IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 30 September 2024, accepted 10 October 2024, date of publication 14 October 2024, date of current version 25 October 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3480814

TOPICAL REVIEW

Contrastive Self-Supervised Learning for Sensor-Based Human Activity Recognition: A Review

HUI CHEN[1], CHARLES GOUIN-VALLERAND[1], KÉVIN BOUCHARD[2], (Member, IEEE), SÉBASTIEN GABOURY[2], (Senior Member, IEEE), MÉLANIE COUTURE[3], NATHALIE BIER[4,5], AND SYLVAIN GIROUX[1]

[1]Computer Science Department, Université de Sherbrooke, Sherbrooke, QC J1K 2R1, Canada
[2]Computer Science and Mathematics Department, Université du Québec à Chicoutimi, Saguenay, QC G7H 2B1, Canada
[3]School of Social Work, Université de Sherbrooke, Sherbrooke, QC J1K 2R1, Canada
[4]School of Rehabilitation, Université de Montréal, Montreal, QC H3N 1X7, Canada
[5]Centre de Recherche de l'Institut Universitaire de Gériatrie de Montréal (CRIUGM), Montreal, QC H3W 1W5, Canada

Corresponding author: Hui Chen (hui.chen@usherbrooke.ca)

This research was supported in part by AGE-WELL NCE, in part by MEDTEQ, in part by the Natural Sciences and Engineering Research Council of Canada (NSERC), in part by the Réseau Québécois de Recherche sur le Vieillissement (RQRV), in part by the Université de Sherbrooke, and in part by the Fondation Luc Maurice and Videotron Inc. The work of Nathalie Bier was supported by the Research Scholar Award from the Fonds de la recherche du Québec–Santé.

ABSTRACT Deep learning models have achieved significant success in human activity recognition, particularly in assisted living and telemonitoring. However, training these models requires substantial amounts of labeled training data, which is time-consuming and costly to acquire in real-world environments. Contrastive self-supervised learning has recently garnered attention in sensor-based activity recognition to mitigate the need for expensive large-scale data collection and annotation. Despite numerous related published papers, there remains a lack of literature reviews highlighting recent advances in contrastive self-supervised learning for sensor-based activity recognition. This paper extensively reviews 43 papers on recent contrastive self-supervised learning methods for sensor-based human activity recognition, excluding those related to video or audio sensors due to privacy concerns. First, we summarize the taxonomy of contrastive self-supervised learning, followed by a detailed description of contrastive learning models used for activity recognition and their main components. Next, we comprehensively review data augmentation methods for sensor data and commonly used benchmark datasets for activity recognition. The empirical performance comparisons of different methods are presented on benchmark datasets in linear evaluation, semi-supervised learning, and transfer learning scenarios. Through these comparisons, we derive significant insights into the selection of contrastive self-supervised models for sensor-based activity recognition. Finally, we discuss the limitations of current research and outline promising research directions for future exploration.

INDEX TERMS Self-supervised learning, contrastive learning, semi-supervised learning, transfer learning, human activity recognition, sensors.

I. INTRODUCTION

Human activity recognition (HAR) plays a vital role in various applications across diverse fields, including assisted living [1], remote monitoring, fitness tracking [2], and smart homes [3]. Sensor-based HAR involves identifying human activities, such as walking and sitting, using wearable and embedded environmental sensors [4]. In recent years, sensor-based HAR has been extensively investigated and achieved tremendous success using expressive deep neural architectures [5]. The impressive performance has potentially driven the growing deployment of HAR systems in real-world settings [4]. However, deep HAR models still face substantial challenges, particularly in acquiring the

The associate editor coordinating the review of this manuscript and approving it for publication was Renato Ferrero.

large-scale, labeled datasets required for training. Collecting comprehensive and annotated data in real-world settings is expensive and labor-intensive, posing a significant barrier to the widespread adoption and implementation of deep learning in HAR [6]. In some cases, annotation can raise privacy issues, especially in medical and other human-related applications [7]. Consequently, it is crucial to develop methods for pre-training sensor-based HAR models with unlabeled data, allowing them to be fine-tuned for various downstream tasks without relying on large amounts of labeled data.

Self-supervised learning (SSL) is an emerging paradigm that learns meaningful representations from massive unlabeled data, reducing the dependency on manual annotations [8]. Due to its ability to address the scarcity of labels, SSL was first popularized in natural language processing (NLP) and has since spread to diverse fields such as computer vision (CV) and audio and speech processing [9], [10]. In recent years, contrastive SSL, a subset of SSL, has become a powerful alternative to traditional supervised learning. It extracts transferable knowledge from abundant unlabeled data by contrasting positive and negative pairs of data samples [11]. These learned representations can then be fine-tuned for various downstream tasks without needing large-scale labeled datasets. Motivated by the immense success of contrastive SSL models in CV domains such as Momentum Contrast (MoCo) [12], SimCLR [11], and Bootstrap Your Own Latent (BYOL) [13], a growing number of researchers have explored the applications in sensor-based HAR, demonstrating performance comparable to that of supervised models [14], [15], especially in label-scarce scenarios. By training encoders in an SSL manner, contrastive SSL facilitates the learning of latent representations that are highly beneficial for various downstream tasks, including classification and prediction in the context of HAR [16]. Furthermore, contrastive SSL offers versatility through its adaptability to semi-supervised learning environments, where only a minimal amount of labeled data is available [17]. This adaptability significantly reduces the reliance on extensive labeled datasets. Additionally, contrastive SSL supports transfer learning, allowing the transfer of learned representations to different datasets, thereby enhancing the scalability and applicability of HAR models across diverse settings and scenarios [18], [19].

Although there has been considerable research on SSL for HAR, there are no comprehensive review papers addressing the latest contrastive SSL methods for sensor-based HAR, and the most suitable strategies for HAR have not been sufficiently investigated. This paper aims to provide an overview of the current achievements and evaluate the effectiveness of various contrastive SSL methods in leveraging sensor data to address these gaps. It identifies challenges and solutions in the current landscape and highlights potential directions for future research, serving as a reference for researchers in the sensor-based HAR field. Due to privacy concerns associated with cameras [20] and microphones, this paper mainly focuses on ambient and wearable sensors. In summary, the main contributions are as follows:

- A comprehensive survey of contrastive SSL for sensor-based HAR is provided, including the taxonomy of contrastive SSL, commonly used models for sensor-based HAR, and important frameworks.
- We offer an in-depth review of recent augmentation functions for sensor data and sensor-based HAR datasets.
- The performance of existing contrastive SSL methods for HAR is quantitatively compared in linear evaluation, semi-supervised learning, and transfer learning. We derive significant insights for selecting effective contrastive SSL methods for sensor-based HAR through these comparisons.
- We point out several open issues, analyze the limitations and discuss possible future directions for contrastive SSL in the sensor-based HAR domain.

The rest of the paper is organized as follows: Section II introduces the search strategy adopted to select the papers. In Section III, we present a taxonomy of contrastive SSL and their models in HAR. The important frameworks for contrastive SSL are illustrated in Section IV. Sections V and VI list commonly used datasets for HAR and compare the recent contrastive SSL models based on the most popular HAR datasets. Section VII discusses the limitations in current research and identifies potential future directions. Finally, Section VIII provides the conclusion.

II. METHOD

The search strategy follows the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines [21] to ensure a comprehensive overview of the satisfied requirements. The search targeted papers published between January 2019 and July 2024 in IEEE Xplore, Scopus, and ACM Digital Library databases as primary sources, including conference and journal papers. To specifically focus on contrastive SSL, the search excluded the broader term *SSL*, which encompasses other approaches such as generative models, as well as the term *unsupervised learning*, which would significantly broaden the scope of identification in the databases. Instead, the selected search keywords include *self-supervised*, *activity recognition*, *sensor*, and *contrastive learning*. The full search queries for the three databases are:

- IEEE Xplore: (“All Metadata”:“self-supervised” OR “All Metadata”:“contrastive learning”) AND (“All Metadata”:activity recognition) AND (“All Metadata”:sensor). Publication year is between 2019 to 2024.
- Scopus: (TITLE-ABS-KEY (“self-supervised” OR “contrastive learning”) AND TITLE-ABS-KEY (activity AND recognition) AND TITLE-ABS-KEY (sensor)) AND PUBYEAR > 2018 AND PUBYEAR < 2025 AND (LIMIT-TO (DOCTYPE, “ar”) OR LIMIT-TO

(DOCTYPE, "cp")) AND (LIMIT-TO (SRCTYPE, "j") OR LIMIT-TO (SRCTYPE, "p")).
- ACM: [Abstract: activity recognition] AND [[Abstract: "self-supervised"] OR [Abstract: "contrastive learning"]] AND [Full Text: sensor] AND [E-Publication Date: (01/01/2019 TO 07/31/2024)].

The initial search, conducted in July 2024, yielded 210 records, which were reduced to 160 unique studies after removing 50 duplicates. After the title and abstract screening, 75 review papers and unrelated studies were excluded. Full-text reading of the remaining 85 studies followed, and we ultimately selected 38 studies. In addition to the database search, a manual search was conducted using Google Scholar, guided by relevant references and the previously mentioned terms. This additional search resulted in the selection of 5 studies, bringing the total number of studies for the review to 43 for data extraction and analysis. The screening process and reasons for exclusion are outlined in Figure 1, along with the inclusion and exclusion criteria:

- **First screen (title and abstract)**: Exclude studies related to video, camera, skeleton, 3D, visual, image, audio, sound, acoustic, speech, emotion/face recognition.
- **Full-text assessment**: Include studies that explicitly implement contrastive SSL on activity recognition, wearable sensors, environmental sensors, and the latest subject paper from the same author. Only peer-reviewed studies were included. Excluded studies were out of scope, had no available PDF, were shorter than seven pages, or had limited experimental results.

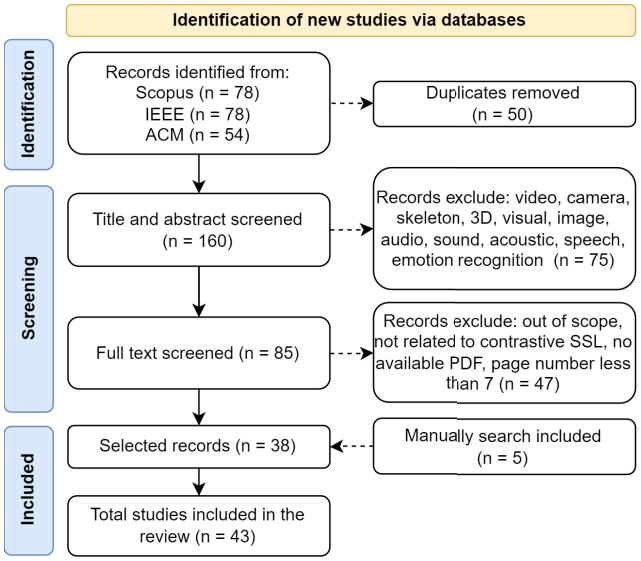


FIGURE 1. The PRISMA diagram for the papers screened and included / excluded in the review.

III. A TAXONOMY OF CONTRASTIVE SELF-SUPERVISED LEARNING FOR HUMAN ACTIVITY RECOGNITION

Recently, contrastive SSL has become a prominent approach for sensor-based HAR. The general HAR workflow consists of four key steps, as shown in Figure 2. First, sensor data are collected from various sources, such as Wi-Fi, accelerometers, gyroscopes, motion sensors, etc. This initial step applies to all categories of contrastive SSL, where data from different sensor types are prepared for further analysis. Next, the sensor data are preprocessed using data cleaning, sliding windows, transformation, and sampling techniques. Preprocessing is essential for all types of contrastive SSL approaches, as it prepares the data for further representation learning. After preprocessing, the unlabeled sensor data are fed into contrastive SSL models (e.g., SimCLR, MoCo) to learn representations during pre-training. Each category in the taxonomy applies different strategies in this phase. Finally, the model is fine-tuned with a limited amount of labeled data for downstream tasks, utilizing a classifier for activity classification. This step applies to all contrastive SSL approaches, where the pre-trained representations are refined using limited labels, improving the accuracy of HAR.

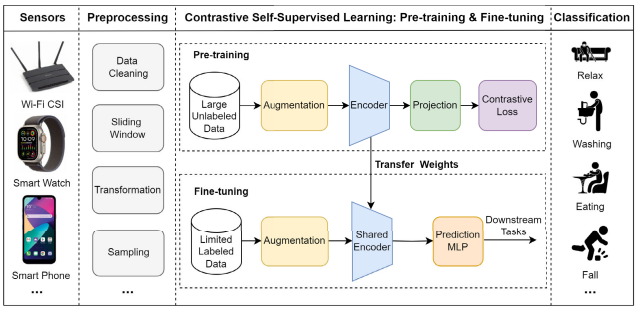


FIGURE 2. Overview of contrastive self-supervised learning for human activity recognition.

Contrastive SSL is a discriminative approach that learns representations by explicitly encouraging the model to distinguish between similar (positive) and dissimilar (negative) pairs of samples in the embedding space [22]. This approach is typically achieved using a contrastive loss function, such as NT-Xent (Normalized Temperature-scaled Cross Entropy Loss) [11] or InfoNCE (Information Noise-Contrastive Estimation) [10] loss, which measures the similarity between different instances of the same data point compared to instances of different data points [10], [23]. The following subsections will introduce recent contrastive learning (CL) methods and their applications in HAR, categorized into instance discrimination, predictive learning, multiview contrast, clustering feature representations, and siamese networks. Each subsection begins by presenting well-known contrastive SSL frameworks, followed by methods that extend or are inspired by these frameworks, applied in HAR using wearable or ambient sensors, as shown in Table 1. Table 1 provides a summary of prominent contrastive SSL applications for sensor-based HAR from selected review papers, with detailed performance metrics provided in Tables 4, 5, and 6, and further explored in the discussion section.

