar, or positive, sample pairs have close representations, while semantically different, or negative, pairs are spaced apart. In multimodal settings, the Contrastive Multiview Coding (CMC) framework [8] forms positive pairs between the different modalities of each data sample, and negative pairs between different modalities of different samples. Since the data is unannotated, this implies that datapoints from negative pairs might correspond to the same class label

in the downstream task [9]. The presence of these false negatives is one of the major drawbacks of self-supervised contrastive learning approaches which rely on negative pairs. We visualize this limitation for inertial and skeleton modalities in Figure 1. Furthermore, as evidenced in [10], CMC uses only inter-modality negatives, although employing intra-modality negatives as well might have a positive impact on the intra-modal alignment of features.

In this paper, we aim not only to adapt contrastive learning to multimodal Human Activity Recognition using wearable sensor and skeletal data but also mitigate the limitations of the classical contrastive learning approaches by introducing a Contrastive Multiview Coding with Cross-Modal Knowledge Mining (CMC-CMKM) framework. The main contributions of this work are listed as follows:

- We implement the contrastive multiview coding (CMC) algorithm [8] to extract robust feature representations from inertial and skeleton data in the SSL settings. Moreover, we compare the designed models with the supervised and unimodal SSL approaches, namely Sim-CLR, built for each modality independently.
- We address the problem of false negative samples by introducing cross-modal knowledge mining techniques. First, we propose using feature representations learnt by unimodal encoders to mine additional positive pairs for the CMC framework, assuming they might otherwise represent false negatives. Besides, we propose using intra-modality positives and negatives to enhance the intra-modal alignment of features.
- Extensive experiments have been carried out on two open-source datasets containing inertial and skeleton modalities, namely UTD-MHAD [11] and MMAct [12].

II. RELATED WORK

A. Unimodal Human Activity Recognition

The most widely-used approaches for performing human activity recognition on inertial data often use CNNs [13] or RNNs [14] or a combination of these types of networks [15]. Recent works have also explored more advanced architectures based on attention mechanisms [16], [17], and deep metric learning, which attempt to learn robust feature embeddings relying on various contrastive loss functions, such as triplet loss [18], in a supervised manner.

Another widely-used input modality for HAR is skeleton data. A powerful technique for processing skeleton data is co-occurrence feature learning, which transposes the data in different ways as it is passed through a CNN, to capture both temporal and spatial relations between joints [19]. Recurrent neural networks have also been proposed as an alternative architecture for classifying skeleton sequences [20]. More recently, a great deal of attention has been given to graph convolutional networks, which explore the intuition of representing spatial and temporal structure of skeleton data as graphs [21]–[23].

B. Multimodal Human Activity Recognition

While promising results have been obtained by leveraging individual modalities for HAR, each modality has its own limitations. Then, combining multiple modalities should lead to more robust predictions in practical use cases. One of the main challenges in performing multimodal HAR is combining (or fusing) all of this different information in a coherent way, in order to provide a single prediction.

To account for the significant difference between input modalities, many multimodal HAR works apply input, feature or decision fusion in various architectures comprising of multiple backbone networks suitable for individual modalities [24]–[26]. More sophisticated recent works propose end-to-end architectures designed specifically for multimodal HAR. For example, Wang et al. propose a multi-view generative framework where GANs are used to generate the feature encodings of one view, given another [27]. This allows their framework to be used for inference even when one of the original modalities is unavailable. In [28], multiple multi-head attention mechanisms are used to encode and fuse features from different modalities. Liu et al. [29] proposed a framework which preserves the semantics of the original data by distilling knowledge from a teacher network which is trained on inertial data, to a student network which only uses RGB data.

C. Contrastive learning for Human Activity Recognition

Recent contrastive SSL methods rely on maximising the latent similarity of augmented views originating from the same data sample [30], [31]. The main issue associated with contrastive learning frameworks is their reliance on negative pairs. Besides the fact that contrastive pre-text tasks which use negative pairs normally require large batch sizes, there is also the issue of false negative pairs which may harm learning. While some works proposed using positive pairs only by introducing additional constraints to avoid trivial solutions [32], [33], others suggested different approaches to mitigate the impact of false negatives [9], [10], [34].

In the field of HAR, contrastive SSL has mainly been applied on individual data sources, such as sensors [35], [36], skeleton data [34] or visual data [37]. Despite the large number of supervised techniques which have achieved good performance on multimodal HAR, very few works, to the best of our knowledge, have addressed this problem using self-supervised learning. Akbari et al. [38] introduced the VATT framework that uses modality-specific and modality-agnostic Transformer encoders for multimodal self-supervised learning using video, audio and text modalities.

In this paper, we adapt the Contrastive Multiview Coding framework [8], previously used in Computer Vision applications, to the problem of multimodal HAR using inertial and skeleton data. Moreover, inspired by ideas suggested in [34] and [10], we introduce a novel cross-modal knowledge mining technique that can be easily plugged into the the CMC framework (CMC-CMKM). It aims to mitigate the impact of false negative pairs by using knowledge from each modality to guide the training process.

