Abstract—Learning time-series representations when only unlabeled data or few labeled samples are available can be a challenging task. Recently, contrastive self-supervised learning has shown great improvement in extracting useful representations from unlabeled data via contrasting different augmented views of data. In this work, we propose a novel Time-Series representation learning framework via Temporal and Contextual Contrasting (TS-TCC) that learns representations from unlabeled data with contrastive learning. Specifically, we propose time-series-specific weak and strong augmentations and use their views to learn robust temporal relations in the proposed temporal contrasting module, besides learning discriminative representations by our proposed contextual contrasting module. Additionally, we conduct a systematic study of time-series data augmentation selection, which is a key part of contrastive learning. We also extend TS-TCC to the semi-supervised learning settings and propose a Class-Aware TS-TCC (CA-TCC) that benefits from the available few labeled data to further improve representations learned by TS-TCC. Specifically, we leverage the robust pseudo labels produced by TS-TCC to realize a class-aware contrastive loss. Extensive experiments show that the linear evaluation of the features learned by our proposed framework performs comparably with the fully supervised training. Additionally, our framework shows high efficiency in few labeled data and transfer learning scenarios.

Index Terms—Augmentation, contextual contrasting, self-supervised learning, semi-supervised learning, temporal contrasting, time-series classification.

I. INTRODUCTION

TIME-SERIES data are being incrementally collected on a daily basis from IoT, wearable devices, and machines sensors for various applications in healthcare, manufacturing, etc. [1]. However, unlike images, time-series data generally do not contain human recognizable patterns to differentiate different classes and thus easily assign class labels to data. Consequently, different time-series applications require trained specialists to perform the challenging data annotation/labeling task. Therefore, little time-series data have been labeled in real-world applications compared to the collected data [2]. With the advance of deep learning techniques, it becomes essential to train deep learning-based models with a massive amount of labeled data to learn useful representations and avoid overfitting. However, applying them to time-series data becomes challenging with the aforementioned labeling limitations.

Self-supervised learning has gained much attention recently for exploiting unlabeled data to learn powerful representations for deep learning models. Compared with models trained on fully labeled data (i.e., supervised models), self-supervised pre-trained models can extract effective representations and achieve comparable performance when fine-tuned with a small percentage of the labels [3], [4]. The recent rebirth of self-supervised learning relied on manually-designed pretext tasks to learn representations about data. For example, Noroozi et al. split the image into smaller patches, shuffled them, and trained the model to reorder these patches to form the original image [3]. Some other pretext tasks assigned pseudo labels to different variations of input samples. For instance, Gidaris et al. applied several angles to the original image and assigned a label to each rotation of input samples. The code is publicly available at https://github.com/emadeldeen24/CA-TCC.

This article has supplementary downloadable material available at https://doi.org/10.1109/TPAMI.2023.3308189, provided by the authors.

Digital Object Identifier 10.1109/TPAMI.2023.3308189

Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.
input sample, which are generated using augmentation techniques [4], [6], [7]. However, these image-based contrastive learning methods may not be able to work well on time-series data for the following reasons. First, unlike images, where features are mostly spatial, the time-series data are mainly characterized by their temporal dependencies [8]. Therefore, applying the aforementioned techniques directly to time-series data may not efficiently address the temporal features of data. Second, some augmentation techniques used for images such as color distortion, generally cannot fit well with time-series data. So far, few works on contrastive learning have been proposed for time-series data. For example, [9], [10] developed contrastive learning methods for bio-signals. However, these two methods are proposed for specific clinical applications and may not generalize to other time-series data.

In this paper, we propose a novel framework that incorporates contrastive learning into self- and semi-supervised learning. Specifically, we propose a Time-Series representation learning framework via Temporal and Contextual Contrasting (TS-TCC) that is trained on totally unlabeled datasets. Our TS-TCC employs two contrastive learning and augmentation techniques to handle the temporal dependencies of time-series data. We propose simple yet efficient data augmentations that can fit any time-series data to create two different, but correlated views of the input samples. These views are then used by the two innovative contrastive learning modules. In the first module, we propose a novel temporal contrasting module to learn robust representations by designing a tough cross-view prediction task. Specifically, for a certain timestep, it utilizes the past latent features of one augmentation to predict the future of the other augmentation. This novel operation will force the model to learn robust representation by a harder prediction task against any perturbations introduced by different timesteps and augmentations. In the second module, we propose contextual contrasting to further learn discriminative representations upon the robust representations learned by the temporal contrasting module. In this contextual contrasting module, we aim to maximize the similarity among different contexts of the same sample while minimizing similarity among contexts of different samples. The pretrained model learns powerful representations about the time-series data regardless of downstream tasks.

We also extend TS-TCC to the semi-supervised settings, where few labeled samples are available, and propose the Class-Aware TS-TCC (CA-TCC). Specifically, we fine-tune the TS-TCC encoder with the few labeled samples and deploy it to generate pseudo labels for the whole unlabeled data. Next, we utilize the class information in the pseudo labels in the supervised contextual contrasting between samples to maximize the similarity between samples having the same class label, while minimizing the similarity between samples from different classes.

This journal paper is an extended version of our previous work [11]. The main adjustments are listed as follows: 1) We extend our TS-TCC model to fit the semi-supervised setting by taking advantage of the available few labeled data to further improve the learned representations. 2) We study the time-series data augmentations, which are considered as a key part of contrastive learning, in more details. We also provide a systematic methodology for selecting the best augmentations for time-series data. 3) We conduct additional experiments to assess our proposed CA-TCC, demonstrating the effectiveness of its learned representations in semi-supervised settings. In summary, the main contributions of this work are as follows.

- A novel contrastive learning framework for time-series representation learning with two variants is developed. The first (TS-TCC) is proposed for learning from unlabeled data, and the second (CA-TCC) is proposed for semi-supervised learning.
- We provide simple yet efficient augmentations that are designed for time-series data in the contrastive learning framework, and provide extensive discussion about their selection.
- We propose a novel temporal contrasting module to learn robust representations from time-series data by designing a tough cross-view prediction task. In addition, we propose a (supervised) contextual contrasting module to further learn discriminative representations upon the robust representations.
- We perform extensive experiments on the two proposed variants of our framework using ten real-world datasets. Experimental results show that both variants are capable of learning effective representations for different scenarios.