TABLE 1. A summary of recent contrastive self-supervised models in HAR. ABC: Attention-Based Classifier; AL: Adversarial Learning; AM: Autoregressive Model; CA: Contrastive Adaptation; CAT: Category; CC: Causal Convolutions; CCL: Contextual Contrasting Loss; CCLoss: Contrastive Center-loss; CE: Cross Entropy; CL: Contrastive Learning; CMCrr: Cross-modality Correlation; CML: Confidential Marginal Loss; CMFL: Cross-Modal Feature Contrastive Loss; CFCL: Cross-Modal Feature Contrastive Learning; CFL: Contrastive Fusion Learning; CFLoss: Contrastive Fusion Loss; CS: Contrastive Supervision; CT: Capsule Transformer; DCT: Deep Convolutional Transformer; DFWS: Device-free Wireless Sensing; Disc: Discrimination; ETE: End-to-End; FC: Fully Connected Layer; FCN: Fully Convolutional Network; FSC: Fewshot Calibration; GRN: Gated Residual Network; GSL: Geometric Structural Loss; GSS: Geometric Self-Supervised; IMD: Intra-modality Discriminator; KD: Knowledge Distillation; LSTM: Long Short-Term Memory Network; MB: Memory Bank; MCAT: Multi-scale Convolution Augmented Transformer; MCL: Multilevel Contrastive Loss; MI: Mutual Information; MIM: Mutual Information Maximization; MLP: Multi-layer Perceptron; MTL: Multi-Task Learning; RMSPO: RMSPropOptimizer; SGD: Stochastic Gradient Descent; SubCAT: Subcategory; SS-TCN: Single-Stage Temporal Convolutional Network; SVM: Support Vector Machine; TCL: Temporal Contrastive Loss; TPN: Transformation Prediction Network; VF: Virtual Fusion; VICReg: Variance-Invariance-Covariance Regularization.

CAT	SubCAT	Method	Framework	Encoder	Classifier	Optimzer	Loss Function
Instance Disc	MB	MoCoHAR [15]	MoCo	DeepConvLSTM	MLP	Adam	InfoNCE
	ETE	STF-CSL [24]	SimCLR	STFNet	Linear	Adam	NT-Xent
		SemiC-HAR [17]	SimCLR	STFNet	Linear	Adam	Contrastive Loss
		Taghanaki et al. [25]	SimCLR	2D-CNN	MLP	Adam	NTXent
		DT [18]	SimCLR	CNN+Transformer	MLP	LARS	NDT-Xent
		SimCLRHAR [15]	SimCLR	DeepConvLSTM	MLP	Adam	NT-Xent
		CSSHAR [19]	SimCLR	1D-CNN+ Transformer	MLP	LARS	NT-Xent
		CSSHAR-TFA [26]	SimCLR	1D-CNN+ Transformer	MLP	LARS	NT-Xent
		AttCLHAR [27]	SimCLR	CNN+LSTM +Attention	MLP	Adam	NT-Xent
		CoTMix [28]	Temporal Mixup +CA	CNN	FC	Adam	InfoNCE
		CoDEm [29]	CL+Attention	Convolution	FC	Adam	CCLoss
		TS-TCC [30]	Temporal+ Contextual CL	CNN	Linear	Adam	TCL+CCL
		DABaCLT [31]	DAB-aware CL	CNN	Linear	Adam	DABMinLoss
		AutoCL [32]	Siamese	FCN	MLP	Adam	NT-Xent
		FusionCL [33]	Time-Frequency Fusion-Augmentation based CL	1D-ResNet	FC	Adam	NT-Xent
		CapMatch [34]	CL+ Feature-based KD	CT	-	RMSPO	KD Loss+ CL Loss+CML
		TFCL [35]	AM+CL	CNN+GRN	Linear	Adam	InfoNCE
		AutoFi [36]	GSS+FSC	CNN	FC	SGD	GSL
		CoS [37]	CS	CNN	FC	SGD	CE
		MS-TCN [38]	Multi-scale CL	SS-TCN	-	-	MCL
Predictive	-	CPCHAR [39]	CPC	1D-Conv	MLP	Adam	InfoNCE
		Enhanced CPC [40]	CPC	CNN	MLP	Adam	InfoNCE
		TSCP2 [41]	CPC	TCN	MLP	-	InfoNCE
Multiview	-	COCOA [7]	Cross-modality CL	CNN	Linear	Adam	CMCrr+IMD
		ColloSSL [42]	Multi-view CL	1D-CNN	FC	SGD	Multi-view Contrastive Loss
		CAGE [43]	Two Stream CNN	CNN	MLP	Adam	CE
		CroSSL [44]	Cross-modal SSL	1D-CNN	MLP	Adam	VICReg
		CMC-TFA [26]	CMC	1D-CNN	MLP	Adam	Contrastive Loss
		ModCL [45]	Intra/Inter-ModCL	ResNet	MLP	Adam	NT-Xent
		AFVF [46]	VF+CL	1D-ResNet	-	Adam	CE+ Multiview NT-Xent
		Cosmo [47]	CFL	CNN+RNN	ABC	-	CFLoss
		MESEN [48]	CFCL	1D-CNN	Linear	-	CMFL
Cluster	-	ClusterCLHAR [14]	SimCLR+Cluster	TPN	Linear	Adam	Cluster-NTXent
Siamese	-	SS-HAR [49]	BYOL	CNN	SVM	SGD	MSE
		DCTCSS [50]	BYOL	DCT	Linear	Adam	MSE
Maximize MI	-	Li et al. [51]	JL+CL	ResNet	MLP	Adam	InfoNCE+CE
		DualConFi [52]	Dual-stream CL	MCAT+1D-CNN	MLP	Adam	NT-Xent
		Chen et al. [53]	MIM-based DFWS	CNN	Neurons	-	Contrastive Loss
Hybrid	-	CogAx [54]	CL+MTL	CNN+FC	-	Adam	Contrastive Loss+CE
		CALDA [55]	CL+AL	CNN	MLP	-	Multiple-positive InfoNCE+CE

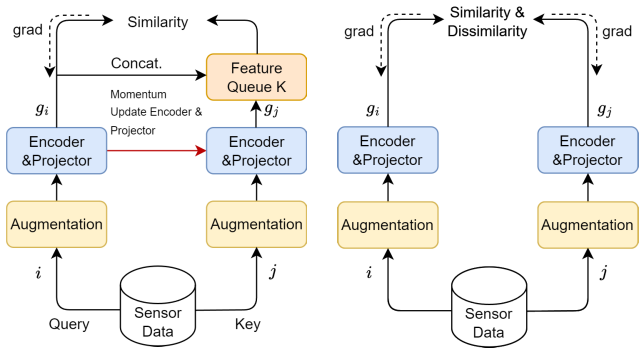
A. INSTANCE DISCRIMINATION

This subsection first introduces the two subcategories of instance discrimination: memory bank and end-to-end. Then, it illustrates the latest applications of instance discrimination for HAR.

1) MEMORY BANK

The Instance Discrimination (InstDisc) approach [56] is designed to learn embedding representations by treating each image as its own class, facilitating instance-level discrimination. This is achieved using noise-contrastive estimation

(NCE) to effectively distinguish between a large number of individual instances. Additionally, InstDisc incorporates a feature memory bank to store the representations of all instances. In this framework, it is not semantic labeling, but rather the inherent similarities within the data, that determines the proximity of different classes. The ultimate objective is to learn a feature space where different augmentations (or instances) of the same image are closely aligned, while instances of different images are distinctly separated.



(a). MoCo (b). SimCLR

FIGURE 3. Comparison of the MoCo and SimCLR structures.

The InstDisc framework serves as a foundation for subsequent advancements in CL, paving the way for more refined methodologies such as MoCo [12], [57], [58], as shown in Figure 3 (a). It outlines the same work and replaces the traditional memory bank with a dynamic dictionary and a queue mechanism, allowing for storing a substantial number of negative samples without necessitating costly memory operations or the recomputation of embeddings for the entire dataset. Furthermore, MoCo introduces a momentum encoder that updates more gradually than the primary encoder. This slow update rate helps maintain the consistency of representations within the dynamic dictionary, thereby stabilizing the training process. MoCo also employs the contrastive loss function InfoNCE to measure sample similarities effectively. Due to its dynamic dictionary, MoCo can utilize a significantly larger set of negative samples compared to InstDisc, facilitating the learning of more discriminative features. This enhancement over InstDisc underscores MoCo's ability to improve self-supervised representation models' learning efficiency and quality.

MoCo v2 [57] draws inspiration from SimCLR [11] to enhance the original MoCo framework. These enhancements include integrating an MLP (multi-layer perceptron) projection head and incorporating more extensive data augmentation techniques, improving performance over MoCo and SimCLR without requiring large training batches. MoCo v3 [58] marks a significant shift by replacing the conventional ResNet [59] backbone with transformer components [60], reflecting the adaptation to Vision Transformer (ViT) models. Alongside this architectural change, MoCo v3 also fine-tunes data augmentation strategies and hyperparameters to suit vision training better. Collectively, these advancements signify substantial progress in the MoCo series.

Pretext-Invariant Representation Learning (PIRL) [61] is another prior work that utilizes a memory bank to store negative samples for contrastive learning. PIRL learns invariant representations based on pretext tasks, substantially improving the semantic quality of the learned image representations.

2) END-TO-END

InvaSpread [62] utilizes a Siamese network to map inputs into low-dimensional, normalized embeddings via a CNN backbone, ensuring that augmented versions of the same instance remain invariant while separating different instances. This is achieved through a unique softmax embedding method applied directly to actual instance features, eliminating the need for storing numerous negative instances in structures like memory banks or queues by using instances from the same batch instead.

SimCLR [11], shown in Figure 3 (b), similarly to InvaSpread, learns visual representations from both positive and negative instances within the same batch. Additionally, it utilizes a larger batch size, a varied set of data augmentations, and a learnable nonlinear projection head. This projection head maps representations to a space where contrastive loss, NT-Xent loss, is applied to enhance the quality of feature learning by maximizing agreement between different augmentations of the same instance. These additional components allow SimCLR to outperform InvaSpread. SimCLR v2 [63] builds upon the original SimCLR by introducing unsupervised pre-training of a larger ResNet model, followed by supervised fine-tuning with a few labeled examples. Additionally, it employs distillation using unlabeled examples to refine and transfer task-specific knowledge. These modifications enhance the efficiency and effectiveness of semi-supervised learning.

3) INSTANCE DISCRIMINATION IN HUMAN ACTIVITY RECOGNITION

Instance discrimination is a dominant contrastive SSL method for sensor-based HAR. MoCoHAR [15] extends MoCo [12] into HAR-suitable frameworks, utilizing DeepConvLSTM as the backbone network and InfoNCE as the contrastive loss function for its pre-training task. STF-CSL [24] introduces a framework called Short-Time Fourier Neural Networks (STFNets)-based contrastive self-supervised representation learning from a time-frequency perspective, rather than purely from a time-domain (CNN-based approach). This framework utilizes both time-domain and frequency-domain data augmentation techniques during SSL. SemiC-HAR [17] efficiently utilizes both labeled and unlabeled data during the pre-training process, combining the advantages of supervised and self-supervised frameworks. Taghanaki et al. [25] design two distinct pipelines: one processes data in the time-frequency

domain, and the other focuses solely on the time domain. Initially, each pipeline employs self-supervised CL for training. Subsequently, each stream is fine-tuned for final classification tasks.

While the original SimCLR framework scales similarities between features of augmented views using a constant temperature parameter, DT [18] introduces a method to dynamically compute temperature values for scaling within a contrastive loss function. This dynamic temperature approach is based on instance-level similarity values extracted by an additional model pre-trained on initial instances beforehand. SimCLRHAR [15] proposes a novel sensor data augmentation method through resampling. This technique introduces variability and simulates realistic activity data by altering the sampling frequency, enhancing the model's ability to generalize across different scenarios.

CSSHAR [19] is first trained on unlabeled data within the pretext task. Later, the encoder parameters are frozen, and only the prediction model is fine-tuned on the entire training set with the original activity labels. CSSHAR-TFA [26] enhances the SimCLR framework by integrating a dynamic time warping (DTW) algorithm within the latent space. This addition forces features to align along the temporal dimension, ensuring that temporal relationships are maintained in the learned representations.

CoTMix [28] proposes a novel temporal mixup strategy to generate two intermediate augmented views for the source and target domains. It then leverages CL to maximize the similarity between each domain and its corresponding augmented view. CoDEm [29] initially uses CL to create domain embeddings that handle inter-domain variability and similarities from raw samples and metadata. CoDEm then learns label embeddings, conditioned on these domain embeddings, and enhanced with a novel attention mechanism, improving activity classification. TS-TCC [30] introduces a time-series representation learning framework using temporal and contextual contrasting to learn from unlabeled data with CL. It utilizes specific augmentations to learn robust temporal relations and discriminative representations.

DABaCLT [31] introduces a data augmentation bias (DAB) aware CL framework for time series representation, leveraging a raw features stream (RFS) to extract features from raw data and incorporating a DAB-minimizing loss function (DABMinLoss) within the contrasting module to minimize the bias of the extracted temporal and contextual features. AutoCL [32] develops an end-to-end auto-augmentation CL method that implements high-performance learning through an auto-augmentation architecture, correlation reduction strategy, and stop-gradient design. CapMatch [34] leverages pseudo-labeling, CL, and feature-based knowledge distillation (KD) techniques to construct similarity learning on both lower and higher-level semantic information extracted from two augmented versions of the data, *weak* and *timecut*. This approach aims to recognize the relationships among the features of classes in the unlabeled data. FusionCL [33] presents a time-frequency fusion-augmentation based CL method for self-supervised HAR, combining time-domain augmentation (TNet) and frequency-domain augmentation (FNet) to generate diverse samples.

The Temporal Fusion Contrastive Learning (TFCL) [35] method integrates a sensor wearer's biometric information into the CL process. This biometric information is extracted through a static encoder and fed into an observed encoder as input values. Subsequently, the static information is re-extracted, along with the latent representation from the observed encoder. The resulting context vector is then used to predict future latent representations and learn temporal features. Finally, the contextual similarity of representations for identical samples is enhanced by evaluating each augmentation method. AutoFi [36] fully utilizes unlabeled low-quality CSI samples to transfer knowledge to user-defined tasks, enabling cross-task transfer in Wi-Fi sensing. Contrastive Supervision (CoS) [37] aims to learn time series augmentation invariances by forcing positive pairs to be nearby and negative pairs to be far apart at different depths of a neural network. The approach generalizes CL in a supervised setting, where the contrastive loss is used to supervise the intermediate layers instead of only the last layer. This approach allows for more effective leveraging of label information to better fuse multi-level features. The Multi-Stage Temporal Convolutional Network (MS-TCN) [38] introduces sample-level and segment-level contrast to learn a well-structured embedding space, resulting in improved activity segmentation and recognition performance.

B. PREDICTIVE LEARNING

This subsection explores contrastive predictive learning and its application in HAR.

1) CONTRASTIVE PREDICTIVE CODING

The Contrastive Predictive Coding (CPC) [10] framework comprises three major components, as shown in Figure 4: the encoder, the autoregressive model, and the future timestep prediction model. CPC utilizes autoregressive models like gated recurrent units (GRU) [64] and self-attention networks [60] for extracting useful representations from high-dimensional data by predicting future latent space sequences. It employs a probabilistic contrastive loss, InfoNCE, to maximize a lower bound on mutual information for positive samples while minimizing it for negative samples. CPC is effective in various domains, such as NLP and image recognition. CPC v2 [65], also inspired by SimCLR, enhances the original CPC by expanding model capacity by increasing the depth and width of the model and replacing batch normalization with layer normalization. Additionally, it extends prediction across all four spatial directions (bottom, up, left, and right) and applies patch-based augmentations to enhance learning effectiveness. These improvements significantly boost performance in various tasks, particularly in

data-efficient image recognition, without relying on labeled data.