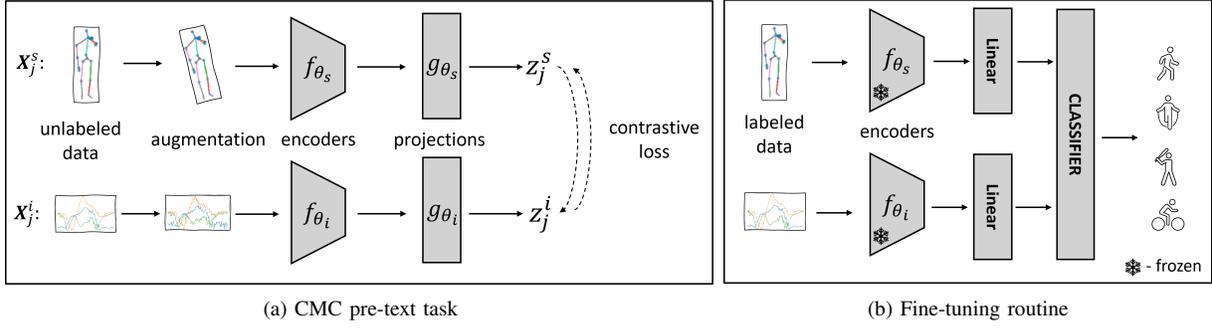


Fig. 2: Multimodal Contrastive Learning stages: pre-training (a) - inertial and skeleton data examples are passed through the modality-specific encoders and projection heads to generate representations used in contrastive loss; fine-tuning (b) - the pre-trained frozen encoders produce features for labeled data, these features are then passed through a mapping linear layer and classifier to get activity labels.

III. METHODOLOGY

A. Problem Definition

Multimodal Human Activity Recognition can be formulated as a classification problem where, given a set of inputs $\{\mathbf{X}^m | m \in M\}$ from a set of modalities M , the objective is to predict the label $y \in Y$ associated with these inputs. The remainder of this paper will focus on two input modalities, namely inertial (or sensor) data and skeleton data.

Inertial signals are generally obtained from wearable devices such as accelerometers, magnetometers or gyroscopes, and have the shape of multivariate time series. At any timestamp t , the input signal $\mathbf{x}_t = [x_t^1, x_t^2, \dots, x_t^S] \in \mathbb{R}^S$ consists of S values obtained from the S available sensor channels. In matrix form, an inertial data sample recorded over T timestamps is denoted as $\mathbf{X}^i = [\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_T] \in \mathbb{R}^{T \times S}$.

Skeleton data is generally provided as a set of 2D or 3D coordinates tracked over time for a number of keypoints located on the human body. For any skeleton sequence, we denote T as the number of frames in the sequence, J as the number of joints and C as the number of data channels (2 or 3). Then, a skeleton sequence $\mathbf{X}^s \in \mathbb{R}^{T \times J \times C}$ consists of T frames where the skeleton data for each frame is described by $\mathbf{P}_t = [p_t^1, p_t^2, \dots, p_t^J]$ and $p_t^j \in \mathbb{R}^C$ is the position of joint j at frame t .

B. Contrastive Learning for Multimodal HAR

The first contribution of this paper is the adaptation of the widely-used Contrastive Multiview Coding framework to the problem of multimodal HAR. It is a contrastive self-supervised learning method which can be used when two or more representations per example are available for each sample [8]. Specifically, the proposed adaptation of CMC contrasts between the feature embeddings obtained from the inertial and skeleton modalities.

Formally, for each sample $\{\mathbf{X}_j^i, \mathbf{X}_j^s\}$ in a training batch of size N (where i and s refer to the inertial and skeleton modalities, respectively), the input data for each modality is augmented with a random modality-specific augmentation, generating $\tilde{\mathbf{X}}_j^i = t_j^i(\mathbf{X}_j^i)$ and $\tilde{\mathbf{X}}_j^s = t_j^s(\mathbf{X}_j^s)$. For CMC, the purpose of data augmentation is simply to improve learning by

extending the size of the dataset. Then, two modality-specific encoders $f_{\theta_i}, f_{\theta_s}$ and projection heads $g_{\theta_i}, g_{\theta_s}$ are used to generate projections $z_j^i = g_{\theta_i}(f_{\theta_i}(\tilde{\mathbf{X}}_j^i))$ and $z_j^s = g_{\theta_s}(f_{\theta_s}(\tilde{\mathbf{X}}_j^s))$. These representations are then treated as a positive pair. This process is illustrated in Figure 2a. The negative pairs are formed by all inter-modal combinations of projections which do not originate from the same input instance. Thus, the loss obtained by treating \mathbf{X}_j^i as an anchor and enumerating over the representations of the other samples \mathbf{X}_k^s is:

$$l_j^{i \rightarrow s} = -\log \frac{\delta(z_j^i, z_j^s)}{\sum_{k=1}^N \delta(z_j^i, z_k^s)}, \quad (1)$$

where $\delta(z_j^i, z_j^s) = \exp(s(z_j^i, z_j^s))/\tau$ and $s(\cdot)$ is the cosine similarity function.

The total loss accumulated over the training batch is calculated as follows:

$$\mathcal{L} = \sum_{j=1}^N (l_j^{i \rightarrow s} + l_j^{s \rightarrow i}) \quad (2)$$

The inertial and skeleton encoders pre-trained within the CMC framework are then frozen and used in the fine-tuning stage as shown in Figure 2b. Specifically, we map inertial and skeleton features to the same size using a single fusion linear layer, including batch normalization and ReLU, concatenate the outputs and pass the resulting feature vector through the classification model.