II. RELATED WORKS

A. Self-Supervised Learning

Self-supervised learning is gaining more attention recently, as it deals with the emerging problem of learning from unlabeled data [12], [13], [14]. The recent advances in self-supervised learning started with applying pretext tasks on images to learn useful representations. Many pretext tasks, that vary in nature, have been proposed according to the target applications. For image-related applications, Zhang et al. trained their model to color the input grayscale image [15]. Misra et al. proposed learning invariant representations based on pretext tasks by encouraging the representation of both the pretext task and the original image to be similar [16]. Zhai et al. developed pretext tasks based on self-supervised training to empower the representations learned in a semi-supervised setting [17].

For video-related tasks, Srivastava et al. proposed different tasks such as input sequence reconstruction, or future sequence prediction to learn representations of video sequences [18]. In addition, Wei et al. proposed training the model to check if the order of the input frame sequences is correct or not [19]. Another direction is to use ranking as a proxy task to solve regression problems [20], in which the authors developed a backpropagation technique for Siamese networks to prevent redundant computations. Although using pretext tasks can improve representation learning, they are found to limit the generality of the learned representations. For example, classifying the different rotation angles of an image may deviate the model from learning features about the color or orientation of objects [21].

On the other hand, contrastive methods intend to learn invariant representations from different augmented views of data.
The different methods vary between each other in their ways of choosing negative samples against positive samples during training. SimCLR [4] considered the augmented views of the original image as the positives, while all the other views of different images within the batch are treated as negatives. This technique benefits from larger batch sizes to accumulate more negative samples. Another approach was to accumulate a large number of the negative samples in a memory bank as proposed in MoCo [7]. The embeddings of these samples are updated with the most recent ones at regular intervals. The next direction was to use contrastive learning without the need for negative samples. For example, BYOL [22] learned representations by bootstrapping representations from two neural networks that interact and learn from each other. SimSiam [23] supported the idea of neglecting the negative samples and relied only on a Siamese network and stop-gradient operation to achieve state-of-the-art performance.

Some methods deployed contrastive learning in semi-supervised settings. For example, FixMatch [24] used weak augmented views to produce pseudo labels for the input image, and if the pseudo label exceeds a confidence threshold, they use it to penalize the prediction of the strong augmented view of the same input image. While all these approaches have successfully improved representation learning for visual data, they may not work well on time-series data that have different properties, such as temporal dependency.

### B. Self-Supervised Learning for Time-Series

Self-supervised representation learning for time-series is becoming more popular recently. Some approaches employed pretext tasks for time-series data. For example, Saeed et al. designed a binary classification pretext task for human activity recognition by applying several transformations on the data and trained the model to classify between the original and the transformed versions [25]. Similarly, Sarkar et al. proposed SSL-ECG, in which ECG representations are learned by applying six transformations to the dataset as pretext tasks, and assigned pseudo labels according to the transformation type [26]. The model is expected to learn useful representations about the data by classifying these transformations. The same approach was used by Saeed et al. as they designed eight auxiliary tasks and learning representations from multi-sensor human activity data [27]. Additionally, Aggarwal et al. learned subject-invariant representations by modeling local and global activity patterns [28].

Inspired by the success of contrastive learning in visual applications, few works have recently leveraged contrastive learning for time-series data. For example, CPC [21] learned representations by predicting the future in the latent space and showed great advances in various speech recognition tasks. Banville et al. studied three self-supervised tasks with two pretext tasks, i.e., relative positioning, temporal shuffling, and one contrastive task that uses CPC to learn representations about clinical EEG data [29]. They found that representations learned by CPC perform the best. This indicates that contrastive learning techniques generally perform better than pretext tasks.

Mohsenvand et al. designed EEG-related augmentations and extended the SimCLR model [4] to different EEG clinical data [9]. TS2Vec [30] proposed two contrastive losses, where the first pulls timesteps from the augmented views closer and pulls the different timesteps from the same time series away, while the second is the instance-wise contrasting.

Some methods also included the frequency domain characteristics to learn time-series representations. For example, Bilinear Temporal-Spectral Fusion (BTSF) [31] applied an iterative bilinear fusion between feature embeddings of both time and frequency representations of time series. Similarly, both TF-C [32] and STFNets [33] learned representations by pushing the time domain and frequency domain representations of the same sample closer to each other, while pushing them apart from representations of other signals.

Existing approaches used either temporal or global features. Differently, we address both types of features in our cross-view temporal and contextual contrasting modules. These modules rely on different views for input data that we provide via time-series-specific augmentations.

### C. Semi-Supervised Learning for Time-Series

In semi-supervised learning, a model is trained on a dataset that is partially labeled, i.e., some samples in the dataset have been labeled, where it may be expensive or time-consuming to label all of the examples in the dataset. There are different approaches for time-series semi-supervised learning.

The first approach is to apply generic regularizations, e.g., ensembling and reconstruction. For example, TEBLSTM [34] applies temporal ensembling based on LSTM output features. It ensembles two losses, i.e., the supervised loss on the labeled data, and an unsupervised loss based on the distance of the current prediction and the previous ensemble output. Also, REG-GAN [35] uses Generative Adversarial Networks (GAN) to learn representations by regenerating the signals of the whole data and applying supervised regression loss on the labeled part.

The second approach is self-training, where the unlabeled samples are assigned pseudo labels. For example, SUCCESS [36] clustered the data, then assigned pseudo labels based on the Dynamic Time Warping (DTW) distance. In addition, SelfMatch [37] deployed Gaussian noised signals to guide the prediction of the augmented view of the same signal. It also applied self-distillation by guiding the lower-level blocks of the feature extractor with the knowledge obtained in the output layer. Last, SemiTime [38] split the signal into two parts, i.e., past and future. Next, it applied contrastive loss to push the past of the signal toward its future, while pulling it away from the future parts of other signals.

### III. METHODS

This section describes the components of the proposed framework in details. Starting with TS-TCC, we first generate two different yet correlated views of the input data based on strong and weak augmentations. Then, a temporal contrasting module is proposed to explore the temporal features of the data with two autoregressive models. These models perform a tough cross-view...
prediction task by predicting the future of one view using the past of the other. We further maximize the agreement between the contexts of the autoregressive models by a contextual contrasting module. These components are illustrated in Fig. 1. For the semi-supervised settings, our proposed CA-TCC starts with TS-TCC pretraining phase and includes three more phases to complete the semi-supervised training, as shown in Fig. 3. In the next subsections, we will introduce each component in more details.