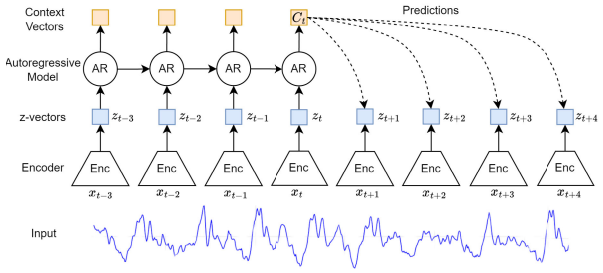


FIGURE 4. Overview of contrastive predictive coding. The architecture comprises three components: the encoder, the autoregressive model, and the prediction model. The figure is adapted from [10].

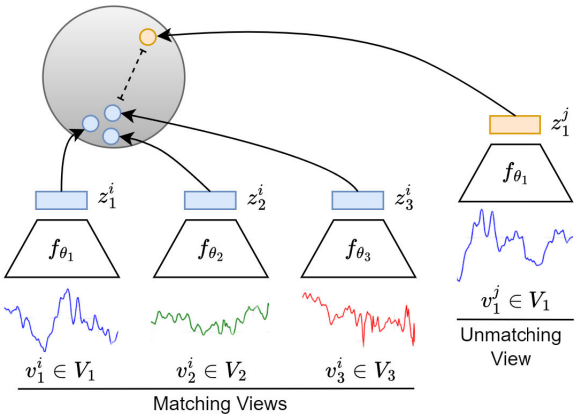


FIGURE 5. Overview of contrastive multiview coding. The figure is adapted from [66].

2) PREDICTIVE LEARNING IN HUMAN ACTIVITY RECOGNITION

CPCHAR [39] adopts the CPC framework to learn effective sensor representations in an unsupervised manner by leveraging the temporal characteristics of body-worn sensor data. This approach predicts multiple future time steps of encoded data from a context vector derived from past data, creating representations that capture high-level information between temporally separated parts of the time-series signal. Building on this, the enhanced CPC [40] proposes improvements to the original CPC framework by systematically enhancing the encoder convolutional layers with higher striding, replacing the GRU aggregator with causal convolutions to summarize the latent vectors into a context vector for the aggregator network, and performing future timestep predictions for each context vector. These enhancements result in improved downstream activity recognition performance. The self-supervised Time Series Change Point Detection (TSCP2) [41] method is based on CPC and exploits the local correlation present within a time series by learning a representation that maximizes the shared information between contiguous time intervals. This approach detects change points in the dataset by identifying transitions between physical activities such as staying, walking, and jogging.

C. MULTIVIEW CONTRAST

This subsection illustrates multiview contrast models and their applications in HAR.

1) MULTIVIEW

Contrastive Multiview Coding (CMC) [66] learns robust representations from multiple views of the same input, using CL loss to maximize the mutual information between different perspectives of the same scene, as shown in Figure 5. It employs a memory bank to store latent features, enabling efficient pairing of negative samples with positive ones without requiring feature recomputation. CMC demonstrates that incorporating more views can enhance the performance of representation capture, and the contrastive objective proves superior to cross-view prediction.

Tian et al. [67] investigate both theoretically and empirically the importance of view selection in CL across multiple data views. They introduce the InfoMin principle, which states that optimal views should share the minimal information necessary to perform well on the downstream task. They suggest minimizing mutual information (MI) between views while preserving task-relevant information, utilizing both unsupervised and semi-supervised frameworks. Their findings also indicate that stronger data augmentation reduces mutual information, aligning with the InfoMin principle and improving downstream classification accuracy.

2) MULTIVIEW CONTRAST IN HUMAN ACTIVITY RECOGNITION

COCOA (Cross mOdality COntrastive leArning) [7] employs a novel objective function to learn high-quality representations from multisensor data by computing cross-correlation between different data modalities and minimizing the similarity between irrelevant instances. This approach, known as cross-modal or cross-view, is a form of multimodal learning where different modalities serve as supervisory signals for each other. ColloSSL (Collaborative Self-Supervised Learning) [42] captures unlabeled sensor data simultaneously from multiple devices, treating these datasets as natural transformations of one another to generate a supervisory signal for representation learning. This method utilizes three technical innovations to adapt conventional SSL algorithms for a multi-device context.

CAGE (Contrastive Accelerometer-Gyroscope Embedding) [43] introduces a two-stream convolutional neural network (CNN) model that processes accelerometer and gyroscope signals independently. Each modality is analyzed separately before its modality-specific features are fused at the feature level for recognition tasks. The model interprets the accelerometer and gyroscope signals, collected simultaneously from sensors worn by a user, as dual physical

representations of an individual's actions. CroSSL (Cross-modal SSL) [44] utilizes masking of intermediate embeddings produced by modality-specific encoders, followed by their aggregation into a global embedding through a cross-modal aggregator. This aggregated embedding is then fed into downstream classifiers, facilitating end-to-end cross-modal learning and efficiently handling missing modalities without prior data preprocessing or negative-pair sampling typical in CL. While each modality has inherent limitations, CMC-FTA [26] enhances the contrastive SSL framework CMC by integrating a Dynamic Time Warping (DTW) algorithm within the latent space to align features along the temporal dimension.

ModCL (Modality Consistency-guided Contrastive Learning) [45] exploits both intramodality and intermodality consistency of wearable device data to construct CL tasks. This approach encourages the recognition model to identify similar patterns and distinguish dissimilar ones. By leveraging these mixed constraint strategies, ModCL learns inherent activity patterns and extracts meaningful, generalized features across different datasets. AFVF (Actual Fusion within Virtual Fusion) [46] takes advantage of unlabeled data from multiple time-synchronized sensors during training. CL is adopted to exploit multimodal correlation, aiding unimodal classification. Cosmo [47] proposes contrastive fusion learning with small data in multimodal HAR applications, effectively extracting both consistent and complementary information across different modalities for efficient fusion by integrating novel fusion-based CL and quality-guided attention mechanisms. MESEN (Multimodal Empowered Unimodal Sensing Framework) [48] utilizes unlabeled multimodal data available during the HAR model design phase for unimodal HAR enhancement during the deployment phase. It exploits this data to extract effective unimodal features for each modality through a multi-task mechanism that integrates cross-modal feature CL and multimodal pseudo-classification aligning.

D. CLUSTERING

This subsection includes two contrastive clustering models, DeepCluster and SwAV, and contrastive clustering applications for HAR.

1) DeepCluster AND SwAV

DeepCluster [68] proposes a novel clustering approach for the large-scale, end-to-end training of convolutional networks. DeepCluster iteratively groups features using a standard clustering algorithm, k-means, and uses the resulting assignments as supervision to update the network weights.

Swapping Assignments between Views (SwAV) [69] clusters data while enforcing consistency between cluster assignments of different augmentations (views) of the same image instead of directly comparing features of augmented images, as shown in Figure 6. It employs a swapping mechanism to predict a view's code from another view's representation. This strategy makes SwAV more computationally efficient than other CL models that require large memory banks or special momentum networks. SwAV also illustrates a multi-crop augmentation strategy to enhance performance without increasing memory requirements.

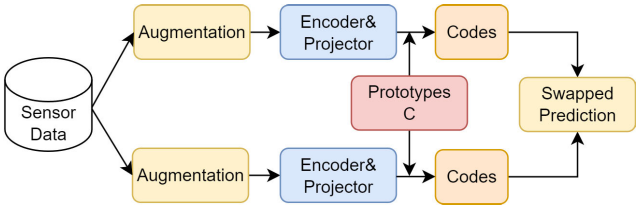


FIGURE 6. The structure of SwAV.

2) CLUSTERING IN HUMAN ACTIVITY RECOGNITION

Instance discrimination can lead to over-clustering, where representations of instances from the same class are overly separated [14]. ClusterCLHAR (Clustering for Contrastive Learning in Human Activity Recognition) [14] proposes a new CL framework that selects negative pairs through clustering in HAR. Initially, ClusterCLHAR clusters the instance representations, considering only those from different clusters as negatives. Subsequently, it introduces a novel contrastive loss function designed to exclude same-cluster instances from the negative pairs, thereby refining the learning process to better preserve intra-class similarity while enhancing inter-class differentiation.

E. SIAMESE NETWORKS

This subsection presents three Siamese networks: BYOL, SimSiam, and Barlow Twins. The applications of Siamese networks for HAR are also introduced.

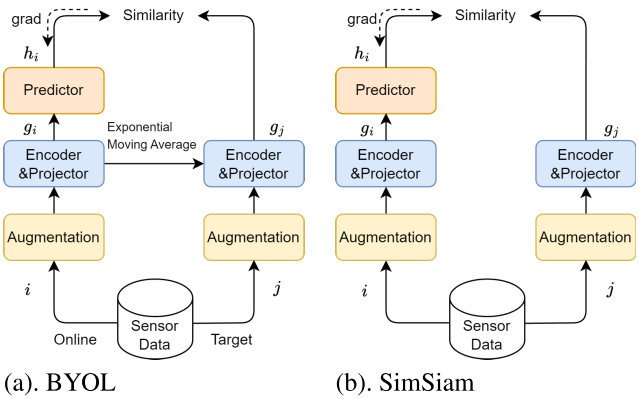


FIGURE 7. The structures of BYOL and SimSiam.

1) BYOL

BYOL [13] is a novel SSL approach for representation learning without negative pairs, as shown in Figure 7 (a). It utilizes two neural networks, an online network and a target network, that interact and learn from each other. BYOL

trains the online network to predict the target network's representation of differently augmented views of the same image. Meanwhile, the target network is updated with a slow-moving average of the online network using a stop-gradient method. The loss between the online predictions and target projections is measured using mean squared error (MSE).

2) SimSiam

Simple Siamese network (SimSiam) [23] learns meaningful representations without negative sample pairs, large batches, or momentum encoders, as shown in Figure 7 (b). It employs two identical networks to process distinct augmentations of the same input, with one predicting the other's features while applying a stop-gradient to prevent collapse. SimSiam demonstrates that simple Siamese architectures for unsupervised representation learning can achieve competitive results without complex networks.

3) BARLOW TWINS

Barlow Twins [70], shown in Figure 8, neither requires large batches nor asymmetry between the network twins, such as gradient stopping or moving average updates on the weights. It utilizes an objective function that prevents collapse by measuring the cross-correlation matrix between the embeddings of two identical networks fed with distorted versions of the same sample, aiming to make this matrix close to the identity.

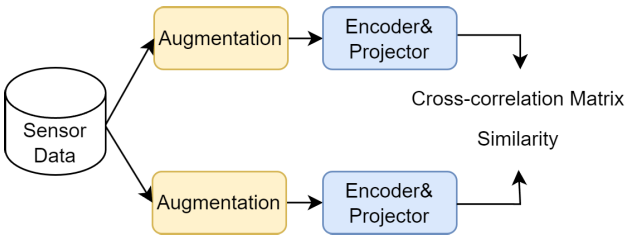


FIGURE 8. The structure of Barlow Twins.

4) SIAMESE NETWORKS IN HUMAN ACTIVITY RECOGNITION

SS-HAR [49] initially takes the unlabeled data generated by data augmentation as the input of the network, explores the supervised information of the unlabeled data through SSL, and implements the obtained backbone network as a feature extractor to extract activity features for subsequent classification. The DCTCSS (Deep Convolutional Transformer-based Contrastive Self-Supervised) [50] model is built under the BYOL framework and utilizes a deep convolutional transformer (DCT) as the backbone. It aims to achieve reliable activity recognition using only a small amount of labeled data.

F. MAXIMIZE MUTUAL INFORMATION

This subsection explores the Deep InfoMax model and the applications of maximizing mutual information in HAR.

1) DEEP InfoMax

Deep InfoMax (DIM) [71] learns unsupervised representations by maximizing mutual information. By incorporating knowledge about the input's locality into the objective, DIM significantly enhances the suitability of representations for downstream tasks. The mutual information maximization procedure allows prioritization of global or local information, which can be tuned to optimize the learned representations for specific applications, such as classification or reconstruction tasks.

2) MAXIMIZE MUTUAL INFORMATION IN HUMAN ACTIVITY RECOGNITION

Li et al. [51] introduce an optimization strategy that leverages mutual information maximization to directly optimize the distance or similarity between samples, thereby controlling the shape of the decision boundary for recognizing complex activities. This approach utilizes contrastive loss to maximize the mutual information between similar samples, emphasizing the crucial role of positive pair selection in effectively learning decision boundaries. The training framework is based on joint learning (JL), which trains labeled samples with the assistance of unlabeled samples to optimize the decision boundary and learn the representation of sensor-based HAR.

DualConFi [52] proposes a novel mutual information-enhanced dual-stream CL model to leverage the advantages of CL techniques in HAR tasks. It consists of two streams of CL (e.g., CL for temporal and spatial information) to extract highly discriminative representations for activity classification. Chen et al. [53] design a CL framework to obtain features of unlabeled samples by maximizing the mutual information between features and corresponding samples, addressing device-free wireless sensing for HAR.

G. HYBRID

This subsection introduces the definition of hybrid contrastive models and the applications of various hybrid models based on CL for HAR.

1) HYBRID MODELS

Hybrid models combine the advantages of various SSL models, such as contrastive, generative, and predictive models, which may operate in parallel or collaborate to enhance self-supervision signals. These models typically require multiple encoders and projection heads [72].

2) HYBRID MODELS IN HUMAN ACTIVITY RECOGNITION

CogAx [54] employs a combined contrastive and multi-task learning framework to derive high-level motion-related representations from accelerometer data. This approach is specifically designed to more effectively correlate with the underlying functional and cognitive health parameters of older adults.

CALDA (Contrastive Adversarial Learning for Multi-Source Time Series Domain Adaptation) [55] integrates the methodologies of CL and adversarial learning (AL) to effectively facilitate multi-source unsupervised domain adaptation (MS-UDA) for time series data. CALDA employs adversarial training to align feature-level distributions across different domains and utilizes CL to enhance accuracy in the target domain by leveraging cross-source label information.

IV. CONTRASTIVE SELF-SUPERVISED LEARNING FRAMEWORK

Although contrastive SSL encompasses a variety of architectures, they share several key components, including data augmentation, encoders, projection heads, classifiers, and loss functions. In this section, we will detail these components.