To provide a comparison with CMC, we have also implemented the SimCLR [31] framework for both modalities independently. Specifically, the pre-text task is performed separately for the inertial and skeleton encoders. In this framework, two random sets of augmentations are applied to each input instance to create positive pairs.

C. Cross-Modal Knowledge Mining

Contrastive multiview coding is a powerful framework for performing multimodal SSL. However, the training procedure and the formulation of the loss function still rely on a set of underlying assumptions which might have a negative impact on the learned representations. First, CMC relies heavily on

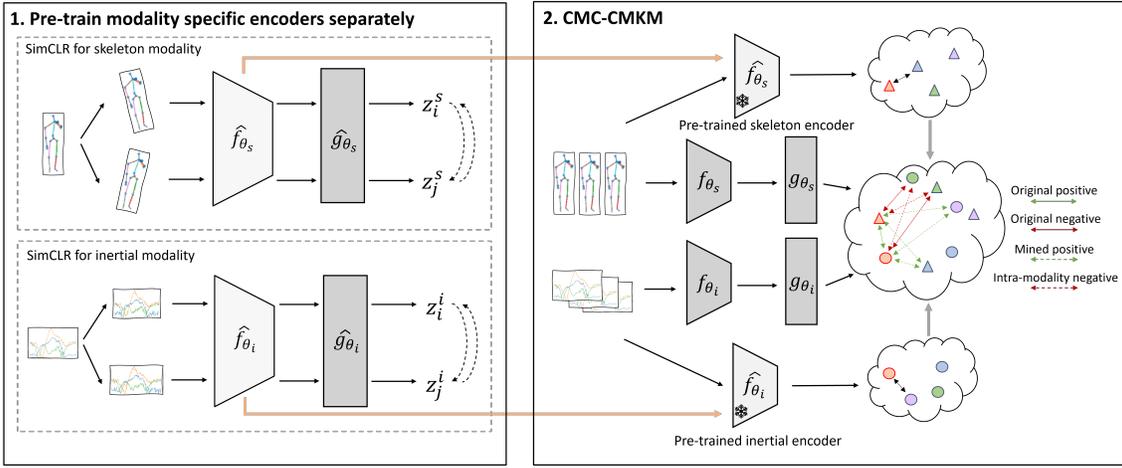


Fig. 3: CMC with Cross-modal knowledge mining (CMKM). First, additional encoders are pre-trained separately using SimCLR. Second, the additional encoders are used to guide additional positive mining within the CMC framework. Besides, CMC-CMKM uses intra-modality negatives. Triangles and circles indicate skeleton and inertial features, respectively, while each color corresponds to one instance. For the orange instance, a number of positive and negative relationships are shown.

negative pairs in its contrastive objective, which makes it more sensitive to batch size. Furthermore, as ground-truth labels are unavailable during the pre-training process, there is no way to prevent the negative pairs from occasionally consisting of false negatives. Finally, as CMC only contrasts between representations from different modalities, the encoders are not explicitly trained to preserve intra-modal similarities.

First, inspired by [34], we upgrade the CMC framework with a novel method for cross-modal and intra-modal positive mining. We take advantage of the cross-modal setting and use knowledge from one modality to guide the training process for the other. Intuitively, if two samples are very similar in one modality, there is a chance that the samples might come from the same underlying action class. Applying this intuition to our framework, we use similarities between representations learnt by encoders pre-trained separately using SimCLR to mine additional positives. Specifically, for each modality, we use a pre-trained encoder to compute a similarity matrix containing the pair-wise similarity values for each pair of sample encodings in a batch. For each instance, we select the top- K most similar samples, for a fixed K value, and we include these mined samples in the positive sets of the instance, in both modalities.

More formally, given a training batch of size N , and intra-modality similarity matrices \mathbf{S}^i and \mathbf{S}^s for the inertial and skeleton modalities, respectively, we define the positive and negative sets as follows:

For the inertial modality:

$$P_j^i = \{z_j^i\} \cup \{z_l^i \in \text{Top}K(\mathbf{S}_j^i)\} \cup \{z_l^i \in \text{Top}K(\mathbf{S}_j^s)\}$$

$$N_j^i = \{z_k^i \mid 0 \leq k \leq N\} \setminus P_j^i$$

For the skeleton modality:

$$P_j^s = \{z_j^s\} \cup \{z_l^s \in \text{Top}K(\mathbf{S}_j^s)\} \cup \{z_l^s \in \text{Top}K(\mathbf{S}_j^i)\}$$

$$N_j^s = \{z_k^s \mid 0 \leq k \leq N\} \setminus P_j^s$$

We note that both positive sets contain embeddings originating from both modalities. The respective loss term for each sample in the inertial modality becomes:

$$l_j^{i \rightarrow s} = -\log \frac{\sum_{z_k^m \in P_j^i} \delta(z_j^i, z_k^m)}{\sum_{z_k^m \in P_j^i \cup N_j^i} \delta(z_j^i, z_k^m)} \quad (3)$$

The loss term for the skeleton modality is defined according to the same logic.