A. Time-Series Data Augmentation

Data augmentation is a key part in the success of contrastive learning methods [4], [22]. Contrastive methods try to maximize the similarity among different views of the same sample while minimizing its similarity with other samples. It is thus important to design proper data augmentations for contrastive learning. Usually, contrastive learning methods use two (random) variants of the same augmentation [4], [9]. Specifically, given a sample \( x \), they produce two views \( x_1 \) and \( x_2 \) sampled from the same augmentation family \( T \), i.e., \( x_1 \sim T \) and \( x_2 \sim T \). However, we argue that producing views from different augmentations can improve the robustness of the learned representations. Consequently, we propose two separate augmentations, such that one augmentation is weak and the other is strong.

Our framework uses strong augmentation to enable the tough cross-view prediction task in the next module, which helps in learning robust representations about the data. The weak augmentation aims to add some small variations to the signal without affecting its characteristics or making major changes in its shape. We include both types of augmentation to introduce variations of data, which increases the model’s generalization ability to perform well in the unseen test set. We discuss more details about the augmentation selection in Section V-F.

Formally, for each input sample \( x \), we denote its strongly augmented view as \( x^s \), and its weakly augmented view as \( x^w \), where \( x^s \sim T_s \) and \( x^w \sim T_w \). These views are then passed to the encoder to extract their high dimensional latent representations. In particular, the encoder has a 3-block convolutional architecture as proposed in [39]. For an input \( x \), the encoder maps \( x \) into a high-dimensional latent representation \( z = f_{enc}(x) \). We define \( z = [z_1, z_2, \ldots, z_T] \), where \( T \) is the total timesteps, \( z_i \in \mathbb{R}^d \), where \( d \) is the feature length. Thus, we get \( z^s \) for the strong augmented views, and \( z^w \) for the weak augmented views, which are then fed into the temporal contrasting module.

B. Temporal Contrasting

The Temporal Contrasting module deploys a contrastive loss to extract temporal features in the latent space with an autoregressive model. Given the latent representations \( z \), the autoregressive model \( f_{ar} \) summarizes all \( z_{c,t} \) into a context vector \( c_t = f_{ar}(z_{c,t}) \), \( c_t \in \mathbb{R}^h \), where \( h \) is the hidden dimension of \( f_{ar} \). The context vector \( c_t \) is then used to predict the timesteps from \( z_{t+1} \) until \( z_{t+k} (1 < k \leq K) \). To predict future timesteps, we use a log-bilinear model that would preserve the mutual information between the input \( x_{t+k} \) and \( c_t \), such that \( f_k(x_{t+k}, c_t) = \exp((W_k(c_t))^T z_{t+k}) \), where \( W_k \) is a linear function that maps \( c_t \) back into the same dimension as \( z \), i.e., \( W_k : \mathbb{R}^h \rightarrow \mathbb{R}^d \).

In our approach, the strong augmentation generates \( c^s_t \) and the weak augmentation generates \( c^w_t \). We propose a tough cross-view prediction task by using the context of the strong augmentation \( c^s_t \) to predict the future timesteps of the weak augmentation \( z^w_{t+k} \) and vice versa. The contrastive loss tries to maximize the dot product between the predicted representation and the true one of the same sample while minimizing the dot product with the other samples \( N_{l,k} \) within the mini-batch. Accordingly, we calculate the two losses \( L^T_{TC} \) and \( L^w_{TC} \) as follows:

\[
L^s_{TC} = -\frac{1}{K} \sum_{k=1}^{K} \log \sum_{n \in N_{l,k}} \exp \left( (W_k(c^s_t))^T z^w_{t+k} \right)
\]

\[
L^w_{TC} = -\frac{1}{K} \sum_{k=1}^{K} \log \sum_{n \in N_{l,k}} \exp \left( (W_k(c^w_t))^T z^s_{t+k} \right)
\]
layers with a non-linearity ReLU function and dropout in between. Pre-norm residual connections, which can produce more stable gradients [41], are adopted in our Transformer. We stack $L$ identical layers to generate the final features. Inspired by the BERT model [42], we add a token ${e} \in \mathbb{R}^k$ to the input, whose state acts as a representative context vector in the output. The operation of the Transformer starts by applying the features $z_{<i}$ to a linear projection $\mathcal{W}_{\text{Tran}}$ layer that maps the features into the hidden dimension, i.e., $\mathcal{W}_{\text{Tran}} : \mathbb{R}^{d \rightarrow h}$. The output of this linear projection is then sent to the Transformer i.e., $\hat{z} = \mathcal{W}_{\text{Tran}}(z_{<i})$. $\hat{z} \in \mathbb{R}^h$. Next, we attach the context vector into the feature vector $\hat{z}$ such that the input features become $\psi_0 = [c; \hat{z}]$, where the subscript 0 denotes being the input to the first layer. Next, we pass $\psi_0$ through Transformer layers as in the following equations:

$$\hat{\psi}_1 = \text{MHA}(\text{Norm}(\psi_{1-1})) + \psi_{1-1}, \quad 1 \leq l \leq L;$$

$$\psi_1 = \text{MLP}(\text{Norm}(\hat{\psi}_1)) + \hat{\psi}_1, \quad 1 \leq l \leq L. \quad (4)$$

Finally, we re-attach the context vector from the final output such that $c_l = \psi_L^0$. This context vector will be the input of the following contextual contrasting module.

### C. Contextual Contrasting

We further propose a contextual contrasting module that aims to learn more discriminative representations. It starts by applying a non-linear transformation to the contexts using a non-linear projection head as in [4]. The projection head maps the contexts into the space where the contextual contrasting is applied.

Given a batch of $N$ input samples, we will have two contexts for each sample from its two augmented views and thus have $2N$ contexts. For a context $c_l^t$, we denote $c_l^{t+}$ as the positive sample of $c_l^t$ that comes from the other augmented view of the same input, and hence, $(c_l^t, c_l^{t'})$ is considered to be a positive pair. Meanwhile, the remaining $(2N - 2)$ contexts from other inputs within the same batch are considered as the negative samples of $c_l^t$, i.e., $c_l$ can form $(2N - 2)$ negative pairs with its negative samples. Therefore, we can derive a contextual contrasting loss to maximize the similarity between the sample and its positive pair, while minimizing its similarity with the negative samples within the mini-batch. As such, the final representations can be discriminative.