A. DATA AUGMENTATION

Data augmentation is a crucial component of CL, enabling the acquisition of high-quality, generalizable representations. Different data augmentation operations can lead to varying performances in learning representations [24]. Analogous to image data enhancement methods, temporal dependency plays a significant role in time-series data [35]. Although most existing data augmentation methods consider the time-domain perspective, applications involving acceleration, vibration, or wireless signal propagation are fundamentally based on signal frequencies. Consequently, these applications often have sparser and more compact representations in the frequency domain [24]. We present an overview of commonly used data augmentation functions in sensor-based contrastive SSL and categorize them into two domains: time and frequency, as shown in Table 2.

Some data augmentation operations, such as *noise*, *scaling*, and *low-pass filtering*, can be applied to tasks in both the time and frequency domains, while others, such as *frequency-domain perturbation* and *phase shift*, are more task-specific. For example, these operations may involve applications using Wi-Fi Channel State Information (CSI). STF-CSL [24] leverages the hypothesis that the underlying physics of measured phenomena in the Internet of Things (IoT) applications are more succinctly and compactly expressed in the frequency domain. It applies data augmentation operations in both the time and frequency domains to enhance its approach.

Due to the sensitivity of contrastive methods to the choice of augmentations, selecting appropriate augmentations is critical for CL techniques [11]. TS-TCC [30] analyzes suitable augmentations for the CL problem, employing both strong and weak augmentations to introduce data variations. This approach enhances the model's generalization ability, improving performance on unseen test sets. SimCLRHAR [15] demonstrates that the performance of data augmentation through resampling is superior to other augmentation functions such as noise addition, rotation, and scaling. This resampling technique introduces variability in domain information and simulates realistic activity data by altering the sampling frequency. The goal is to maximize the coverage of the sampling space, thereby enhancing the model's ability to generalize across different scenarios.

AutoCL [32] aims to reduce the augmentation burden by proposing an end-to-end auto-augmentation CL method for wearable-based HAR. However, data augmentation methods during training can distort the raw data distribution. This discrepancy between the representations learned from augmented data in CL and those obtained from supervised learning leads to an incomplete understanding of the information in the real data by the trained encoder [31]. To mitigate the influence of data augmentation bias (DAB), DABaCLT [31] introduces a DAB-aware CL framework for time series representation. This framework leverages a Raw Features Stream (RFS) to extract features from raw data, combined with augmented data to create positive and negative pairs for DAB-aware CL. Additionally, it introduces a DAB-minimizing loss function (DABMinLoss) within the contrastive module to minimize the DAB of the extracted temporal and contextual features.

B. ENCODER

An encoder extracts representation vectors from augmented data examples for subsequent classification tasks.

CNNs are most popularly used in HAR to learn hierarchical representations from time-series data, as illustrated in Table 1. The CNN configurations vary to optimize the parameters that differ across models, including the number of convolutional layers, kernel sizes, filter counts, dropout rates, and max pooling layers. Taghanaki et al. [25] employ a 2D CNN in a scalogram encoder for time-frequency signals represented as 2D matrices, while the signal learner uses a 1D convolution-based neural network encoder to process augmented raw signals in the time domain.

CNNs are often combined with LSTMs (Long Short-Term Memory Networks) to effectively capture temporal dependencies, which is crucial for accurately recognizing activities from sequential sensor data in HAR. Zou et al. [76] extract local features of time-series data through convolution blocks and global features through an LSTM module as the backbone. SimCLRHAR [15] employs DeepConvLSTM as the base encoder. ClusterCLHAR [14] utilizes the Transformation Prediction Network (TPN) as the backbone network, known for its fast inference and superior performance in supervised learning compared to DeepConvLSTM.

The self-attention mechanisms in transformer encoders allow each position to attend to other positions in the sequence, which is especially beneficial for HAR. CSSHAR [19] combines a one-dimensional CNN with a transformer encoder to enhance CL performance. The DT [18] encoder comprises three CNN layers followed by positional encoding and transformer self-attention, training on augmented views of data instances to generate view-level features and similarities between them.

TABLE 2. A summary of the surveyed papers on data augmentation functions. A: Accelerometer; C: CSI; Dom.: Domain; E: ECG; F: Frequency; G: Gyroscope; HFC: High Frequency Component; LFC: Low Frequency Component; T: Time; W: Wi-Fi.

Method	Dom.	Sensor	Description	Reference
Random Truncation	T	A, G	Randomly set a part of the signal as 0 and splice the remaining signal.	[49]
Cropping	T	A, G	Randomly crop the raw sample according to a certain time window size.	[73]
Geometric Scaling	T	A, G	Randomly scaling each axis of the signal.	[49]
Permutation	T	A, G	Split signals into a number of intervals and randomly permutes them.	[16], [18], [19], [24], [25], [32], [35], [37], [52], [74]
Shuffle	T	A, G	Channel shuffling based on axis dimension.	[25], [35], [37]
Channel Shuffle	T	A, G	Shuffle channels of multivariate time-series data.	[16], [19]
Rotation	T	A, G	A method to simulate different sensor positions involves applying a random 3-D axis and rotation to the sample.	[15], [16], [18], [19], [24], [25], [35], [37], [46], [74]
Inversion/Negate	T	C	Multiply the input data with -1.	[15], [16], [24], [25], [35], [52]
Horizontal Flipping	T	A, G	Reverse data along the time-direction.	[24]
Channel Sobel	T	C	Process random channels with 1D Sobel operator for consistent spatial augmentations along the temporal dimension with unprocessed channels.	[52]
Time flip/Reversing	T	C	The entire window of the sample is flipped in the time direction.	[15], [16], [25], [35], [37], [52]
Magnitude Warping	T	A, G	Warp the magnitude of a window-length data with a smooth scalar around 1 which can add smoothly varying noise to samples.	[46], [74]
Time-warping	T	A, G	Generat a new data example with the same label by stretching and warping the time intervals of the input time-series.	[24], [25], [35], [46], [74]
Random Sampling	T	A, G	Similar to time-warping, but uses sub-samples for interpolation.	[37], [74]
Down-sampling	T	A, G	Sample signal by a random time interval.	[49]
Linear Upsampling	T	A, G	Interpolating physical states.	[15]
Magnify	T	A, G	Multiply by a random scalar to magnify the size of the data in the window to simulate stronger amplitude motion.	[15]
Jitter/Noise	T&F	A, G, C, E	Add random noise (Gaussian) to signals.	[15], [16], [18], [19], [24], [25], [32], [35], [37], [49], [74]
Scaling	T&F	A, G	Multiply by normally distributed values.	[15], [16], [18], [19], [24], [25], [32], [35], [37], [74]
Low-pass Filter	T&F	A, G, C	Filtered by various low-pass filters.	[24], [49]
Amplitude and Phase Perturbation	F	A, G, C	Randomly select segments of the frequency domain data are perturbed by Gaussian noise.	[24]
Phase Shift	F	A, G, C	Perturb phase spectrum values by Gaussian noise.	[24], [75]
Frequency-domain + Perturbation	F	W	Replace the values of all points in the selected segment with Gaussian distribution.	[17]
HFC	F	A, G	Split the low and high frequency components and reserve high frequency components.	[75]
LFC	F	A, G	Split the low and high frequency components and reserve low frequency components.	[75]

Additionally, some specialized models are employed as encoders. STF-CSL [24] utilizes an encoder based on the STFNet architecture for frequency domain learning. An STFNet block corresponds to one layer in a CNN, and the encoder is constructed by stacking multiple STFNet blocks. Each layer processes multi-dimensional time-series data, computing multiple Short-Term Fourier Transform (STFT) representations with varying time-frequency resolutions. FusionCL [33] adopts a 3-layer 1D ResNet as the backbone for its encoder, capturing temporal dependencies and generating informative embedding.

C. PROJECTOR AND CLASSIFIER

The projector serves as a crucial component following the encoder in CL, assisting in the encoder's training throughout the self-supervised process. A nonlinear projection head, typically implemented using a MLP, enhances the quality of representations derived from the encoder [14], [18], [19], [29]. AutoCL [32] employs an MLP for the projector, which includes a fully connected layer with 256 neurons and a ReLU activation function, followed by another fully connected layer with 128 neurons, concluding with a softmax activation function acting as the classifier.

In the fine-tuning phase, a classifier is added to perform downstream tasks such as classification, detection, or segmentation. Classifiers commonly employ MLPs, single fully connected layers, and linear classifiers, as indicated in Table 1. SS-HAR [49] demonstrates that the support vector machine (SVM) achieves superior classification outcomes with fewer labeled samples compared to other classifiers, such as K-nearest neighbors (K-NNs) and random forests (RF).

D. CONTRASTIVE LOSS FUNCTIONS

The contrastive loss measures the similarities between samples in the embedding space [12]. Cosine similarity is

the most common basis for various contrastive loss functions. It measures the similarity between two vectors in an inner product space.

$$\text{sim}(i, j) = \frac{i \cdot j}{\|i\| \, \|j\|} \tag{1}$$

SimCLR [11], shown in Figure 3 (b), combines cosine similarity with cross-entropy loss to form the NT-Xent loss, which is used for learning representations from normalized embeddings and an appropriately adjusted temperature parameter. The contrastive function aims to maximize similarity between positive data pairs and minimize similarity for negative data pairs [19]. A mini-batch contains N examples and defines the contrastive prediction task on pairs of augmented examples derived from the mini-batch, resulting in $2N$ data points. The loss for a positive pair of samples (i, j) can be defined as follows [3], [11]:

$$l(i, j) = -\log \frac{\exp(\text{sim}(g_i, g_j)/\tau)}{\sum_{k=1}^{2N} \mathrm{I}_{[k \neq i]} \exp(\text{sim}(g_i, g_k)/\tau)} \tag{2}$$

$$\mathcal{L}_{NT-Xent} = \frac{1}{2N} \sum_{k=1}^{N} [l(2k-1, 2k) + l(2k, 2k-1)] \tag{3}$$

where τ serves as the temperature parameter to scale the similarity scores, the function $\text{sim}(g_i, g_j)$ represents the cosine similarity between two entities. When k does not equal i, the indicator function $\mathrm{I}_{[k \neq i]}$ returns 1, ensuring that the loss calculation excludes self-comparisons. The total loss is computed across all positive pairs in the mini-batch, both $l(2k-1, 2k)$ and $l(2k, 2k-1)$, where $2k-1$ and $2k$ form positive pairs, and $2k-1$ combined with other views forms negative pairs.

SimSiam, shown in Figure 7 (b), uses contrastive loss to minimize the two views g_j and h_i through negative cosine similarity (NCS) [23]:

$$\mathcal{D}(h_i, g_j) = -\frac{h_i}{\|h_i\|_2} \cdot \frac{g_j}{\|g_j\|_2}. \tag{4}$$

where $\|\cdot\|_2$ is ℓ_2 norm. g_j denotes representations in the embedding space after encoder and h_i is output of predictor in another latent space. The projection head's output, g_j, is applied with a stop-gradient operation to prevent model collapse, ensuring that gradients are not propagated back. This procedure is symmetrically adopted for both views. The total symmetrized loss can be calculated as follows [23]:

$$\mathcal{L} = \frac{1}{2}\mathcal{D}(h_i, \text{stopgrad}(g_j)) + \frac{1}{2}\mathcal{D}(h_j, \text{stopgrad}(g_i)). \tag{5}$$

CPC and MoCo [10], [12] employ the InfoNCE loss, which is based on noise-contrastive estimation (NCE) [77] as their contrastive loss function. The MoCo framework is shown in Figure 3 (a). The purpose is to make representations of positive samples closer while pushing representations of negative samples farther apart, as illustrated below [12]:

$$\mathcal{L}_{InfoNCE} = -\log \frac{\exp(g_i \cdot g_+/\tau)}{\sum_{j=0}^{K} \exp(g_i \cdot g_j/\tau)} \tag{6}$$

where τ denotes the temperature hyper-parameter and controls the separation degree of positive and negative pairs. The keys of a dictionary include the encoded query g_i and a set of samples $\{g_0, g_1, g_2, \ldots\}$. The symbol g_+ represents a key that matches g_i in the dictionary and can be treated as the positive counterpart of g_i. The denominator comprises the sum of a positive sample g_+ and K negative samples.

The BYOL, illustrated in Figure 7 (a), contrastive loss is calculated using the mean squared error (MSE) between the normalized predictions of the online network, h_i, and the projections of the target network, g_j, after processing different augmentations of the same input. This aims to minimize the distance between the representations produced by the online network and the target network as follows [13]:

$$\mathcal{L}_{i,j} \triangleq \left\|\bar{h}_i - \bar{g}_j\right\|_2^2 = 2 - 2 \cdot \frac{< h_i, \ g_j >}{\|h_i\|_2 \cdot \left\|g_j\right\|_2}. \tag{7}$$

where $< h_i, \ g_j >$ represents the dot product between h_i and g_j and $\|\cdot\|_2$ denotes the ℓ_2 norm. Another view j, is passed through the online network to obtain the online network prediction h_j, which combines with the target network projection g_i to calculate the symmetrized loss. The losses from both perspectives comprise the total loss.

V. BENCHMARKS FOR SENSOR-BASED HUMAN ACTIVITY RECOGNITION

This section surveys the datasets available for sensor-based HAR using contrastive SSL, as shown in Table 3. There are primarily two categories of datasets used for HAR: wearable and ambient sensor-based datasets. Most datasets consist of data collected from one or multiple modalities.

A. WEARABLE SENSOR-BASED DATASETS

Wearable sensors are typically integrated into smart devices or designed as portable units unobtrusively attached to or worn on the human body [78]. These devices commonly include sensors such as accelerometers, gyroscopes, magnetometers, and physiological sensors. These sensors can be positioned at one or multiple locations on the human body, such as the chest, arms, and legs, to serve various purposes [79].

- **Accelerometers**. The accelerometer is standard hardware used by most smartphone and smartwatch manufacturers. It measures the linear acceleration of the device in three-dimensional space, including the force of gravity. Data retrieved from the accelerometer can be processed to detect sudden changes in movement [80]. The USC-HAD project [79] utilizes a customized sensing platform called MotionNode (MN) to capture human activity signals. This multi-modal sensor integrates a 3-axis accelerometer, a 3-axis gyroscope, and a 3-axis magnetometer. Accelerometer-based datasets make up a significant portion of the listed datasets, such as UCI-HAR, USC-HAD, and WISDM, as shown in Table 3.
- **Gyroscopes**. The gyroscope is another standard piece of smartphone hardware that measures orientation and

angular velocity, operating based on the principles of angular momentum. Signals retrieved from the gyroscope can be processed to detect position and rotation angle, which are crucial for enhancing activities (fall detection) performance [80], [81].