Additionally, we exploit intra-modality negatives to encourage the model to better align features inside each modality. This is done by adding the similarities of intra-modality negatives to the denominator of the loss term. Finally, the proposed loss function for each sample can be formally written as follows:

$$l_j^{i \rightarrow s} = -\log \frac{\sum_{z_k^m \in P_j^i} \delta(z_j^i, z_k^m)}{\sum_{z_k^m \in P_j^i \cup N_j^i} \delta(z_j^i, z_k^m) + \sum_{z_k^i \in N_j^s} \delta(z_j^i, z_k^i)} \quad (4)$$

We also note that the addition of intra-modality negatives is consistent with positive mining, as the mined samples are also implicitly excluded from the intra-modality negative sets. An illustrative example of the proposed improvements is shown in Figure 3. The whole pre-training routine is summarized in Algorithm 1.

D. Backbone Models

The proposed framework requires two encoders to extract features from inertial and skeleton signals. For inertial data, the encoder f_{θ_i} is the transformer-like encoder described in the CSSHAR framework [36]. The input data is passed through a one-dimensional CNN with batch normalization and a ReLU non-linearity, then through a positional encoding layer and a transformer encoder consisting of multiple self-attention blocks, as described in the original Transformer architecture [39].

Algorithm 1: Model pre-training using cross-modal knowledge mining

Data: unlabelled dataset $\{\mathbf{X}_k^i, \mathbf{X}_k^s\}_{k=1}^N$, where N is the number of training samples

Input: inertial encoders $\hat{f}_{\theta_i}, f_{\theta_i}$ and projections heads $\hat{g}_{\theta_i}, g_{\theta_i}$, skeleton encoders $\hat{f}_{\theta_s}, f_{\theta_s}$ and projections heads $\hat{g}_{\theta_s}, g_{\theta_s}$

stage 1: unimodal pre-training

pre-train encoder \hat{f}_{θ_i} and projection head \hat{g}_{θ_i} using SimCLR, then discard \hat{g}_{θ_i} and freeze \hat{f}_{θ_i} ;

pre-train encoder \hat{f}_{θ_s} and projection head \hat{g}_{θ_s} using SimCLR, then discard \hat{g}_{θ_s} and freeze \hat{f}_{θ_s} ;

stage 2: main multimodal pre-training

for each training batch $\{\mathbf{X}_k^i, \mathbf{X}_k^s\}_{k=1}^n$ **do**

 obtain augmented samples $\{\tilde{\mathbf{X}}_k^i, \tilde{\mathbf{X}}_k^s\}_{k=1}^n$;

 compute projections $\{z_k^i\}_{k=1}^n, \{z_k^s\}_{k=1}^n$;

 # positive mining

 for $k \in \{1, \dots, n\}, l \in \{1, \dots, n\}$, compute

$\mathbf{S}_{k,l}^i = s(\hat{f}_{\theta_i}(\tilde{\mathbf{X}}_k^i), \hat{f}_{\theta_i}(\tilde{\mathbf{X}}_l^i))$,

$\mathbf{S}_{k,l}^s = s(\hat{f}_{\theta_s}(\tilde{\mathbf{X}}_k^s), \hat{f}_{\theta_s}(\tilde{\mathbf{X}}_l^s))$;

 define sets $P_j^i, N_j^i, P_j^s, N_j^s$;

 # contrastive loss

for $k \in \{1, \dots, n\}$ **do**

 | compute $l_k^{i \rightarrow s}, l_k^{s \rightarrow i}$ according to Equation 4;

end

 compute total loss $\mathcal{L} = \sum_{k=1}^N (l_k^{i \rightarrow s} + l_k^{s \rightarrow i})$;

 update $f_{\theta_i}, g_{\theta_i}, f_{\theta_s}, g_{\theta_s}$ to minimize \mathcal{L} ;

end

For the skeleton modality, we picked the lightweight convolutional co-occurrence feature learning network [19]. It uses a two-stream input (of positions and motions) and comprises of a series of convolutional blocks, with ReLU non-linearities and max-pooling applied to certain layers. A key element of this architecture is a transpose block which is inserted into the network between two intermediate layers, and which rearranges the data such that the joints become the input channels of subsequent convolutions. This allows the network to learn features in a hierarchical manner, from point-level features describing each joint, to co-occurrence features which capture the relationship between the different joints in a sequence.

IV. IMPLEMENTATION DETAILS

A. Datasets

In this paper, two open-source multimodal datasets were used to evaluate the performance of the proposed approaches, namely UTD-MHAD [11] and MMAAct [12]. Skeleton and inertial modalities were extracted and used from both datasets. **UTD-MHAD.** The dataset contains data collected by 10 subjects performing 27 activities, 4 trials for each. The three-dimensional joint coordinates were recorded with a Kinect

camera, while the inertial data was collected using one wearable device with accelerometer and gyroscope. We follow the original evaluation protocol, using odd-numbered subjects for training and even-numbered subjects for testing and reporting test accuracy.

MMAAct. The dataset consists of 36 activities performed by 20 subjects in different scenes. For skeleton data, we employ the 2D keypoints present in the challenge version of the dataset¹. The sensor data comes from a smartwatch (accelerometer) and a smartphone (accelerometer, gyroscope, orientation) located in the subject’s pocket. We follow the cross-subject and cross-scene evaluation protocols. Specifically, for the former we use the samples from the first 16 subjects for training and the others for testing, while for the latter we reserve all samples collected in the occlusion scene for testing, and train on all subjects and all other scenes. As per the authors’ recommendation, we report the F1 score obtained on the test set.