Equation (6) defines the contextual contrasting loss function $L_{SCC}$. Given a context $c_l^t$, we divide its similarity with its positive sample $c_l^{t+}$ by its similarity with all the other $(2N - 1)$ samples, including the positive pair and $(2N - 2)$ negative pairs, to normalize the loss.

$$\ell(i, i^+) = -\log \frac{\exp \left( \text{sim} \left( c_l^t, c_l^{t+} \right) / \tau \right)}{\sum_{m=1}^{N \setminus \{i\}} \exp \left( \text{sim} \left( c_l^t, c_l^{t''} \right) / \tau \right)}, \quad (5)$$

$$L_{SCC} = \frac{1}{2N} \sum_{k=1}^{2N} [\ell(2k - 1, 2k) + \ell(2k, 2k - 1)], \quad (6)$$

where $\text{sim}(u, v) = \|u^T v\|/\|u\|\|v\|$ denotes the dot product between $\ell_2$ normalized $u$ and $v$ (i.e., cosine similarity), $\|m \setminus \{i\}\| \in \{0, 1\}$ is an indicator function, evaluating to 1 iff $m \neq i$, and $\tau$ is a temperature parameter.

The overall self-supervised loss is the combination of the two temporal contrasting losses and the contextual contrasting loss as follows.

$$L_{unsup} = \lambda_1 \cdot (L_{TC} + L_{wTC}) + \lambda_2 \cdot L_{SCC}, \quad (7)$$

where $\lambda_1$ and $\lambda_2$ are fixed scalar hyperparameters denoting the relative weight of each loss.

### D. Class-Aware TS-TCC

We propose a second variant of our framework, called the Class-Aware TS-TCC (CA-TCC), to operate in semi-supervised settings. In this variant, we attempt to exploit the available few labeled data to further improve the representation learned by TS-TCC. Our semi-supervised learning framework is performed in four phases, as shown in Fig. 3. In Phase 1, we start with a randomly-initialized encoder and used it in TS-TCC self-supervised pretraining. In Phase 2, we fine-tune the pretrained encoder with the few labeled data. In Phase 3, we deploy the fine-tuned encoder to create pseudo labels for the entire unlabeled dataset. Finally, in Phase 4, we train the class-aware semi-supervised framework, i.e., CA-TCC, with the pseudo labels.

In the class-aware semi-supervised training, we replace the unsupervised contextual contrasting with a supervised contextual contrasting module. This module benefits from the pseudo-labeled data to  train a supervised contrastive loss [43]. This loss considers samples with the same class label as positive pairs, while samples from different classes are considered as negative pairs. This is different from the unsupervised contrastive loss that only forms positive pairs from the augmented views of the sample, as depicted in Fig. 4. This difference can affect the performance of the model, as the unsupervised contrastive loss may treat samples having a similar class as negative pairs.

Formally, assuming that the dataset consists of $N$ labeled samples $\{x_k, y_k\}_{k=1...N}$, then after applying augmentations, the dataset becomes of $2N$ samples, $\{\hat{x}_i, \hat{y}_i\}_{i=1...2N}$ such that $\hat{x}_{2k}$ and $\hat{x}_{2k-1}$ are the two views of $x_k$, and similarly $\hat{y}_k = \hat{y}_{2k} = \hat{y}_{2k-1}$. Also assuming that $i \in I \equiv \{1 \ldots 2N\}$ represents the index of an arbitrary augmented sample, and $A(i) \equiv I \setminus \{i\}$, the supervised contextual contrasting loss can be expressed as:

$$L_{SCC} = \sum_{i \in I} \frac{1}{|P(i)|} \sum_{p \in P(i)} \ell(i, p), \quad (8)$$

$$P(i) = \{p \in A(i) : \hat{y}_p = \hat{y}_i\}, \quad (9)$$

where $P(i)$ is the set of indices of all samples with the same class as the sample $\hat{x}_i$ in the batch, and $(i, p)$ for any $p \in P(i)$ is thus a positive pair. $|P(i)|$ is the cardinality of $P(i)$. Thus, the overall loss can be expressed as:

$$L_{semi} = \lambda_3 \cdot (L_{TC} + L_{wTC}) + \lambda_4 \cdot L_{SCC}, \quad (10)$$

where $\lambda_3$ and $\lambda_4$ are fixed scalar hyperparameters denoting the relative weight of each loss.
Fig. 3. Four phases for CA-TCC semi-supervised training. In Phase 1, TS-TCC is trained with fully unlabeled data. Next, we use the available few labeled samples to fine-tune the pretrained encoder in Phase 2. Following that, in Phase 3, we generate pseudo labels for the unlabeled data with the fine-tuned TS-TCC encoder. Finally, Phase 4 includes deploying the pseudo-labeled data in the semi-supervised training.

Fig. 4. (a) Unsupervised contrasting versus (b) supervised contrasting. The unsupervised contrasting forms positive pairs only from the augmented views of the sample, which may deteriorate the performance. In contrast, supervised contrasting considers the semantics of the data and forms positive pairs among all the samples having the same class label. In CA-TCC, we deploy the pseudo-labeled data to apply the supervised contrasting in the semi-supervised training.

IV. EXPERIMENTAL SETUP

A. Datasets

To comprehensively evaluate our proposed models, we adopted ten publicly available real-world datasets that cover different time-series applications. Additionally, we investigated the transferability of our learned features on a fault diagnosis dataset.

1) Human Activity Recognition: We employed UCI HAR dataset\textsuperscript{1} [44], which contains sensor readings for 30 subjects performing 6 activities (i.e., walking, walking upstairs, downstairs, standing, sitting, and lying down). The data were collected by a mounted Samsung Galaxy S2 device on the users’ waist, with a sampling rate of 50 Hz.

2) Sleep Stage Classification: We adopted Sleep-EDF dataset\textsuperscript{2} [45], which includes whole-night polysomnography (PSG) sleep recordings for 20 subjects. In particular, each recording contains two electroencephalogram (EEG) channels namely Fpz-Cz and Pz-Oz, with a sampling rate of 100 Hz. In this work, we used a single EEG channel (i.e., Fpz-Cz), following previous studies [46], [47]. Sleep stage classification refers to classifying the input EEG signal into one of five classes: Wake (W), Non-rapid eye movement (N1, N2, N3), and Rapid Eye Movement (REM).

3) Epilepsy Seizure Prediction: The Epileptic Seizure Recognition dataset\textsuperscript{3} [48] consists of EEG recordings from 500 subjects, where the brain activity was recorded for each subject for 23.6 seconds. The original dataset is labeled with five classes, but as four of them do not include epileptic seizures, we merged them into one class and treat it as a binary classification problem.