- **Magnetometers**. Magnetometers detect the strength and direction of the Earth's magnetic field as a vector quantity [82].
- **Physiological Sensors**. Physiological sensors measure physical parameters related to the body's physiological state, such as heart rate, body temperature, blood pressure, and respiratory rate [83]. The most popular physiological sensors are electromyograms (EMG) and electrocardiograms (ECG), which are commonly employed in portable biomedical devices. EMG measures the electrical activity of muscles, while ECG measures the electrical activity of the heart [78].

B. AMBIENT SENSOR-BASED DATASETS

Ambient sensors are installed in the environment at fixed locations to recognize activities. These sensors primarily include Wi-Fi, radar, magnetic contact sensors, motion sensors, ultrasound, and visible light.

- **Wi-Fi**. Wi-Fi is a local-area wireless communication technology that transmits signals from a transmitter to a receiver, such as a smartphone or a Wi-Fi access point [78], [84]. The pattern of Wi-Fi signal transmission can be used to extract and analyze the location or orientation of activities. For example, Chen et al. [53] design a CL framework to recognize human hand gestures using Wi-Fi CSI data. DualConFi [52] introduces a new dual-stream CL model designed to process and learn from raw CSI data in a self-supervised manner.
- **Radar**. Radar technology can extract the context of activities through range (the physical distance to the radar), time (the evolution of the subject's location), and velocity (body motion) [85]. The SSCL framework [86] introduces a self-supervised CL approach that leverages physics-aware augmented radar micro-Doppler signatures for HAR, such as vacuuming the floor and walking with a cane.
- **Magnetic contact sensors**. When two magnetic sensors are separated, an alarm is triggered. These sensors are commonly used as door/window sensors to monitor the opening and closing of windows, front doors, cabinet doors, refrigerator doors, and more. Additionally, these sensors can be utilized for indoor positioning and contact sensing, contributing to activity recognition in smart homes [87].
- **Motion sensors**. Infrared motion sensors can be installed in indoor environments to recognize daily activities or distinguish proximity during room transitions. For example, the AttCLHAR framework [27] leverages the SimCLR framework and a self-attention mechanism to recognize daily activities, including eating, sleeping, and meal preparation, using the CASAS Aruba and Milan datasets [88].
- **Ultrasound**. EI [84] conducts an experiment to evaluate the effect of human activities on ultrasound signals. The sound signal is transmitted from the generator to the receiver, comprising both sound waves traveling directly through the Line-of-Sight (LOS) and those reflected by surrounding objects, including human bodies in the room.
- **Visible light**. The optical system can employ photoresistors to capture in-air body gestures, as these components are capable of detecting changes in illuminance (lux) resulting from body interactions [84].

C. ACTIVITY CATEGORIES

According to the taxonomy of activities presented in [89] and [78], Table 3 categorizes human activities into different types based on their application domains rather than by complexity. These categories include locomotion (e.g., walking, sitting), entertainment (e.g., watching TV, playing basketball), health-related activities (e.g., falls, frontal elevation of arms), daily activities (e.g., eating soup, brushing teeth), gestures (e.g., push, clap), and activity transitions (e.g., stand-to-sit, sit-to-stand).

Wearable and ambient sensors exhibit different strengths in recognizing actions and activities. Wearable sensors, attached to the body, are particularly effective at detecting fine-grained actions (e.g., pushing) and physical activities (e.g., walking, running) [49]. However, they face challenges in recognizing non-physical activities (e.g., watching TV) and complex activities (e.g., cooking), which require environmental context or additional sensors on multiple body parts. In contrast, ambient sensors, positioned in the environment, are more suitable for recognizing high-level activities involving environmental interactions (e.g., using an oven or refrigerator, motions in the kitchen) [3] to infer cooking activities, but are less capable of capturing specific bodily actions (e.g., frontal elevation of arms). Therefore, selecting appropriate sensor types based on the specific action or activity level is crucial. Combining wearable and ambient sensors, along with carefully determining the required number of sensors, can improve the accuracy of recognizing human activities.

VI. APPLICATION PERFORMANCE COMPARISONS

The performance of contrastive SSL for HAR has reached and even surpassed that of supervised learning methods. Different contrastive SSL approaches improve activity recognition from different perspectives. To identify the most effective models, this section compares recent contrastive SSL models' performances across widely used public datasets, including UCI-HAR, MobiAct, Motionsense, USC-HAD, and PAMAP2, for linear evaluation, semi-supervised learning, and transfer learning. Models were selected based on their evaluation of at least two of these datasets, ensuring a feasible

TABLE 3. Summary of commonly used datasets for human activity recognition. A: Accelerometer; AT: Activity Transitions; AS: Ambient Sensors; C: CSI; DA: Daily Activity; E: ECG; EMG: Electromyograms; EN: Entertainment; FMCWR: Frequency Modulated Continuous Wave Radar; G: Gyroscope; GE: Gesture; HR: Heart Rate Monitor; HRA: Health-Related Activity; LA: Linear Acceleration Sensor; LM: Locomotion; M: Magnetometers; MN: MotionNode; OS: Object Sensor; Sub.: Subject; SP: Smartphone; SW: Smartwatch; W: Wi-Fi; WD: Wearable Device.

Dataset	Device	Modality	Sub.	Activities (Category/Number of Activities)	Reference
UCI-HAR [80]	SP	A, G	30	Walking, walking upstairs, downstairs, sitting, laying, standing (LM/6).	[7], [14], [15], [17]–[19], [24], [26], [28], [30]–[32], [34], [37], [39], [43], [46], [48]–[50], [55], [75], [90]
USC-HAD [79]	MN	A, G, M	14	Walking forward, standing, sleeping, elevator up, and so on (LM, DA/12).	[14], [15], [18], [19], [26], [33], [39], [41], [47]–[49], [51], [74]
WISDM [91]	SP, SW	A, G	51	Walking, kicking soccer ball, brushing teeth, eating soup, and so on (LM, EN, DA/18).	[17], [24], [28], [33], [34], [37], [49], [55]
SHAR [92]	SP	A	30	StandingUpFL, LyingDownFS, and so on (LM/9); FallingBackSC, and so on (HRA/8).	[29], [90]
UCIHHAR [93]	SP, SW	A, G	9	Biking, sitting, standing, and so on (LM/6).	[17], [24], [25], [40], [48], [55], [75], [90]
Opportunity [94]	WD, OS, AS	A	4	Standing, walking, sitting, lying (LM/4).	[7], [42], [49]
PAMAP2 [95]	IMU, HR	A, G, M	9	Sitting, watching TV, house cleaning, and so on (LM, EN, DA/18).	[7], [32], [33], [37], [40], [42], [43], [46], [49], [51], [74]
RealWorld [96]	SP, SW	A	15	Jumping, standing, and so on (LM/8).	[42]
MobiAct [97]	SP	A, G	66	Standing, walking, and so on (LM/ 12); Front-knees-lying, and so on (HRA/4).	[15], [17]–[19], [24]–[26], [35], [39], [40], [43]
Motionsense [98]	SP	A, G	24	Walking, jogging, going up and down the stairs, sitting and standing (LM/6).	[14]–[17], [24], [25], [39], [40], [48]
DLR [99]	IMU	A	16	Walking, running, and so on (LM/7).	[35]
HAPT [100]	SP	A, G	30	Walking, standing, and so on (LM/6); stand-to-sit, sit-to-stand, and so on (AT/6).	[25], [34], [40]
UniMiB [92]	SP, SW, WD	A	30	StandingUpFS, fallingRight, fallingBack, and so on (HRA/17).	[33], [37], [51], [75]
DSADS [101]	Xsens MTx	A, G, M	8	Sitting, standing, cycling exercise bike, playing basketball (LM, EN/19).	[32], [82]
Shoaib [102]	SM	A, G, LA, M	10	Sitting, jogging, biking, and so on (LM, EN/7).	[48]
HASC [103]	SM	A	540	Stay, walk, jogging, skip, stair-up and stair-down (DA/6).	[41]
Hospital [104]	Inertial Sensor	A, G	12	Lying, stand up (sit to stand), sitting, walking, and so on. (LM/7).	[38]
mHealth [105]	WD	A, E, G, M	10	Waist bends forward, Frontal elevation of arms, Knees bending (crouching) (HRA/12).	[40], [43], [50]
Myogym [106]	Armband	A, EMG, G	10	Pushups, Bench Dip, and so on (EN/30).	[40]
HAR [107]	SM	A, G, GPS, M	19	Inactive, Active, Walking and Driving (DA/4).	[76]
Aruba/Milan [88]	AS	Motion	1	Relax, eating, sleeping, and so on (DA/11).	[3], [27]
EI [84]	W	C	40	Moving a suitcase, walking, and so on (DA/7).	[17], [24]
	iPad	Ultrasound	40	Moving a suitcase, walking, and so on (DA/7).	
	X60 [108]	mmWave	19	Moving a suitcase, walking, and so on (DA/7).	
	Lamp	Visible Light	18	Drawing an anticlockwise circle, and so on (DA/4).	
RSHA [85]	FMCWR	Radar	50	Walk, drink, fall, and so on (LM, HRA/6).	[53]
Widar3.0 [109]	W	C	16	Push, sweep, clap, slide, and so on (GE/6).	[53]
Office Room [110]	W	C	6	Lie down, walk, run, and so on (LM/6).	[36], [52]
SSCL [86]	TI AWR1642	Doppler Radar	10	Vacuuming the floor, walking with a cane, and so on (DA/14).	[86]
Falldefi [111]	W	C	3	Trip, slip, walk, and so on (DA/15).	[51]

basis for performance comparison. Due to space limitations, not all reviewed methods are included in the comparisons.

A. LINEAR EVALUATION

Linear evaluation assesses the quality of features learned through contrastive SSL in the latent space. Typically, models are pre-trained on a large unlabeled dataset. After pre-training, the encoder weights are frozen, and the projection head is dropped. A linear classifier is then attached and fine-tuned on top of the encoder's output with limited annotation. However, as the models in prominent papers utilize varying evaluation criteria and apply different proportions of annotations from the same datasets, we can only make a rough comparison of their performance metrics, as shown in Table 4.

ClusterCLHAR [14] achieves the best performance on the UCI-HAR, Motionsense, and USC-HAD datasets with mean F1 scores of 94.68%, 94.02%, and 85.01%, respectively, using only 10% labeled data. For the MobiAct dataset, SimCLRHAR [15] reaches a mean F1 score of 94.08% with 10% labeled data. Although STF-CSL [24] attains a higher F1 score of 95% on MobiAct, it required 50% labeled data. On the PAMAP2 dataset, COCOA [7] achieves an average macro F1 score of 96.3% with only 10% labeled

TABLE 4. Performance comparison in linear evaluation scenarios. Acc: accuracy; AF1: Average F1-scores; AMF1: Average Macro F1; MF1: Mean F1; WF1: Weighted F1. The best performance is highlighted in bold.

Model	UCI-HAR (%)	MobiAct (%)	Motionsense (%)	USC-HAD (%)	PAMAP2 (%)
COCOA [7]	90.9 (AMF1, 10%)	-	-	-	**96.3** (AMF1, 10%)
Enhanced CPC [40]	-	78.07 (F1)	89.35 (F1)	-	58.19 (F1)
STF-CSL [24]	92.5 (F1, 20%)	95 (F1, 50%)	95(F1, 50%)	-	-
CSSHAR [19]	91.14 (MF1)	81.13 (MF1)	-	57.76 (MF1)	-
CSSHAR-TFA [26]	90.7 (AF1)	83.23 (AF1)	-	72.74 (AF1)	-
SimCLRHAR [15]	80.21 (MF1, 10%)	**94.08** (MF1, 10%)	85.84 (MF1, 10%)	84.84 (MF1, 10%)	-
MoCoHAR [15]	91.89 (MF1, 10%)	83.38 (MF1, 10%)	91.55 (MF1, 10%)	78.51 (MF1, 10%)	-
ClusterCLHAR [14]	**94.68** (MF1, 10%)	-	**94.02** (MF1, 10%)	**85.01** (MF1, 10%)	-
DT [18]	93 (Macro F1, 20%)	82.02 (Macro F1, 20%)	-	55.78 (Macro F1, 20%)	-

TABLE 5. Performance comparison in semi-supervised learning scenarios. Acc: Accuracy; AF1: Average F1-scores; AMF1: Average Macro F1; MF1: Mean F1; WAF1: Weight-Average F1. The best performance is highlighted in bold.

Model	UCI-HAR (%)	MobiAct (%)	Motionsense (%)	USC-HAD (%)	PAMAP2 (%)
FusionCL [33]	-	-	-	**89.68** (Acc, 20%)	92.54 (Acc, 20%)
Li et al. [51]	-	-	-	93.84 (F1, 80%)	97.48 (F1, 70%)
SS-HAR [49]	**96.74** (Acc, 30%)	-	-	88.99 (Acc, 30%)	96.05 (Acc)
AutoCL [32]	94.69 (Acc, 20%)	-	-	-	90.47 (Acc, 20%)
COCOA [7]	95.8 (AMF1, 10%)	-	-	-	**98.5** (AMF1, 10%)
CoS [37]	95 (Acc, 20%)	-	-	-	90 (Acc, 20%)
CAGE [43]	92.49 (WF1, 10%)	89.84 (WF1, 14%)	-	-	83.57 (WF1, 7.5%)
CSSHAR-TFA [26]	90 (AF1, 60/class)	70.2 (AF1, 60/class)	-	56 (AF1, 60/class)	-
DDLearn [74]	-	-	-	77.36 (Acc, 20%)	82.95 (Acc, 20%)
SimCLRHAR [15]	95.26 (MF1, 10%)	95.38 (MF1, 10%)	**97.05** (MF1, 10%)	87.58 (MF1, 10%)	-
MoCoHAR [15]	95.49 (MF1, 10%)	**95.45** (MF1, 10%)	96.32 (MF1, 10%)	85.84 (MF1, 10%)	-
ClusterCLHAR [14]	**95.91** (MF1, 10%)	-	96.44 (MF1, 10%)	**87.86** (MF1, 10%)	-
SS-HAR [49]	95.21 (Acc, 10%)	-	-	85.86 (Acc, 10%)	-
DT [18]	90 (AF1, 30/class)	70 (AF1, 100/class)	-	54 (AF1, 100/class)	-
SemiC-HAR [17]	92.64 (WAF1, 10%)	94.79 (WAF1, 10%)	93.93 (WAF1, 10%)	-	-
CPC [39]	84.5 (MF1, 60/class)	62.5 (MF1, 60/class)	85 (MF1, 60/class)	44 (MF1, 60/class)	-
Enhanced CPC [40]	-	70 (F1, 80/class)	85 (F1, 80/class)	-	54 (F1, 80/class)
CSSHAR [19]	80 (MF1, 60/class)	62.5 (MF1, 60/class)	-	39.5 (MF1, 60/class)	-

data using all the available modalities of the PAMAP2 dataset. ClusterCLHAR demonstrates the best comparative performance across multiple models due to its strong results on three different datasets.