B. Hyperparameters

In this subsection, we describe the hyperparameters used to pre-train and fine-tune the proposed models as well as specific details regarding the architectures of the modality-specific encoders. To optimize the parameters of the models, we use the Adam optimizer with a learning rate of 0.001 which is reduced twice when learning stagnates for more than 20 epochs.

Inertial encoder. The inertial encoder implemented in this paper is adapted from CSSHAR [36]. Specifically, first, input data is passed through 3 one-dimensional CNN blocks consisting of [32, 64, 128] feature maps with kernel size 5. The obtained feature maps are then used as an input for 2 self-attention blocks with 2 heads each.

Skeleton encoder. The skeleton encoder adopted for the experiments follows a hierarchical co-occurrence learning architecture [19]. We implement the model as described in the original paper, only replacing dropout layers with batch normalization layers.

Initial pre-processing. We re-sample all input sequences to 50 timesteps, for both inertial and skeleton data. Additionally, we normalise joint positions in all skeleton sequences based on the first frame of each sequence, following a standard normalisation procedure [24].

Pre-training. Prior to pre-training the models, we apply a set of random augmentations for inertial and skeleton modalities. The inertial augmentations, as proposed in [36], applied to each input instance are sampled randomly (with 75% probability) from the set of augmentations {jittering, scaling, rotation} for UTD and {jittering, scaling, permutation, channel shuffle} for MMAAct. For the skeleton modality, we use {jittering, random resized crops, scaling, rotation, shearing} on both datasets. Jittering is always applied, while the other augmentations are applied with a 75% probability. We pre-train with SimCLR for 300 epochs, and with CMC-CMKM

¹challenge dataset: <https://mmap19.github.io/challenge/>

top- K	UTD-MHAD	MMAct (F1-score)	
	(Accuracy)	x-subject	x-scene
0	94.88	83.36	79.06
1	97.67	84.51	82.91
2	96.05	81.92	81.73
3	96.05	82.41	82.77
4	94.65	82.64	81.4
5	94.88	82.96	82.84

TABLE I: Ablation for different k in cross-modal positive mining. The row with $k = 0$ refers to the model using intra-modality negatives only.

for 100 epochs. For the SimCLR experiments on inertial data, we use batch sizes of 128 and 64 and temperature values of 0.05 and 0.2 for UTD-MHAD and MMAct, respectively. For skeleton SimCLR, we use batches of 32 and 128 samples, and temperatures of 0.5 and 0.2. For the multimodal experiments using CMC-CMKM, we used a batch size of 64 for UTD-MHAD and 128 for MMAct, and temperature values of 0.1 for both datasets.

Fine-tuning. For the fine-tuning routine, which remains the same for all multimodal approaches (Figure 2b), we implement modality-specific fusion layers (1 layer per modality), including batch normalization and ReLU, that map inertial and skeleton embeddings to the same size of 256. The obtained features are then concatenated and passed through a simple linear classifier with softmax activation. We train the modality specific fusion layers and linear classifier for 100 epochs using the labels of the downstream task.

V. EVALUATIONS

Our code has been made publicly available on GitHub². All experiments have been run on a single Nvidia Quadro RTX 5000 card. One epoch of CMC pre-training on MMAct takes approximately 8-9 seconds, while one epoch of CMC-CMKM pre-training takes 14-15 seconds. It is also worth mentioning that CMC-CMKM pre-training requires models pre-trained in the unimodal settings, with unimodal pre-training taking approximately 15-17 seconds per epoch. Finally, fine-tuning for both CMC and CMC-CMKM takes approximately 4-5 seconds per epoch.

A. Learning Feature Representations

In order to evaluate the representations learnt by the proposed multimodal approaches, we perform linear evaluation on top of the fused features extracted by the pre-trained modality-specific encoders. Specifically, the whole annotated datasets are used to fine-tune the fusion layer and linear classifier as shown in Figure 2b.

Number of mined positives. First, we explore how the number of mined positives k affects the performance of models in the proposed cross-modal knowledge mining protocol. The results of this experiment are presented in Table I. In Table I, when K is set to 0, only the intra-modality negatives are used. As can be seen from the table, the proposed method

²<https://github.com/razzu/cmc-cmkm>

Modality	Approach	UTD-MHAD	MMAct (F1-score)	
		(Accuracy)	x-subject	x-scene
Inertial	SimCLR	72.09	52.89	59.23
Inertial	Supervised	76.74	61.22	78.86
Skeleton	SimCLR	95.11	75.82	67.80
Skeleton	Supervised	94.65	82.50	70.58
Multimodal	CMC	96.04	82.05	79.97
Multimodal	CMC-CMKM	97.67	84.51	<u>82.91</u>
Multimodal	Supervised	<u>97.21</u>	<u>84.05</u>	87.36

TABLE II: Linear evaluation results: the highest results are highlighted in bold, the second highest are underlined.