4) Fault Diagnosis (FD): Fault Diagnosis dataset\textsuperscript{4} [49] was collected from sensor readings of bearing machine under four different working conditions. Each working condition can be considered as a separate domain as it has different characteristics from the other working conditions (e.g., rotational speed and load torque) [50]. Each domain has three classes; two fault classes (i.e., inner fault and outer fault) and one healthy class. We use this dataset to conduct the transferability experiment only and show the effectiveness of our method in transfer learning scenarios.

5) UCR Repository Datasets: We included seven UCR datasets\textsuperscript{5} in our experiments. The selected datasets are suitable for our few labels experiments, and they satisfy the condition that 1% of samples allow all the classes to be present during the training phase. The included datasets are: Wafer, FordA, FordB, PhalangesOutlinesCorrect (POC), ProximalPhalanxOutlineCorrect (PPOC), StarLightCurves, and ElectricDevices. We included a detailed discussion about each of these datasets in the supplementary materials, available online.

Table I summarizes the statistical details of each dataset, i.e., the number of training samples (# Train) and testing samples (# Test).

\textsuperscript{1}https://archive.ics.uci.edu/ml/datasets/human+activity+recognition+using+smartphones
\textsuperscript{2}https://physionet.org/physiobank/database/sleep-edf/
\textsuperscript{3}https://archive.ics.uci.edu/ml/datasets/Epileptic+Seizure+Recognition
\textsuperscript{4}https://mb.uni-paderborn.de/en/kat/main-research/datacenter/bearing-datacenter/data-sets-and-download
\textsuperscript{5}https://timeseriesclassification.com/dataset.php
TABLE I  
BRIEF DESCRIPTION OF THE 10 DATASETS USED IN OUR EXPERIMENTS  

| Dataset               | # Train | # Test  | Length | # Channel | # Class |
|-----------------------|---------|---------|--------|-----------|---------|
| HAR                   | 7,352   | 2,947   | 128    | 9         | 6       |
| Sleep-EDF             | 25,612  | 8,910   | 3,000  | 1         | 5       |
| Epilepsy              | 9,200   | 2,300   | 178    | 1         | 2       |
| FD                    | 8,184   | 2,728   | 5,120  | 1         | 3       |
| Wafer                 | 3,000   | 6,174   | 152    | 1         | 2       |
| FordA                 | 1,320   | 3,601   | 500    | 1         | 2       |
| FordB                 | 610     | 3,636   | 500    | 1         | 2       |
| POC                   | 1,400   | 385     | 80     | 1         | 2       |
| PPOC                  | 600     | 291     | 80     | 1         | 2       |
| StarLightCurves       | 1,000   | 8,286   | 1,024  | 1         | 3       |
| ElectricDevices       | 8,926   | 7,711   | 96     | 1         | 7       |

For the FD dataset, we mentioned the settings of one working condition, as they are the same for the four working conditions.

Test), the length of the sample, the number of sensor channels (# Channel) and the number of classes (# Class).

B. Implementation Details

We split the data into 60%, 20%, and 20% for training, validation, and testing with considering subject-wise split for the Sleep-EDF dataset to avoid overfitting. Experiments were repeated 5 times with 5 different seeds, and we reported the performance with mean and standard deviation. The pretraining and downstream tasks were done for 40 epochs, as we noticed that the performance did not improve with further training. We applied a batch size of 128 (which was reduced in few-labeled data experiments as data size may be less than 128). We used Adam optimizer with a learning rate of 3e-4, weight decay of 3e-4, $\beta_1 = 0.9$, and $\beta_2 = 0.99$. For the strong augmentation, we set the number of segments $M$ as $M_{EP}=12$, $M_{EDF}=20$, and $M=10$ for all other datasets, (see Section V-F), while for the weak augmentation, we set the scaling ratio to 2. In TS-TCC, we set $\lambda_1 = 1$, while we achieved good performance when $\lambda_2$ is around 1. Particularly, we set $\lambda_2 = 0.7$ in our experiments. For CA-TCC, was set $\lambda_3 = 0.01$ and $\lambda_4 = 0.7$ (see Section V-E). In the Transformer, we set the $L = 4$, and the number of heads as 4. We tuned $h \in \{32, 50, 64, 100, 128, 200, 256\}$ and set $h_{EDF}=64$ and $h=100$ for all other datasets. We also set its dropout to 0.1. In contextual contrasting, we set $\tau = 0.2$. Lastly, we built our model using PyTorch 1.7 and trained it on NVIDIA GeForce RTX 2080 Ti GPU.

V. EXPERIMENTAL RESULTS

We conduct several experiments to demonstrate the efficacy of our proposed TS-TCC and CA-TCC models. We measure the performance using two metrics namely the accuracy and the macro-averaged F1-score (MF1) to better verify performance on the imbalanced datasets. These metrics are defined as follows.

\[
\text{Accuracy} = \frac{\sum_{i=1}^{K} TP_i}{N},
\]

\[
\text{MF1-score} = \frac{1}{K} \sum_{i=1}^{K} \frac{2 \times \text{Precision}_i \times \text{Recall}_i}{\text{Precision}_i + \text{Recall}_i},
\]

where $\text{Precision}_i = \frac{TP_i}{TP_i + FP_i + \text{FN}_i}$ and $\text{Recall}_i = \frac{TP_i}{TP_i + \text{FN}_i}$. $TP_i$, $FP_i$, and $\text{FN}_i$ denote the True Positives, False Positives, and False Negatives for the $i$-th class respectively, $N$ is the total number of samples, and $K$ is the number of classes in the dataset.

Although the results of these two equations are expected to be in a range from 0 to 1, we report them in this paper as a percentage.

A. Results of Self-Supervised TS-TCC Model

1) Comparison With Baseline Approaches: We compare our proposed TS-TCC against the following baselines.

1) Random Initialization: by training a linear classifier on top of a frozen and randomly initialized encoder, and this represents the lower bound.

2) Supervised: supervised training of both: the encoder and the classifier.

3) SSL-ECG [26]: an auxiliary-based self-supervised learning algorithm, based on classifying signal transformations.

4) CPC [21]: a contrastive self-supervised learning method that predicts future timesteps in the embedding space.

5) SimCLR [4]: a contrastive self-supervised learning method that deploys the unsupervised contrastive loss based on data augmentations. It is worth noting that, we use our time-series specific augmentations to pretrained SimCLR, as it was originally designed for images.