B. SEMI-SUPERVISED LEARNING

The semi-supervised learning scenario involves pre-training on a large unlabeled dataset and fine-tuning on a small amount of labeled data. Similar to linear evaluation, the selected models employ varying evaluation criteria and different proportions of annotations, as shown in Table 5.

For the UCI-HAR dataset, ClusterCLHAR reaches the best performance with a mean F1 score of 95.91% using only 10% labeled data, closely followed by COCOA with an average macro F1 score of 95.8%. In contrast, the BYOL-based method SS-HAR achieves a higher accuracy of 96.74% but requires 30% labeled data. On the MobiAct dataset, both SimCLRHAR and MoCoHAR perform impressively, achieving mean F1 scores of 95.38% and 95.45%, respectively, each with 10% labeled data. For the Motionsense dataset, SimCLRHAR leads with a mean F1 score of 97.05%, while MoCoHAR and ClusterCLHAR also show strong results with mean F1 scores of 96.32% and 96.44%, respectively. On the USC-HAD dataset, FusionCL reaches an accuracy of 89.68% with 20% labeled data, whereas ClusterCLHAR records a mean F1 score of 87.86% with just 10% labeled data.

For the PAMAP2 dataset, COCOA achieves an average macro F1 score of 98.5% with only 10% labeled data, surpassing Li et al. [51], who reach an F1 score of 97.48% with about 70% labeled data, demonstrating COCOA's superior ability to learn representations with less annotation and using multiple modalities. Generally, a higher proportion of labels significantly enhances performance. Overall, ClusterCLHAR, COCOA, MoCoHAR, and SimCLRHAR demonstrate competitive performance across various datasets.

C. TRANSFER LEARNING

Transfer learning involves transferring knowledge from one dataset to another. Encoders are initially pre-trained on one unlabeled dataset and subsequently evaluated on a different, unseen dataset. Table 6 presents a comparison of transfer learning performance across various models.

FusionCL pre-trains the SSL encoder on the PAMAP2 dataset and then fine-tunes it on the USC-HAD dataset with 20% labeled data, achieving an accuracy of 89.68%. This performance surpasses that of CSSHAR and DT models pre-trained on the MobiAct dataset. DT achieves the best performance when transferring from MobiAct to UCI-HAR,

TABLE 6. Performance comparison in transfer learning scenarios. Acc: Accuracy; M: MobiAct; MF1: Mean F1; O: OPPORTUNITY; P: PAMAP2. The best performance is highlighted in bold.

Model	Dataset	Performance (%)
FusionCL [33]	P→ USC-HAD	**89.68** (Acc, 20%)
CSSHAR [19]	M → USC-HAD	48.73 (MF1)
	M → UCI-HAR	88.26 (MF1)
DT [18]	M → USC-HAD	49.90 (F1)
	M → UCI-HAR	**90.35** (F1)
ModCL [45]	P → UCI-HAR	79.81 (Acc, 20%)
	P → UniMiB-SHAR	**77.17** (Acc, 20%)
	P→ WISDM	**91.96** (Acc, 20%)
	O → UCI-HAR	76.75 (Acc, 20%)
	O → UniMiB-SHAR	76.85 (Acc, 20%)
	O → WISDM	90.44 (Acc, 20%)

with an F1 score of 90.35%. ModCL is initially pre-trained on both the PAMAP2 and OPPORTUNITY datasets. It is then fine-tuned on the UniMiB-SHAR, WISDM, and UCI-HAR datasets, using 20% labeled data for training and 80% for testing. The transfer learning performance of ModCL from PAMAP2 to UniMiB-SHAR and WISDM reaches an accuracy of 77.17% and 91.96%, respectively. These results demonstrate that ModCL, FusionCL, and DT can effectively learn feature representations on unlabeled datasets and transfer this knowledge to new, unseen data. They also indicate that the choice of the pre-training dataset significantly impacts the downstream performance in fine-tuned scenarios.

In summary, contrastive SSL can learn and transfer general, informative representations from one or multiple sensors for HAR, achieving competitive performance with minimal or no labeled data. This significantly reduces the need for extensive annotation in real-world environments. Such performance enables a wide range of healthcare [16] applications, including monitoring the daily activities of older adults [43], detecting abnormalities like falls [41], [112], early detection of health condition changes such as cognitive decline [83], [113], exercise supervision, and rehabilitation monitoring [52]. These advancements hold great potential for improving care and support in healthcare settings.

VII. DISCUSSION AND FUTURE DIRECTIONS

In this section, we identify the limitations of the reviewed contrastive SSL models and propose several research directions for future exploration. The subsections first discuss model architecture, then explore data augmentation techniques and the selection of positive and negative samples within the frameworks, and finally address the challenges of handling multimodal sensor data and the highly sensitive nature of personal data.

A. INFLUENCE OF MODEL ARCHITECTURE ON THE PERFORMANCE OF CONTRASTIVE LEARNING

As discussed in Section III, various models may share similar encoders, optimizers, classifiers, and loss functions. However, architectural differences significantly impact their performance, often due to variations in associated loss functions. For example, although MoCoHAR [15] and SimCLRHAR [15] both utilize the DeepConvLSTM encoder, an MLP classifier and the Adam optimizer, they employ different architectures and loss functions, resulting in distinct performances in linear evaluation, as shown in Table 4. Properly selecting the architecture and loss function can significantly enhance model performance, indicating substantial opportunities for further exploration and contribution in this field. A comprehensive comparison of well-known contrastive SSL frameworks, such as SimCLR, MoCo, CPC, BYOL, and SimSiam, across multiple wearable and ambient sensor datasets for HAR, using either the same or different encoders, optimizers, and classifiers, could provide deeper insights into their performance on sensor-based data.

B. SELECTION OF DATA AUGMENTATION FUNCTIONS FOR PRE-TRAINING

Data augmentation is crucial in CL through pairwise comparisons. Different augmentation methods and their combinations can achieve the best performance across various sensor modalities, such as accelerometers and Wi-Fi. TS-TCC [30] employs temporal and contextual contrasting for time-series data, combining time-series-specific weak augmentations (e.g., scaling, time shifting) with strong augmentations (e.g., permutation, jitter + permutation) to robustly learn temporal relations for optimal performance in the temporal contrasting module. Manually selecting suitable augmentations could be time-consuming. AutoAL [32] conducts end-to-end auto-augmentation to identify effective augmentation strategies without manual effort. However, data augmentation methods risk distorting the distribution of raw data and causing DAB. CoS [37] highlights a critical issue: distortions induced by augmentation can be further magnified by the intermediate layers of a network, potentially severely harming the semantic structure of the original activity instance.

Therefore, selecting suitable augmentations for specific sensor modalities, while minimizing manual labor and considering potential distortions to the invariant properties of raw data, remains an area requiring further exploration.

C. SELECTION OF POSITIVE AND NEGATIVE SAMPLES

Intuitively, positive pairs are typically formed from two views of the same sample created through various data augmentations, while negative pairs are formed from different samples. The goal of this selection is to maintain invariance while being discriminative against negative samples.

Instead of generating augmented or temporal positive samples, COCOA [7] utilizes synchronous data segments across various sensors and modalities, leveraging the unique perspectives of each modality to create positive samples. It also observes that larger batch sizes increase the risk of false negative pairs from the same class being misidentified as different, degrading the quality of learned representations.

Positive pairs are defined by timely aligned readings across all sensors (inter-modal), while negative pairs come from temporally distant samples within the same sensor (intra-modal). In a multi-device setting, ColloSSL [42] introduces a two-step method for selecting positive and negative samples: device Selection, which employs Maximum Mean Discrepancy (MMD) to choose devices dynamically, and contrastive Sampling, which utilizes time-synchronized samples from these selected devices. AFVF [46] considers partially overlapping windows as hard negatives (i.e., points that are difficult to distinguish from an anchor point) rather than false negatives, as [114] argued that hard negatives are beneficial to CL.

ClusterCLHAR [14] highlights that instance discrimination could lead to over-clustering, where representations of instances from the same class become excessively separated. Consequently, sample representations with the same cluster labels are no longer considered as negative pairs in contrastive loss computation. Despite not using negative pairs, SS-HAR [49], based on the BYOL framework, still achieves competitive performance.

The selection or construction of positive and negative pairs for calculating contrastive loss remains a promising avenue for research.

D. MULTIMODAL CONTRASTIVE SELF-SUPERVISED LEARNING

Multimodal SSL requires integrating different types of sensor data and devices to learn representations based on correlations between various modalities.

Most recent HAR systems that utilize multimodal sensors primarily employ data-level (early) fusion, which aggregates signals from different sensors into a single multichannel input. However, this approach fails to account for the fact that different sensors capture distinct physical properties and produce unique patterns. To address this, CAGE [43] proposes a two-stream encoder that processes accelerometer and gyroscope signals separately, fusing modality-specific features at the feature level (late) for recognition. AFVF [46] proves that early fusion consumes fewer resources during training compared to late fusion, while late fusion often performs better as it involves a dedicated feature extractor for each sensor. In contrast, late fusion models may fail to consider correlations across modalities adequately. COCOA [7] enhances feature extraction using modality-specific encoders and a customized loss function to align latent embeddings across different modalities. CroSSL [44] aggregates the modality-specific embeddings and addresses the challenge of missing modalities with a masking-based technique.

Thus, the trade-off between early fusion and late fusion models, particularly their robustness in handling missing modalities and the selection of the proper number of sensors for multimodal CL in real-world environments, remains an open problem for further exploration.

E. PRIVACY PROTECTION

Sensor-based HAR data is highly sensitive regarding privacy, making it challenging for researchers and practitioners to access. ColloSSL [42] addresses privacy issues associated with user-owned sensor devices, often due to manufacturers' reluctance to offload raw sensor data to a centralized cloud server. It proposes two potential solutions: first, training models on a trusted edge device, such as a home router, to ensure that user data remains on-premises. Secondly, exploring federated SSL [115] approaches that train the feature extractor locally on each device, while only aggregating the gradients on a central server. Considering data privacy while performing contrastive SSL for HAR in real-world environments is a promising research direction. Cosmo [47] mentions future work to incorporate the federated learning paradigm to further improve the performance of Cosmo while protecting user privacy. Each node in federated learning would run the first stage of Cosmo using unlabeled multimodal data.

VIII. CONCLUSION

This article provides an extensive review of recent contrastive SSL approaches for sensor-based HAR. We thoroughly explain contrastive SSL, covering common network architectures and detailing the main network components. Additionally, we present a comprehensive table of popular benchmarks for HAR and conduct an empirical comparison of recent contrastive SSL models. Some significant insights illuminate the selection of contrastive SSL for sensor-based HAR. Finally, we identify and discuss open problems in current methodologies and outline future research directions for HAR in real-world applications.

REFERENCES

[1] V. Bianchi, M. Bassoli, G. Lombardo, P. Fornacciari, M. Mordonini, and I. De Munari, "IoT wearable sensor and deep learning: An integrated approach for personalized human activity recognition in a smart home environment," *IEEE Internet Things J.*, vol. 6, no. 5, pp. 8553–8562, Oct. 2019.

[2] N. Phukan, S. Mohine, A. Mondal, M. S. Manikandan, and R. B. Pachori, "Convolutional neural network-based human activity recognition for edge fitness and context-aware health monitoring devices," *IEEE Sensors J.*, vol. 22, no. 22, pp. 21816–21826, Nov. 2022.

[3] H. Chen, C. Gouin-Vallerand, K. Bouchard, S. Gaboury, M. Couture, N. Bier, and S. Giroux, "Leveraging self-supervised learning for human activity recognition with ambient sensors," in *Proc. ACM Conf. Inf. Technol. Social Good*, Sep. 2023, pp. 324–332.

[4] M. Schiemer, L. Fang, S. Dobson, and J. Ye, "Online continual learning for human activity recognition," *Pervas. Mobile Comput.*, vol. 93, Jun. 2023, Art. no. 101817.

[5] V. Lafontaine, K. Bouchard, J. Maítre, and S. Gaboury, "Denoising UWB radar data for human activity recognition using convolutional autoencoders," *IEEE Access*, vol. 11, pp. 81298–81309, 2023.

[6] J. Wang, Y. Chen, S. Hao, X. Peng, and L. Hu, "Deep learning for sensor-based activity recognition: A survey," *Pattern Recognit. Lett.*, vol. 119, pp. 3–11, Mar. 2019.

[7] S. Deldari, H. Xue, A. Saeed, D. V. Smith, and F. D. Salim, "COCOA: Cross modality contrastive learning for sensor data," *Proc. ACM Interact., Mobile, Wearable Ubiquitous Technol.*, vol. 6, no. 3, pp. 1–28, Sep. 2022.

[8] P. H. Le-Khac, G. Healy, and A. F. Smeaton, "Contrastive representation learning: A framework and review," *IEEE Access*, vol. 8, pp. 193907–193934, 2020.

[9] A. Mohamed, H.-Y. Lee, L. Borgholt, J. D. Havtorn, J. Edin, C. Igel, K. Kirchhoff, S.-W. Li, K. Livescu, L. Maaløe, T. N. Sainath, and S. Watanabe, "Self-supervised speech representation learning: A review," *IEEE J. Sel. Topics Signal Process.*, vol. 16, no. 6, pp. 1179–1210, Oct. 2022.

[10] A. van den Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," 2018, *arXiv:1807.03748*.

[11] T. Chen, S. Kornblith, M. Norouzi, and G. E. Hinton, "A simple framework for contrastive learning of visual representations," in *Proc. 37th Int. Conf. Mach. Learn.*, vol. 119, 2020, pp. 1597–1607.

[12] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 9726–9735.

[13] J.-B. Grill, "Bootstrap your own latent a new approach to self-supervised learning," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 21271–21284.

[14] J. Wang, T. Zhu, L. Chen, H. Ning, and Y. Wan, "Negative selection by clustering for contrastive learning in human activity recognition," *IEEE Internet Things J.*, vol. 10, no. 12, pp. 10833–10844, Oct. 2023.

[15] J. Wang, T. Zhu, J. Gan, L. L. Chen, H. Ning, and Y. Wan, "Sensor data augmentation by resampling in contrastive learning for human activity recognition," *IEEE Sensors J.*, vol. 22, no. 23, pp. 22994–23008, Dec. 2022.

[16] C. Ian Tang, I. Perez-Pozuelo, D. Spathis, and C. Mascolo, "Exploring contrastive learning in human activity recognition for healthcare," 2020, *arXiv:2011.11542*.

[17] D. Liu and T. Abdelzaher, "Semi-supervised contrastive learning for human activity recognition," in *Proc. 17th Int. Conf. Distrib. Comput. Sensor Syst. (DCOSS)*, Jul. 2021, pp. 45–53.