Pre-text	x-subject		x-scene	
	inertial	skeleton	inertial	skeleton
Unimodal SimCLR	51.77	66.37	52.98	66.81
CMC	55.26	73.98	57.33	74.26
CMC-CMKM	56.66	75.77	57.44	75.32

TABLE III: Activity retrieval accuracy scores on MMAct.

shows optimal performance in all three scenarios, when K is equal to 1, meaning that one extra positive is mined from each modality.

Comparison to baselines. In Table II, we also compare the performance of the proposed CMC-CMKM approach to other SSL models, pre-trained in the unimodal setting with SimCLR and in the multimodal setting with standard CMC. Additionally, we include the performance of identical encoders trained in a supervised end-to-end manner.

According to the obtained performance scores, it is clear that multimodal methods, both SSL and supervised, are much more powerful than the unimodal ones. Furthermore, SSL approaches on multimodal data outperform all the unimodal models trained in the supervised settings. Besides this, the introduced CMC-CMKM approach outperforms CMC by about 2% in all three scenarios. Moreover, it shows performance comparable to the multimodal supervised model on UTD-MHAD and MMAct (cross-subject) and outperforms it by 0.5%. For the cross-scene scenario on MMAct, the CMC-CMKM is the closest one to the supervised model.

Activity retrieval. We employ an activity retrieval scenario for modality-specific encoders pre-trained in unimodal and multimodal manner on MMAct. Namely, given an input skeleton or inertial signal stream from the test set, we aim to find the most similar example of the same modality in the training set using learnt feature representations. In this scenario, instead of the default fine-tuning routine, we use kNN ($k = 1$) to predict activities in the test set using inertial and skeleton encoders pre-trained with the contrastive SSL approaches. In other words, we match each feature embedding from the test set with the closest one from the training set using cosine similarity. The accuracy scores for this scenario are shown in Table III. According to the obtained results, the encoders pre-trained in multimodal settings significantly outperform models pre-trained in the unimodal manner. Moreover, the proposed CMC-CMKM methods demonstrates better performance than the original CMC indicating improved intra-modality feature alignment for both modalities.

Qualitative analysis of features. In order to assess the separation between classes visually, we project the feature embeddings into a two-dimensional space using t-SNE [40]

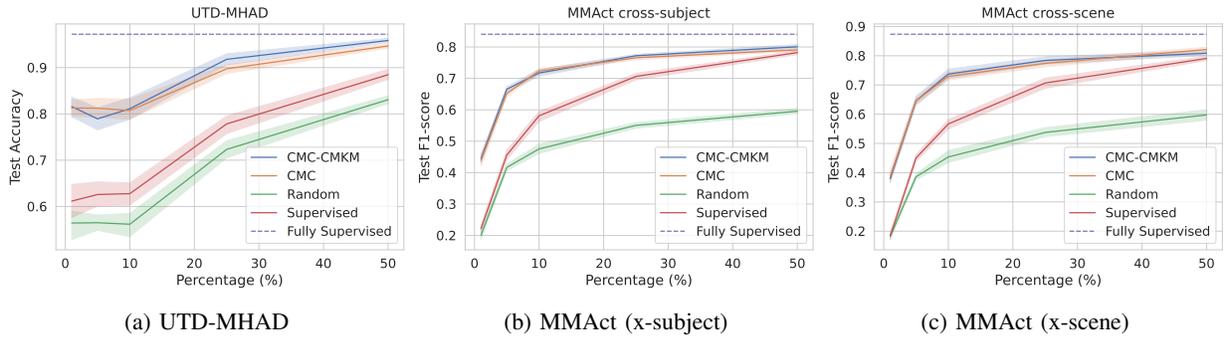


Fig. 4: Average values of performance metrics with 95% confidence intervals for the semi-supervised learning scenario.

and visualize them in Figure 5. For the multimodal approach, we concatenate inertial and skeleton features before feeding them to the t-SNE. From the obtained diagrams, it is clear that features learnt on multimodal data contribute to better class separation.

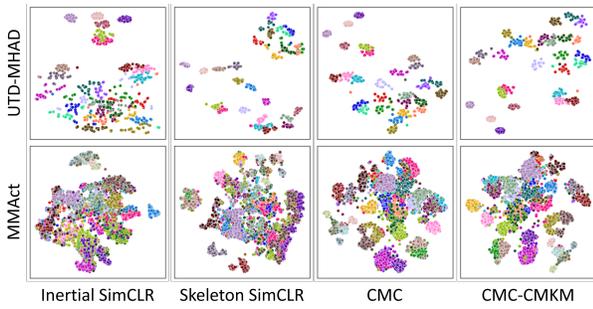


Fig. 5: Visualization of the learnt representations using t-SNE.

B. Semi-supervised Learning Scenario

In a more realistic scenario, vast amounts of labeled training data might not be available. In this case, one can still use the unannotated dataset to pre-train models in the SSL manner. For semi-supervised learning evaluation, we limit the amount of labels available for training. Specifically, we perform an experiment where only a random percentage $p \in \{1\%, 2\%, 5\%, 10\%, 25\%, 50\%\}$ of labels is present. In this experiment, we compare the performance of the proposed multimodal SSL models to the supervised and random models. Namely, we pre-train CMC and CMC-CMKM using the whole unannotated dataset and fine-tune the fusion linear layers and the linear classifier with annotated data. Besides, we train a supervised multimodal model with identical encoders using these data. Finally, we also fine-tune the fusion network and the linear classifier for randomly initialized encoders. For each value of p the experiment is repeated 10 times and the average performance values with the corresponding confidence intervals are presented in Figure 4. We also include, as a horizontal dashed line, the performance of a fully supervised multimodal network ($p=100\%$).