To evaluate the performance of SSL-ECG, CPC, SimCLR, and TS-TCC, we first pretrain each of them with the fully unlabeled data, then use a portion of the labeled data to evaluate their performance. We follow the standard linear evaluation scheme [4], [21], in which we train a linear classifier (single fully connected layer) on top of a frozen self-supervised pretrained encoder model.

Table II shows the linear evaluation results of our approaches against the baseline methods using 1% and 5% of the labeled data. The results show that TS-TCC achieves the best overall performance across all 10 datasets in both the 1% and 5% label scenarios. Specifically, TS-TCC is the best performer on six datasets and the second best on two datasets with only a small margin behind the best methods. This demonstrates that our proposed cross-view prediction task can effectively improve representation learning for various time-series datasets. It is worth noting that contrastive methods (e.g., CPC, SimCLR, and our TS-TCC) generally achieve better results than the pretext-based method (i.e., SSL-ECG), which reflects the power of invariant features learned by contrastive methods. Additionally, CPC shows better results than SimCLR, being the best on one dataset and the second best on five datasets. This highlights the importance of learning temporal dependencies in time-series data for better self-supervised representation learning, as implemented in CPC and TS-TCC.

2) Performance Under Different Labeling Budgets: We investigate our TS-TCC performance under different fractions of few labels, by fine-tuning the pretrained model using 1%, 5%, 10%, 50%, and 75% of randomly selected instances of the training data. Fig. 5 shows the results of our fine-tuned TS-TCC along with the supervised training under the aforementioned
TABLE II
RESULTS OF FINE-TUNING SELF-SUPERVISED PRETRAINED MODELS WITH 1% AND 5% OF LABELS

| 1% of labeled data | 5% of labeled data |
|---------------------|---------------------|
| **Datasets** | **Accuracy** | **MF1-score** | **Accuracy** | **MF1-score** | **Accuracy** | **MF1-score** | **Accuracy** | **MF1-score** |
| HAR | 39.8±3.8 | 34.2±5.0 | 19.6±2.7 | 14.1±0.7 | 35.4±2.4 | 26.2±2.7 | 35.4±2.4 | 26.2±2.7 |
| Sleep-EDF | 35.5±3.0 | 20.7±3.4 | 36.5±3.0 | 20.7±3.4 | 36.5±3.0 | 20.7±3.4 | 36.5±3.0 | 20.7±3.4 |
| Epilepsy | 50.1±1.6 | 39.1±1.2 | 53.1±1.2 | 45.1±1.2 | 53.1±1.2 | 45.1±1.2 | 53.1±1.2 | 45.1±1.2 |
| Wakes | 36.8±2.2 | 47.5±4.4 | 51.9±2.6 | 48.0±3.4 | 44.6±4.2 | 60.5±7.9 | 44.6±4.2 | 60.5±7.9 |
| ForeA | 82.1±1.6 | 38.6±4.0 | 62.0±0.8 | 40.0±2.1 | 62.3±1.8 | 41.2±4.9 | 62.3±1.8 | 41.2±4.9 |
| ForeB | 48.9±1.6 | 59.5±4.8 | 51.9±2.6 | 74.8±0.9 | 74.8±0.9 | 74.8±0.9 | 74.8±0.9 | 74.8±0.9 |
| StarlightCurves | 89.3±1.6 | 65.4±3.6 | 87.8±0.9 | 71.4±0.9 | 78.3±0.9 | 72.0±0.8 | 78.3±0.9 | 72.0±0.8 |
| ElectricDevices | 58.0±1.4 | 49.1±3.6 | 57.6±1.1 | 47.5±1.0 | 60.4±1.1 | 50.6±0.9 | 60.4±1.1 | 50.6±0.9 |

**Average** | 52.8±2.0 | 45.4±2.0 | 52.8±2.0 | 45.4±2.0 | 52.8±2.0 | 45.4±2.0 | 52.8±2.0 | 45.4±2.0 |

**Best results across each row are in bold, while the second-best results are underlined.**

settings on three datasets. We evaluate the performance with MF1 because of the imbalance in the Sleep-EDF dataset. In particular, TS-TCC fine-tuning (i.e., red curves in Fig. 5) means that we fine-tune the pretrained encoder with the few labeled samples.

We observe that supervised training performs poorly with limited labeled data, while our TS-TCC fine-tuning achieves significantly better performance than supervised training. For example, with only 1% of labeled data, TS-TCC fine-tuning can still achieve 69.5% and 89.2% for HAR and Epilepsy datasets respectively, compared with 47.6% and 74.8% for the supervised training on the same datasets. Furthermore, with only 10% of labeled data, our fine-tuned TS-TCC can achieve comparable performance with the supervised training with 100% of labeled data in the three datasets, demonstrating its effectiveness under different semi-supervised settings.

As shown in Fig. 5, TS-TCC fine-tuning is able to achieve impressive performance with few labeled data. Hence, TS-TCC is expected to generate high-quality pseudo labels in Fig. 3, which motivates us to propose the variant CA-TCC.

B. Results of Semi-Supervised CA-TCC Model

1) Comparison With Baseline Approaches: To examine the efficacy of our CA-TCC in semi-supervised settings, we compare it against the following semi-supervised learning baselines.

   1) **Mean-Teacher** [51]: it consists of two models. The first is the student model, which is the base regular model. The second is the teacher model, which is the averaging of the student model weights using the Exponential Moving Average (EMA) over the training steps.

   2) **DivideMix** [52]: it discards the labels that are likely to be noisy from the training and leverages them as unlabeled data to regularize the model against overfitting.

   3) **SemiTime** [38]: it learns the temporal relations in unlabeled data by splitting the signal into past-future pairs, then contrasting the past of one sample with two future splits from another sample.

   4) **FixMatch** [24]: it generates pseudo labels for the weak augmented view of the signal, and uses it to produce pseudo labels for the strong augmented view if it exceeds a confidence threshold.

Fig. 5. Comparison between supervised training versus TS-TCC fine-tuning on three datasets with different fractions of few labeled data in terms of MF1-score.
Notably, the four baselines train a cross-entropy loss on the labeled portion of the samples. Also, for FixMatch, we apply our proposed weak and strong augmentations for training. Table IV compares the performance of CA-TCC against semi-supervised baselines across 10 different datasets. Overall, the average F1-score of our CA-TCC across the 10 datasets outperforms the average F1-score of the second-best performing semi-supervised baseline with 3.8% and 4.9% in 1% and 5% of labels, respectively. Specifically, our CA-TCC ranks first in seven datasets and second in two other datasets. The consistent superior performance of our CA-TCC across different datasets and labeling budgets demonstrates its effectiveness in maximizing the utilization of labeled data in semi-supervised settings.