[18] B. Khaertdinov, S. Asteriadis, and E. Ghaleb, "Dynamic temperature scaling in contrastive self-supervised learning for sensor-based human activity recognition," *IEEE Trans. Biometrics, Behav., Identity Sci.*, vol. 4, no. 4, pp. 498–507, Oct. 2022.

[19] B. Khaertdinov, E. Ghaleb, and S. Asteriadis, "Contrastive self-supervised learning for sensor-based human activity recognition," in *Proc. IEEE Int. Joint Conf. Biometrics (IJCB)*, Aug. 2021, pp. 1–8.

[20] S. K. Yadav, K. Tiwari, H. M. Pandey, and S. A. Akbar, "A review of multimodal human activity recognition with special emphasis on classification, applications, challenges and future directions," *Knowl.-Based Syst.*, vol. 223, Jul. 2021, Art. no. 106970.

[21] M. J. Page et al., "The PRISMA 2020 statement: An updated guideline for reporting systematic reviews," *BMJ*, vol. 372, no. n71, pp. 178–189, 2021. [Online]. Available: https://www.bmj.com/content/372/bmj.n71

[22] A. Jaiswal, A. R. Babu, M. Z. Zadeh, D. Banerjee, and F. Makedon, "A survey on contrastive self-supervised learning," *Technologies*, vol. 9, no. 1, p. 2, Dec. 2020.

[23] X. Chen and K. He, "Exploring simple Siamese representation learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 15745–15753.

[24] D. Liu, T. Wang, S. Liu, R. Wang, S. Yao, and T. Abdelzaher, "Contrastive self-supervised representation learning for sensing signals from the time-frequency perspective," in *Proc. Int. Conf. Comput. Commun. Netw. (ICCCN)*, Jul. 2021, pp. 1–10.

[25] S. R. Taghanaki, M. Rainbow, and A. Etemad, "Self-supervised human activity recognition with localized time-frequency contrastive representation learning," *IEEE Trans. Human-Mach. Syst.*, vol. 53, no. 6, pp. 1027–1037, Dec. 2023.

[26] B. Khaertdinov and S. Asteriadis, "Temporal feature alignment in contrastive self-supervised learning for human activity recognition," in *Proc. IEEE Int. Joint Conf. Biometrics (IJCB)*, Oct. 2022, pp. 1–9.

[27] H. Chen, C. Gouin-Vallerand, K. Bouchard, S. Gaboury, M. Couture, N. Bier, and S. Giroux, "Enhancing human activity recognition in smart homes with self-supervised learning and self-attention," *Sensors*, vol. 24, no. 3, p. 884, Jan. 2024.

[28] E. Eldele, M. Ragab, Z. Chen, M. Wu, C.-K. Kwoh, and X. Li, "Contrastive domain adaptation for time-series via temporal mixup," *IEEE Trans. Artif. Intell.*, vol. 5, no. 3, pp. 1185–1194, Jun. 2024.

[29] A. Z. M. Faridee, A. Chakma, Z. Hasan, N. Roy, and A. Misra, "CoDEm: Conditional domain embeddings for scalable human activity recognition," in *Proc. IEEE Int. Conf. Smart Comput. (SMARTCOMP)*, Jun. 2022, pp. 9–18.

[30] E. Eldele, M. Ragab, Z. Chen, M. Wu, C.-K. Kwoh, X. Li, and C. Guan, "Self-supervised contrastive representation learning for semi-supervised time-series classification," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 12, pp. 15604–15618, Dec. 2023.

[31] Y. Zheng, Y. Luo, H. Shao, L. Zhang, and L. Li, "DABaCLT: A data augmentation bias-aware contrastive learning framework for time series representation," *Appl. Sci.*, vol. 13, no. 13, p. 7908, Jul. 2023.

[32] Q. Wu, J. Shen, F. Fan, Y. Gu, C. Xu, and Y. Chen, "Auto-augmentation contrastive learning for wearable-based human activity recognition," in *Proc. IEEE Int. Conf. Bioinf. Biomed. (BIBM)*, vol. 33, Dec. 2023, pp. 1537–1544.

[33] C. Guo, Y. Zhang, Y. Chen, W. Yang, Q. Wu, and Z. Wang, "Time-frequency fusion-augmentation based contrastive learning for self-supervised human activity recognition," in *Proc. IEEE Smart World Congr. (SWC)*, vol. 189, Aug. 2023, pp. 1–8.

[34] Z. Xiao, H. Tong, R. Qu, H. Xing, S. Luo, Z. Zhu, F. Song, and L. Feng, "CapMatch: Semi-supervised contrastive transformer capsule with feature-based knowledge distillation for human activity recognition," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Dec. 27, 2023, doi: 10.1109/TNNLS.2023.3344294.

[35] I. Kim, J. Lim, and J. Lee, "Human activity recognition via temporal fusion contrastive learning," *IEEE Access*, vol. 12, pp. 20854–20866, 2024.

[36] J. Yang, X. Chen, H. Zou, D. Wang, and L. Xie, "AutoFi: Toward automatic Wi-Fi human sensing via geometric self-supervised learning," *IEEE Internet Things J.*, vol. 10, no. 8, pp. 7416–7425, Apr. 2023.

[37] D. Cheng, L. Zhang, C. Bu, H. Wu, and A. Song, "Learning hierarchical time series data augmentation invariances via contrastive supervision for human activity recognition," *Knowl.-Based Syst.*, vol. 276, Sep. 2023, Art. no. 110789.

[38] S. Xia, L. Chu, L. Pei, W. Yu, and R. C. Qiu, "Multi-level contrast network for wearables-based joint activity segmentation and recognition," in *Proc. GLOBECOM IEEE Global Commun. Conf.*, Dec. 2022, pp. 566–572.

[39] H. Haresamudram, I. Essa, and T. Plötz, "Contrastive predictive coding for human activity recognition," *Proc. ACM Interact., Mobile, Wearable Ubiquitous Technol.*, vol. 5, no. 2, pp. 1–26, Jun. 2021.

[40] H. Haresamudram, I. Essa, and T. Plötz, "Investigating enhancements to contrastive predictive coding for human activity recognition," in *Proc. IEEE Int. Conf. Pervasive Comput. Commun. (PerCom)*, Mar. 2023, pp. 232–241.

[41] S. Deldari, D. V. Smith, H. Xue, and F. D. Salim, "Time series change point detection with self-supervised contrastive predictive coding," in *Proc. Web Conf.*, New York, NY, USA, Jun. 2021, pp. 3124–3135.

[42] Y. Jain, C. I. Tang, C. Min, F. Kawsar, and A. Mathur, "ColloSSL: Collaborative self-supervised learning for human activity recognition," *Proc. ACM Interact., Mobile, Wearable Ubiquitous Technol.*, vol. 6, no. 1, pp. 1–28, Mar. 2022.

[43] I. Koo, Y. Park, M. Jeong, and C. Kim, "Contrastive accelerometer–gyroscope embedding model for human activity recognition," *IEEE Sensors J.*, vol. 23, no. 1, pp. 506–513, Jan. 2023.

[44] S. Deldari, D. Spathis, M. Malekzadeh, F. Kawsar, F. D. Salim, and A. Mathur, "CroSSL: Cross-modal self-supervised learning for time-series through latent masking," in *Proc. 17th ACM Int. Conf. Web Search Data Mining*, vol. 33, Mar. 2024, pp. 152–160.

[45] C. Guo, Y. Zhang, Y. Chen, C. Xu, and Z. Wang, "Modality consistency-guided contrastive learning for wearable-based human activity recognition," *IEEE Internet Things J.*, vol. 11, no. 12, pp. 21750–21762, Jun. 2024.

[46] D.-A. Nguyen, C. Pham, and N.-A. Le-Khac, "Virtual fusion with contrastive learning for single-sensor-based activity recognition," *IEEE Sensors J.*, vol. 24, no. 15, pp. 25041–25048, Aug. 2024.

[47] X. Ouyang, X. Shuai, J. Zhou, I. W. Shi, Z. Xie, G. Xing, and J. Huang, "Cosmo: Contrastive fusion learning with small data for multimodal human activity recognition," in *Proc. 28th Annu. Int. Conf. Mobile Comput. Netw.*, vol. 9, Oct. 2022, pp. 324–337.

[48] L. Xu, C. Gu, R. Tan, S. He, and J. Chen, "MESEN: Exploit multimodal data to design unimodal human activity recognition with few labels," in *Proc. 21st ACM Conf. Embedded Netw. Sensor Syst.*, vol. 35, Nov. 2023, pp. 1–14.

[49] Y. Zhou, C. Xie, S. Sun, X. Zhang, and Y. Wang, "A self-supervised human activity recognition approach via body sensor networks in smart city," *IEEE Sensors J.*, vol. 24, no. 5, pp. 5476–5485, Jun. 2024.

[50] Y. Sun, X. Xu, X. Tian, L. Zhou, and Y. Li, "Efficient human activity recognition: A deep convolutional transformer-based contrastive self-supervised approach using wearable sensors," *Eng. Appl. Artif. Intell.*, vol. 135, Sep. 2024, Art. no. 108705.

[51] Y. Li, J. Wu, A. Fang, and W. Li, "A simple optimization strategy via contrastive loss for recognizing human activity using wearable sensors," *IEEE Sensors J.*, vol. 23, no. 18, pp. 21588–21598, Sep. 2023.

[52] K. Xu, J. Wang, L. Zhang, H. Zhu, and D. Zheng, "Dual-stream contrastive learning for channel state information based human activity recognition," *IEEE J. Biomed. Health Informat.*, vol. 27, no. 1, pp. 329–338, Jan. 2023.

[53] B. Chen, J. Wang, Y. Lv, Q. Gao, M. Pan, and Y. Fang, "Device-free wireless sensing with few labels through mutual information maximization," *IEEE Internet Things J.*, vol. 11, no. 6, pp. 10513–10524, Mar. 2024.

[54] S. R. Ramamurthy, S. Chatterjee, E. Galik, A. Gangopadhyay, N. Roy, B. Mitra, and S. Chakraborty, "CogAx: Early assessment of cognitive and functional impairment from accelerometry," in *Proc. IEEE Int. Conf. Pervasive Comput. Commun. (PerCom)*, Mar. 2022, pp. 66–76.

[55] G. Wilson, J. R. Doppa, and D. J. Cook, "CALDA: Improving multi-source time series domain adaptation with contrastive adversarial learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 12, pp. 14208–14221, Dec. 2023.

[56] Z. Wu, Y. Xiong, S. X. Yu, and D. Lin, "Unsupervised feature learning via non-parametric instance discrimination," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3733–3742.

[57] X. Chen, H. Fan, R. Girshick, and K. He, "Improved baselines with momentum contrastive learning," 2020, *arXiv:2003.04297*.

[58] X. Chen, S. Xie, and K. He, "An empirical study of training self-supervised vision transformers," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 9620–9629.

[59] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[60] A. Vaswani, "Attention is all you need," in *Proc. NIPS*, 2017, pp. 6000–6010.

[61] I. Misra and L. van der Maaten, "Self-supervised learning of pretext-invariant representations," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 6706–6716.

[62] M. Ye, X. Zhang, P. C. Yuen, and S.-F. Chang, "Unsupervised embedding learning via invariant and spreading instance feature," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 6203–6212.

[63] T. Chen, S. Kornblith, K. Swersky, M. Norouzi, and G. E. Hinton, "Big self-supervised models are strong semi-supervised learners," in *Proc. Annu. Conf. Neural Inf. Process. Syst.*, 2020, pp. 22243–22255.

[64] K. Cho, B. van Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder–decoder for statistical machine translation," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2014, pp. 1724–1734.

[65] O. Henaff, "Data-efficient image recognition with contrastive predictive coding," in *Proc. ICML*, Jul. 2020, pp. 4182–4192.

[66] Y. Tian, D. Krishnan, and P. Isola, "Contrastive multiview coding," 2019, *arXiv:1906.05849*.

[67] Y. Tian, C. Sun, B. Poole, D. Krishnan, C. Schmid, and P. Isola, "What makes for good views for contrastive learning?" in *Proc. Neural Inf. Process. Syst. (NeurIPS)*, Dec. 2020, pp. 6827–6839.

[68] M. Caron, P. Bojanowski, A. Joplin, and M. Douze, "Deep clustering for unsupervised learning of visual features," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 139–156.

[69] M. Caron, I. Misra, J. Mairal, P. Goyal, P. Bojanowski, and A. Joulin, "Unsupervised learning of visual features by contrasting cluster assignments," in *Proc. NIPS*, Dec. 2020, pp. 9912–9924.

[70] J. Zbontar, L. Jing, I. Misra, Y. LeCun, and S. Deny, "Barlow twins: Self-supervised learning via redundancy reduction," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 12310–12320.

[71] R. D. Hjelm, A. Fedorov, S. Lavoie-Marchildon, K. Grewal, P. Bachman, A. Trischler, and Y. Bengio, "Learning deep representations by mutual information estimation and maximization," in *Proc. Int. Conf. Learn. Represent.*, New Orleans, LA, USA, May 2019, pp. 1–24.

[72] J. Yu, H. Yin, X. Xia, T. Chen, J. Li, and Z. Huang, "Self-supervised learning for recommender systems: A survey," *IEEE Trans. Knowl. Data Eng.*, vol. 36, no. 1, pp. 335–355, Jan. 2024.

[73] T. T. Um, F. M. J. Pfister, D. Pichler, S. Endo, M. Lang, S. Hirche, U. Fietzek, and D. Kulić, "Data augmentation of wearable sensor data for Parkinson's disease monitoring using convolutional neural networks," in *Proc. 19th ACM Int. Conf. Multimodal Interact.*, Nov. 2017, pp. 216–220.

[74] X. Qin, J. Wang, S. Ma, W. Lu, Y. Zhu, X. Xie, and Y. Chen, "Generalizable low-resource activity recognition with diverse and discriminative representation learning," in *Proc. 29th ACM SIGKDD Conf. Knowl. Discovery Data Mining*, vol. 13, Aug. 2023, pp. 1943–1953.

[75] H. Qian, T. Tian, and C. Miao, "What makes good contrastive learning on small-scale wearable-based tasks?" in *Proc. 28th ACM SIGKDD Conf. Knowl. Discovery Data Mining*, vol. 54, Aug. 2022, pp. 3761–3771.

[76] Y. Zou, Y. Zhang, and X. Zhao, "Self-supervised time series classification based on LSTM and contrastive transformer," *Wuhan Univ. J. Natural Sci.*, vol. 27, no. 6, pp. 521–530, Dec. 2022.

[77] M. Gutmann and A. Hyvarinen, "Noise-contrastive estimation: A new estimation principle for unnormalized statistical models," in *Proc. 13th Int. Conf. Artif. Intell. Statist.*, 2010, pp. 297–304.