According to the obtained figures, the multimodal SSL approaches are much more robust, especially when very limited amounts of annotated data are available. For both datasets, the

SSL models outperform the identical supervised models by more than 10% when less than 10% of data is annotated. What is more, the proposed SSL models almost reach the performance of the fully supervised model when more than 25% of labeled data is available.

VI. CONCLUSIONS AND FUTURE WORK

In this paper, we adopted a set of modality-specific network architectures for encoding inertial and skeleton data and implemented modern unimodal and multimodal self-supervised frameworks, adapting them to the problem of HAR. Furthermore, we proposed a novel framework named CMC-CMKM, which addresses issues related to the CMC pre-training by injecting modality-specific knowledge into the learning process. Extensive experiments have shown that the proposed multimodal SSL frameworks outperform unimodal supervised approaches and show satisfactory performance comparing to multimodal fully-supervised models.

As for the future work, additional data modalities can be used in combination with different backbone architectures. Moreover, while the problems related to negative pairs have been mitigated to some extent using CMC-CMKM framework, there is an entire class of self-supervised approaches which does not rely on negative pairs and it might be useful to explore how they can be adapted to multimodal settings.

REFERENCES

- [1] H. D. Mehr and H. Polat, "Human activity recognition in smart home with deep learning approach," in *2019 7th International Istanbul Smart Grids and Cities Congress and Fair (ICSG)*, 2019, pp. 149–153.
- [2] Y. Du, Y. Lim, and Y. Tan, "A novel human activity recognition and prediction in smart home based on interaction," *Sensors*, vol. 19, no. 20, 2019. [Online]. Available: <https://www.mdpi.com/1424-8220/19/20/4474>
- [3] M. Panwar, D. Biswas, H. Bajaj, M. Jobges, R. Turk, K. Maharatna, and A. Acharyya, "Rehab-net: Deep learning framework for arm movement classification using wearable sensors for stroke rehabilitation," *IEEE Transactions on Biomedical Engineering*, vol. PP, pp. 1–1, 02 2019.
- [4] R. Grzeszick, J. M. Lenk, F. M. Rueda, G. A. Fink, S. Feldhorst, and M. ten Hompel, "Deep neural network based human activity recognition for the order picking process," in *Proceedings of the 4th International Workshop on Sensor-Based Activity Recognition and Interaction*, ser. iWOAR '17. New York, NY, USA: Association for Computing Machinery, 2017.
- [5] R. Vleugels, B. Van Herbruggen, J. Fontaine, and E. De Poorter, "Ultra-wideband indoor positioning and imu-based activity recognition for ice hockey analytics," *Sensors*, vol. 21, no. 14, 2021. [Online]. Available: <https://www.mdpi.com/1424-8220/21/14/4650>