2) CA-TCC Model Analysis: We discuss the different aspects that yield the improved performance of the CA-TCC. Specifically, we regard the improved performance of CA-TCC to two main factors. The first is the effective representations learned by TS-TCC in Phase 1, which enables generating high-quality pseudo labels. This is demonstrated in Table II, where TS-TCC outperforms other self-supervised learning baselines. Therefore, it was anticipated that the fine-tuned TS-TCC encoder should be superior when used to produce the pseudo labels. The second factor is the supervised contrastive loss in CA-TCC, which uses the pseudo-labeled data to include more positive pairs in the contrastive loss. Nevertheless, our CA-TCC is flexible in architecture and can anticipate different models throughout its different phases. Next, we attempt to validate the aforementioned factors and support our conclusions.

a) Quality of pseudo labels: To investigate the effect of pseudo labels, we compare the quality of the pseudo labels generated by each baseline. Specifically, we replace our TS-TCC (in Phase 1) with SSL-ECG, SimCLR, and CPC. Then, for each baseline, we fine-tune the encoder and use it to generate pseudo labels (Phases 2 and 3) and then, conduct the semi-supervised training. Fig. 6 shows the accuracy of the generated pseudo labels (generated in Phase 3) compared to the true labels. We find that the pseudo labels generated by TS-TCC are more accurate than those generated by the other baselines. In addition, Table V reports the performance with different baselines (deployed in Phase 1), while fixing CA-TCC (in Phase 4) for two datasets with different scales, i.e., HAR and Sleep-EDF. We observe that using TS-TCC in both phases results in the best performance, supporting our conclusion and justification for the first reason for improved performance.

b) Effect of supervised contrastive loss: In this experiment, we compare the performance of the semi-supervised training
**Fig. 6.** Accuracy of generated pseudo labels with different self-supervised learning methods applied to HAR and Sleep-EDF datasets. Fine-tuning in Phase 2 is performed with 1% of labels.

**TABLE V**

| Unsupervised pretraining | Semi-supervised training | HAR | Sleep-EDF |
|--------------------------|--------------------------|-----|-----------|
|                          |                          | Acc | F1-score  | Acc | F1-score  |
| SSL-EEG                  | CA (TS-TCC)              | 70.6±0.9 | 66.4±0.8 | 76.1±0.4 | 67.1±0.6 |
| SimCLR                   | CA (TS-TCC)              | 73.2±1.2 | 70.5±1.2 | 78.3±0.7 | 68.1±0.5 |
| CPC                      | CA (TS-TCC)              | 72.7±0.7 | 69.6±0.8 | 79.2±0.4 | 70.0±0.6 |
| TS-TCC                   | CA (TS-TCC)              | 77.9±0.8 | 76.2±0.1 | 79.4±0.1 | 70.8±0.5 |
| SSL-EEG                  | CA (SimCLR)              | 69.5±1.3 | 66.7±1.1 | 75.5±0.2 | 68.2±0.9 |
| SimCLR                   | CA (SimCLR)              | 67.6±1.4 | 64.3±1.2 | 77.5±0.2 | 67.9±0.6 |
| CPC                      | CA (SimCLR)              | 72.3±1.9 | 69.1±2.6 | 79.2±0.1 | 70.1±0.7 |
| TS-TCC                   | CA (SimCLR)              | 75.0±1.5 | 73.1±2.0 | 79.9±0.3 | 70.2±0.3 |

Pretraining in the first phase is fully unsupervised, while the second phase is class-aware (TS-TCC or SimCLR) with generated pseudo labels from fine-tuning with 1% of labels.

(Phase 4) when using the supervised contrastive loss against the unsupervised contrastive loss. The supervised contrastive loss considers samples having the same class label as positive pairs and samples from different classes as negative pairs. In contrast, the unsupervised contrastive loss only forms positive pairs from the augmented views of the sample, and all other samples in the mini-batch are considered as negative pairs. This difference can affect the performance of the model, as the unsupervised contrastive loss may treat samples having a similar class as negative pairs.

Fig. 7 compares the performance of the unsupervised and the supervised contrastive losses in terms of accuracy and F1-score on the HAR and Sleep-EDF datasets. The results indicate that the use of the supervised contrastive loss, which includes more positive pairs, results in improved performance and better representation learning on both datasets.

c) Varying **CA-TCC model architecture:** In this section, we evaluate the performance of different combinations of self-supervised learning algorithms in Phases 1 and 4. Specifically, we use different self-supervised algorithms in Phase 1, while in Phase 4, we deploy only two baselines that can use the supervised contrastive loss, i.e., SimCLR and TS-TCC. The experiments are conducted on the HAR and Sleep-EDF datasets using 1% of labels.

The results, shown in Table V, indicate that using class-aware TS-TCC in the semi-supervised training with any baseline in Phase 1 consistently outperforms class-aware SimCLR for both datasets. This superior performance of TS-TCC can be regarded to its ability to learn temporal relations in time-series data while also benefiting from the supervised contrasting, unlike class-aware SimCLR which only learns through contrasting positive and negative pairs.

In summary, our framework is flexible and allows using various models in different training phases. However, we chose to only use TS-TCC as it is the best-performing self-supervised learning model among the baselines and it is able to handle class-aware training in semi-supervised settings.

**C. Transfer Learning Experiment**

We further examine the transferability of the learned features by designing a transfer learning experiment. We use Fault Diagnosis (FD) dataset for the evaluation under the transfer learning setting. Recall that the FD dataset has four working conditions, which are considered as four domains (denoted as domains A, B, C, and D). Here, we train the model on the data from one condition (i.e., source domain) and test it on another condition (i.e., target domain). We adopt three training schemes on the source domain, namely, 1) Supervised training, 2) TS-TCC fine-tuning, and 3) CA-TCC fine-tuning. In TS-TCC and CA-TCC fine-tuning, we fine-tune our pretrained encoder using the labeled data in the source domain.