[78] F. Gu, M.-H. Chung, M. Chignell, S. Valaee, B. Zhou, and X. Liu, "A survey on deep learning for human activity recognition," *ACM Comput. Surv.*, vol. 54, no. 8, pp. 1–34, Oct. 2021.

[79] M. Zhang and A. A. Sawchuk, "USC-HAD: A daily activity dataset for ubiquitous activity recognition using wearable sensors," in *Proc. ACM Conf. Ubiquitous Comput.*, Sep. 2012, pp. 1036–1043.

[80] D. Anguita, A. Ghio, L. Oneto, X. Parra, and J. L. ReyesOrtiz, "A public domain dataset for human activity recognition using smartphones," in *Proc. 21st Int. Eur. Symp. Artif. Neural Netw., Comput. Intell. Mach. Learn.*, 2013, pp. 437–442.

[81] Q. Li, J. A. Stankovic, M. A. Hanson, A. T. Barth, J. Lach, and G. Zhou, "Accurate, fast fall detection using gyroscopes and accelerometer-derived posture information," in *Proc. 6th Int. Workshop Wearable Implant. Body Sensor Netw.*, Jun. 2009, pp. 138–143.

[82] B. Barshan and M. C. Yüksek, "Recognizing daily and sports activities in two open source machine learning environments using body-worn sensor units," *Comput. J.*, vol. 57, no. 11, pp. 1649–1667, Nov. 2014.

[83] F. Demrozi, G. Pravadelli, A. Bihorac, and P. Rashidi, "Human activity recognition using inertial, physiological and environmental sensors: A comprehensive survey," *IEEE Access*, vol. 8, pp. 210816–210836, 2020.

[84] W. Jiang, C. Miao, F. Ma, S. Yao, Y. Wang, Y. Yuan, H. Xue, C. Song, X. Ma, D. Koutsonikolas, W. Xu, and L. Su, "Towards environment independent device free human activity recognition," in *Proc. 24th Annu. Int. Conf. Mobile Comput. Netw.*, Oct. 2018, pp. 289–304.

[85] F. Fioranelli, S. A. Shah, H. Li, A. Shrestha, S. Yang, and J. L. Kernec, "Radar sensing for healthcare," *Electron. Lett.*, vol. 55, no. 19, pp. 1022–1024, Aug. 2019.

[86] M. M. Rahman and S. Z. Gurbuz, "Self-supervised contrastive learning for radar-based human activity recognition," in *Proc. IEEE Radar Conf. (RadarConf)*, May 2023, pp. 1–6.

[87] S. Dernbach, B. Das, N. C. Krishnan, B. L. Thomas, and D. J. Cook, "Simple and complex activity recognition through smart phones," in *Proc. 8th Int. Conf. Intell. Environments*, Jun. 2012, pp. 214–221.

[88] D. J. Cook, A. S. Crandall, B. L. Thomas, and N. C. Krishnan, "CASAS: A smart home in a box," *Comput. J.*, vol. 46, no. 7, pp. 62–69, Jul. 2013.

[89] O. D. Incel, M. Kose, and C. Ersoy, "A review and taxonomy of activity recognition on mobile phones," *BioNanoScience*, vol. 3, no. 2, pp. 145–171, Jun. 2013.

[90] C. Xu, Y. Li, D. Lee, D. H. Park, H. Mao, H. Do, J. Chung, and D. Nair, "Augmentation robust self-supervised learning for human activity recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2023, pp. 1–5.

[91] G. Weiss, *WISDM Smartphone and Smartwatch Activity and Biometrics Dataset*. Irvine, CA, USA: UCI Machine Learning Repository, 2019, doi: 10.24432/C5HK59.

[92] D. Micucci, M. Mobilio, and P. Napoletano, "UniMiB SHAR: A dataset for human activity recognition using acceleration data from smartphones," *Appl. Sci.*, vol. 7, no. 10, p. 1101, Oct. 2017.

[93] A. Stisen, H. Blunck, S. Bhattacharya, T. S. Prentow, M. B. Kjærgaard, A. Dey, T. Sonne, and M. M. Jensen, "Smart devices are different: Assessing and MitigatingMobile sensing heterogeneities for activity recognition," in *Proc. 13th ACM Conf. Embedded Netw. Sensor Syst.*, Nov. 2015, pp. 127–140.

[94] D. Roggen, A. Calatroni, M. Rossi, T. Holleczek, K. Förster, G. Tröster, P. Lukowicz, D. Bannach, G. Pirkl, A. Ferscha, J. Doppler, C. Holzmann, M. Kurz, G. Holl, R. Chavarriaga, H. Sagha, H. Bayati, M. Creatura, and J. del R Millan, "Collecting complex activity datasets in highly rich networked sensor environments," in *Proc. 7th Int. Conf. Networked Sens. Syst. (INSS)*, Jun. 2010, pp. 233–240.

[95] A. Reiss and D. Stricker, "Introducing a new benchmarked dataset for activity monitoring," in *Proc. 16th Int. Symp. Wearable Comput.*, Jun. 2012, pp. 108–109.

[96] T. Sztyler and H. Stuckenschmidt, "On-body localization of wearable devices: An investigation of position-aware activity recognition," in *Proc. IEEE Int. Conf. Pervasive Comput. Commun. (PerCom)*, Mar. 2016, pp. 1–9.

[97] C. Chatzaki, M. Pediaditis, G. Vavoulas, and M. Tsiknakis, "Human daily activity and fall recognition using a smartphone acceleration sensor," in *Proc. Info. Commun. Technol. Ageing Well e-Health Conf. (ICT4AWE)*, Apr. 2017, pp. 100–118.

[98] M. Malekzadeh, R. G. Clegg, A. Cavallaro, and H. Haddadi, "Protecting sensory data against sensitive inferences," in *Proc. 1st Workshop Privacy Design Distrib. Syst.*, Apr. 2018, pp. 1–6.

[99] K. Frank, M. J. Vera Nadales, P. Robertson, and T. Pfeifer, "Bayesian recognition of motion related activities with inertial sensors," in *Proc. 12th ACM Int. Conf. Adjunct Papers Ubiquitous Comput. Adjunct*, Sep. 2010, pp. 445–446.

[100] J.-L. Reyes-Ortiz, L. Oneto, A. Sama, X. Parra, and D. Anguita, "Transition-aware human activity recognition using smartphones," *Neurocomputing*, vol. 171, pp. 754–767, Jan. 2016.

[101] B. Barshan and K. Altun, *Daily and Sports Activities*. Irvine, CA, USA: UCI Machine Learning Repository, 2013, doi: 10.24432/C5C59F.

[102] M. Shoaib, S. Bosch, O. Incel, H. Scholten, and P. Havinga, "Fusion of smartphone motion sensors for physical activity recognition," *Sensors*, vol. 14, no. 6, pp. 10146–10176, Jun. 2014.

[103] N. Kawaguchi, N. Ogawa, Y. Iwasaki, K. Kaji, T. Terada, K. Murao, S. Inoue, Y. Kawahara, Y. Sumi, and N. Nishio, "HASC challenge: Gathering large scale human activity corpus for the real-world activity understandings," in *Proc. 2nd Augmented Hum. Int. Conf.*, Mar. 2011, pp. 1–5.

[104] R. Yao, G. Lin, Q. Shi, and D. C. Ranasinghe, "Efficient dense labelling of human activity sequences from wearables using fully convolutional networks," *Pattern Recognit.*, vol. 78, pp. 252–266, Jun. 2018.

[105] O. Banos, "mHealthDroid: A novel framework for agile development of mobile health applications," in *Proc. Int. Workshop Ambient Assist. Living*. Cham, Switzerland: Springer, 2014, pp. 91–98.

[106] H. Koskimäki, P. Siirtola, and J. Röning, "MyoGym: Introducing an open gym data set for activity recognition collected using Myo armband," in *Proc. ACM Int. Joint Conf. Pervasive Ubiquitous Comput. ACM Int. Symp. Wearable Comput.*, Sep. 2017, pp. 537–546.

[107] D. Garcia-Gonzalez, D. Rivero, E. Fernandez-Blanco, and M. R. Luaces, "A public domain dataset for real-life human activity recognition using smartphone sensors," *Sensors*, vol. 20, no. 8, p. 2200, Apr. 2020.

[108] S. K. Saha, Y. Ghasempour, M. K. Haider, T. Siddiqui, P. De Melo, N. Somanchi, L. Zakrajsek, A. Singh, O. Torres, D. Uvaydov, J. M. Jornet, E. Knightly, D. Koutsonikolas, D. Pados, and Z. Sun, "x60: A programmable testbed for wideband 60 GHz WLANs with phased arrays," in *Proc. 11th Workshop Wireless Netw. Testbeds, Experim. Eval. CHaracterization*, Oct. 2017, pp. 75–82.

[109] Y. Zhang, Y. Zheng, K. Qian, G. Zhang, Y. Liu, C. Wu, and Z. Yang, "Widar3.0: Zero-effort cross-domain gesture recognition with Wi-Fi," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 11, pp. 8671–8688, Nov. 2022.

[110] S. Yousefi, H. Narui, S. Dayal, S. Ermon, and S. Valaee, "A survey on behavior recognition using WiFi channel state information," *IEEE Commun. Mag.*, vol. 55, no. 10, pp. 98–104, Oct. 2017.

[111] S. Palipana, D. Rojas, P. Agrawal, and D. Pesch, "FallDeFi: Ubiquitous fall detection using commodity Wi-Fi devices," *Proc. ACM Interact., Mobile, Wearable Ubiquitous Technol.*, vol. 1, no. 4, pp. 1–25, Jan. 2018.

[112] F. Hussain, F. Hussain, M. Ehatisham-ul-Haq, and M. A. Azam, "Activity-aware fall detection and recognition based on wearable sensors," *IEEE Sensors J.*, vol. 19, no. 12, pp. 4528–4536, Jun. 2019.

[113] Y. Wang, S. Cang, and H. Yu, "A survey on wearable sensor modality centred human activity recognition in health care," *Expert Syst. Appl.*, vol. 137, pp. 167–190, Dec. 2019.

[114] J. Robinson, C.-Y. Chuang, S. Sra, and S. Jegelka, "Contrastive learning with hard negative samples," 2020, *arXiv:2010.04592*.

[115] X. Ouyang, Z. Xie, J. Zhou, J. Huang, and G. Xing, "ClusterFL: A similarity-aware federated learning system for human activity recognition," in *Proc. 19th Annu. Int. Conf. Mobile Syst., Appl., Services*, Jun. 2021, pp. 54–66.

HUI CHEN received the B.S. degree in electrical and information engineering from Jiangsu University of Technology and the master's degree in communication and electronic engineering from East China Normal University. She is currently pursuing the Ph.D. degree in computer science with the Université de Sherbrooke, Canada. Her major research interests include self-supervised learning, activity recognition, telemonitoring, and smart homes.

CHARLES GOUIN-VALLERAND received the B.Sc. and M.Sc. degrees in computer science from the Université de Sherbrooke, Canada, in 2005 and 2008, respectively, and the joint Ph.D. degree in computer science from Sorbonne-UPMC, France, and the Université de Sherbrooke, in 2011.

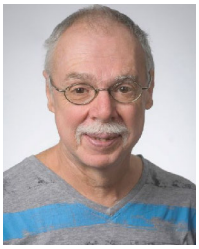
From 2011 to 2012, he was a Postdoctoral Fellow with the Human-Computer Interaction Institute of Carnegie Mellon Université, USA. He was an Associate Professor of computer science with the Tele-Universitédu Québec, from 2012 to 2019. He has been a Full Professor with a cross-appointment in computer science and information technology with the Université de Sherbrooke, since 2019. His research interests include the development and evaluation of assistive technologies for people with cognitive deficiencies by developing human activity recognition system, ambient intelligence framework, assistive systems based on mixed reality, and multimodal interactions in mixed reality and ambient intelligence.

KÉVIN BOUCHARD (Member, IEEE) received the Ph.D. degree in computer science, in 2014. He is a Full Professor with the Université du Quebec á Chicoutimi. He went on to conduct a postdoctoral fellowship as a Project Scientist with the Center for SMART Health, University of California at Los Angeles. During this time, he co-led a deployment project of ambient technology for a rehabilitation center in Santa Monica. He has, among other things, worked on RFID localization, ambient sensing, activity recognition with UWB, and wearable technologies. He has published over 120 papers and has been supervising/co-supervising more than 40 graduate students. In addition to his academic work, he serves as a President for a non-profit organization called "Regroupement quebecois des maladies orphelines," which supports individuals affected by rare diseases. His research interests mainly include the exploitation of artificial intelligence and machine learning for the development of technologies for health.

SÉBASTIEN GABOURY (Senior Member, IEEE) received the Ph.D. degree in mathematics from the Royal Military College of Canada, Kingston, ON, Canada, in 2012. He is currently a Full Professor and the Canada Research Chair of ambient intelligence and wearable devices for technology support in health with the Université du Quebec á Chicoutimi, Canada. He is the Head of the Ambient Intelligence Laboratory for the Recognition of Activities. His research is sponsored by Canada Research Chair, the Natural Sciences and Engineering Research Council of Canada, Quebec Research Fund on Nature and Technologies, and the Canadian Foundation for Innovation. He has published more than 200 articles and has supervised or co-supervised more than 40 graduate students. His research interests include assistive technologies for older adults, individuals with cognitive impairment, and those with neuromuscular diseases.

MÉLANIE COUTURE received the Ph.D. degree. She is the current Chair and the Research Chair in older adult mistreatment and an Associate Professor with the School of Social Work, Université de Sherbrooke. She was a Researcher in social gerontology for more than a decade. Her research promotes the co-design and integration of clinical and organizational innovations for the prevention and management of mistreatment situations in the context of caregiving, common living environments, and in the use of technologies for aging in place.

NATHALIE BIER is a Full Professor of occupational therapy with the Université de Montreal, Canada. She is also a Researcher with the Research Center of the Institut Universitaire de Geriatrie de Montreal and the Associate Scientific Director. The main goal of her research program is to better understand the impact of cognitive deficits in aging and dementia on everyday function, and to develop non-pharmacological approaches to promote aging in place, such as the use of cognitive rehabilitation, e-health, and community mobilization. Her work is carried out using collaborative research approaches, such as action design research and living labs.



SYLVAIN GIROUX received the Ph.D. degree in computer science from the University of Montreal, in 1993. He is currently a Professor with the Department of Computer Science, University of Sherbrooke, Canada. His professional experience is well-balanced between academic institutions and private corporations. His current research interests include cognitive assistance, smart homes, the Internet of Things, ontologies, activity recognition, augmented reality, cryptocurrencies, aging at home, and transdisciplinary research.