- [6] Z. Sun, J. Liu, Q. Ke, and H. Rahmani, "Human action recognition from various data modalities: A review," *ArXiv*, vol. abs/2012.11866, 2020.
- [7] S. Yadav, K. Tiwari, H. Pandey, and S. Ali Akbar, "A review of multimodal human activity recognition with special emphasis on classification, applications, challenges and future directions," *Knowledge-Based Systems*, vol. 223, p. 106970, 04 2021.
- [8] Y. Tian, D. Krishnan, and P. Isola, "Contrastive multiview coding," in *Computer Vision – ECCV 2020*, A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm, Eds. Cham: Springer International Publishing, 2020, pp. 776–794.
- [9] C. Chuang, J. Robinson, Y. Lin, A. Torralba, and S. Jegelka, "Debiased contrastive learning," in *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., 2020.
- [10] M. Zolfaghari, Y. Zhu, P. Gehler, and T. Brox, "Crossclr: Cross-modal contrastive learning for multi-modal video representations," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2021, pp. 1450–1459.
- [11] C. Chen, R. Jafari, and N. Kehtarnavaz, "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)*, 2015, pp. 168–172.
- [12] Q. Kong, Z. Wu, Z. Deng, M. Klinkigt, B. Tong, and T. Murakami, "Mmact: A large-scale dataset for cross modal human action understanding," in *The IEEE International Conference on Computer Vision (ICCV)*, October 2019.
- [13] J. Yang, M. N. Nguyen, P. P. San, X. Li, and S. Krishnaswamy, "Deep convolutional neural networks on multichannel time series for human activity recognition," in *IJCAI*, 2015.
- [14] N. Y. Hammerla, S. Halloran, and T. Plötz, "Deep, convolutional, and recurrent models for human activity recognition using wearables," in *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, ser. IJCAI'16. AAAI Press, 2016, p. 1533–1540.
- [15] C. Xu, D. Chai, J. He, X. Zhang, and S. Duan, "Innohar: A deep neural network for complex human activity recognition," *IEEE Access*, vol. 7, pp. 9893–9902, 2019.
- [16] M. Zeng, H. Gao, T. Yu, O. J. Mengshoel, H. Langseth, I. Lane, and X. Liu, "Understanding and improving recurrent networks for human activity recognition by continuous attention," in *Proceedings of the 2018 ACM International Symposium on Wearable Computers*, ser. ISWC '18. New York, NY, USA: Association for Computing Machinery, 2018, p. 56–63. [Online]. Available: <https://doi.org/10.1145/3267242.3267286>
- [17] S. Mahmud, M. T. H. Tonmoy, K. K. Bhaumik, A. K. M. M. Rahman, M. A. Amin, M. Shoyaib, M. A. H. Khan, and A. A. Ali, "Human activity recognition from wearable sensor data using self-attention," in *ECAI 2020 - 24th European Conference on Artificial Intelligence, 29 August-8 September 2020, Santiago de Compostela, Spain*. IOS Press, 2020, pp. 1332–1339. [Online]. Available: <https://doi.org/10.3233/FAIA200236>
- [18] B. Khaertdinov, E. Ghaleb, and S. Asteriadis, "Deep triplet networks with attention for sensor-based human activity recognition," in *2021 IEEE International Conference on Pervasive Computing and Communications (PerCom)*, 2021, pp. 1–10.
- [19] C. Li, Q. Zhong, D. Xie, and S. Pu, "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*, ser. IJCAI'18. AAAI Press, 2018, p. 786–792.
- [20] S. Song, C. Lan, J. Xing, W. Zeng, and J. Liu, "An end-to-end spatio-temporal attention model for human action recognition from skeleton data," in *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, ser. AAAI'17. AAAI Press, 2017, p. 4263–4270.
- [21] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in *Thirty-second AAAI conference on artificial intelligence*, 2018.
- [22] K. Cheng, Y. Zhang, X. He, W. Chen, J. Cheng, and H. Lu, "Skeleton-based action recognition with shift graph convolutional network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.
- [23] Y.-F. Song, Z. Zhang, C. Shan, and L. Wang, "Efficientgcn: Constructing stronger and faster baselines for skeleton-based action recognition," *arXiv:2106.15125*, 2021.
- [24] P. Khair, P. Kumar, and J. Imran, "Combining cnn streams of rgb-d and skeletal data for human activity recognition," *Pattern Recognition Letters*, vol. 115, pp. 107–116, 2018, multimodal Fusion for Pattern Recognition. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0167865518301636>
- [25] R. Memmesheimer, N. Theisen, and D. Paulus, "Gimme signals: Discriminative signal encoding for multimodal activity recognition," 10 2020, pp. 10394–10401.
- [26] A. Das, P. Sil, P. K. Singh, V. Bhateja, and R. Sarkar, "Mmharsemnet: A multi-modal human activity recognition model," *IEEE Sensors Journal*, vol. 21, no. 10, pp. 11569–11576, 2021.
- [27] L. Wang, Z. Ding, Z. Tao, Y. Liu, and Y. Fu, "Generative multi-view human action recognition," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 6211–6220.
- [28] M. M. Islam and T. Iqbal, "Hamlet: A hierarchical multimodal attention-based human activity recognition algorithm," *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 10285–10292, 2020.
- [29] Y. Liu, K. Wang, G. Li, and L. Lin, "Semantics-aware adaptive knowledge distillation for sensor-to-vision action recognition," *IEEE Transactions on Image Processing*, vol. PP, 06 2021.
- [30] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 9729–9738.
- [31] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *Proceedings of the 37th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, H. D. III and A. Singh, Eds., vol. 119. PMLR, 13–18 Jul 2020, pp. 1597–1607. [Online]. Available: <http://proceedings.mlr.press/v119/chen20j.html>
- [32] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. H. Richemond, E. Buchatskaya, C. Doersch, B. A. Pires, Z. D. Guo, M. G. Azar, B. Piot, K. Kavukcuoglu, R. Munos, and M. Valko, "Bootstrap your own latent: A new approach to self-supervised learning," 2020.
- [33] X. Chen and K. He, "Exploring simple siamese representation learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 15750–15758.
- [34] L. Li, M. Wang, B. Ni, H. Wang, J. Yang, and W. Zhang, "3d human action representation learning via cross-view consistency pursuit," in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*. Computer Vision Foundation / IEEE, 2021, pp. 4741–4750.
- [35] H. Haresamudram, I. Essa, and T. Plöetz, "Contrastive predictive coding for human activity recognition," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 5, pp. 1–26, 2021.
- [36] B. Khaertdinov, E. Ghaleb, and S. Asteriadis, "Contrastive self-supervised learning for sensor-based human activity recognition," in *2021 IEEE International Joint Conference on Biometrics (IJCB)*, 2021, pp. 1–8.
- [37] T. Han, W. Xie, and A. Zisserman, "Self-supervised co-training for video representation learning," in *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 5679–5690.
- [38] H. Akbari, L. Yuan, R. Qian, W.-H. Chuang, S.-F. Chang, Y. Cui, and B. Gong, "VATT: Transformers for multimodal self-supervised learning from raw video, audio and text," in *Thirty-Fifth Conference on Neural Information Processing Systems*, 2021. [Online]. Available: <https://openreview.net/forum?id=RzYrn625bu8>
- [39] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.
- [40] L. van der Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of Machine Learning Research*, vol. 9, pp. 2579–2605, 11 2008.