Table III shows the performance of the three training schemes under 12 cross-domain scenarios. Clearly, our pretrained TS-TCC model consistently outperforms the supervised pretraining in 8 out of 12 cross-domain scenarios. Similarly, with only 1% of labels in each source domain for training, we find that CA-TCC model outperforms the supervised pretraining in 9 out of 12 cross-domain scenarios. We find that TS-TCC model can achieve at least \( \sim 7\% \) improvement in 7 out of 8 winning scenarios (except for D \( \rightarrow \) B scenario). Similarly, CA-TCC model can achieve at least \( \sim 8\% \) improvement in 7 out of 9 winning scenarios. Overall, our two proposed approaches can improve the transferability of learned representations over the supervised training by \( \sim 4\% \) and 6% in terms of accuracy.
A. Ablation Study

We study the effectiveness of each component in our proposed CA-TCC model. Specifically, we derive different model variants for comparison as follows. First, we train the Temporal Contrastive (TC) module without the cross-view prediction task, where each branch predicts the future timesteps of the same augmented view. This variant is denoted as ‘TC only’. Second, we train the TC module with adding the cross-view prediction task, which is denoted as ‘TC + X-Aug’. Third, we train the proposed TS-TCC model, which is denoted as ‘TC + X-Aug + CC’. Finally, we train the proposed CA-TCC model, which is denoted as ‘TC + X-Aug + SCC’. We also study the effect of using a single family of augmentations on the performance of TS-TCC and CA-TCC. In particular, for an input \( x \), we generate two different views \( x_1 \) and \( x_2 \) from the same augmentation type, i.e., \( x_1 \sim T_w \) and \( x_2 \sim T_w \) when using either the weak augmentation or the strong augmentation. We show the linear evaluation results in terms of accuracy and macro F1-score with only 5\% throughout these experiments.

Table VI shows this ablation study on the three datasets. Clearly, the proposed cross-view prediction task generates robust features and thus improves the performance by more than 6\% accuracy on HAR datasets, and \( \sim 2\% \) on Sleep-EDF and Epilepsy datasets. Additionally, the contextual contrasting module further improves the performance, as it helps the features to be more discriminative. More improvement was achieved by using supervised contextual contrasting in CA-TCC, which supports the importance of considering more positive samples from the same class to generate more discriminative features. By studying the effect of augmentations on TS-TCC, we find that generating different views from the same augmentation type is not helpful with HAR and Sleep-EDF datasets. For these complex datasets, using only weak augmentations may not make a tough cross-view prediction task, leading to close results to the variant ‘TC only’. Counterpart, using only strong augmentations deviates the model from recognizing the original data while testing. However, the less complex Epilepsy dataset can still achieve comparable performance with only one augmentation. For CA-TCC, we find that it consistently outperforms the results of TS-TCC, showing its effectiveness to improve the representations with the available few labeled samples. For example, we find that it highly improves the performance of using only weak or strong augmentations in both HAR and Sleep-EDF datasets.

E. Sensitivity Analysis

We perform sensitivity analysis on the HAR dataset to study five parameters namely, the number of predicted future timesteps \( K \) in the temporal contrastive module, \( \lambda_1 \) and \( \lambda_2 \) in (7), and \( \lambda_3 \) and \( \lambda_4 \) in (10). In specific, we used the linear evaluation experiment with full labels to assess the performance.

Fig. 8(a) shows the effect of \( K \) on the overall performance of TS-TCC and CA-TCC, where the x-axis is the percentage \( K/d \), and \( d \) is the length of the features. Clearly, increasing the percentage of the predicted future timesteps improves the performance. However, larger percentages can harm the performance as it reduces the amount of past data used for training the autoregressive model. We observe that predicting 40\% of the total feature length performs the best, and thus we set \( K \) as \( d \times 40\% \) in our experiments. The same conclusion can be drawn for both variants of our framework. Fig. 8(b) shows the results of varying \( \lambda_1 \) and \( \lambda_2 \) in TS-TCC (7) in a range between 0.001 and 1000 respectively. We first fix \( \lambda_1 = 1 \) and change the values of \( \lambda_2 \). We observe that our model achieves good performance when \( \lambda_2 \approx 1 \), where the model performs best with \( \lambda_2 = 0.7 \). Consequently, we fix \( \lambda_2 = 0.7 \) and tune the value of \( \lambda_1 \) as in Fig. 8(b), where we find that our model achieves the best performance when \( \lambda_1 = 1 \). We also find that as \( \lambda_1 < 10 \), our model is less sensitive to its value, while it is more sensitive to different values of \( \lambda_2 \).

The proper selection of augmentations is crucial for contrastive learning techniques, as contrastive methods are sensitive to the choice of augmentations [4]. Many studies showed that composing multiple data augmentation operations achieves better performance for image data [4], [43]. However, the selection

---

**Table VI**

| Component          | HAR                  | Sleep-EDF             | Epilepsy              |
|--------------------|----------------------|-----------------------|-----------------------|
| TC only            | 68.16±1.15           | 66.89±1.11            | 88.29±1.29            |
| TC + X-Aug         | 74.22±1.03           | 72.18±0.99            | 90.51±0.43            |
| TS-TCC (TC + X-Aug + CC) | 77.58±1.78 | 76.66±1.96            | 93.12±0.31            |
| CA-TCC (TC + X-Aug + SCC) | 88.27±0.38 | 88.29±0.34            | 94.52±0.12            |
| TS-TCC (Weak only) | 67.39±1.73           | 65.54±2.42            | 93.22±0.11            |
| CA-TCC (Weak only) | 85.68±0.26           | 84.77±0.25            | 91.97±0.19            |
| TS-TCC (Strong only) | 50.37±1.18          | 43.05±1.42            | 92.49±0.62            |
| CA-TCC (Strong only) | 59.59±0.06 | 53.30±0.49            | 90.60±0.20            |

We also show the effect of using two weak or two strong augmentations on their performance. It is clear that using a combination of weak and strong augmentations yields the best performance. The results are obtained with the linear evaluation experiment on 5\% of labeled data on three datasets.
Fig. 8. Sensitivity analysis experiments on HAR dataset. Figure (a) shows the effect of changing the percentage of the predicted future timesteps, where we notice a close performance among our two variants. Figure (b) shows the impact of the different combinations of $\lambda_1$ and $\lambda_2$ on TS-TCC performance. Last, figure (c) shows the effect of the different combinations of $\lambda_3$ and $\lambda_4$ on CA-TCC performance.

Fig. 9. Sample from each adopted dataset after normalization along with its weak and strong augmented views. The first row is the original samples, the second row shows the weak augmented views, and the third row shows the strong augmented views.

The proper augmentations that well-fit time-series data is less explored and still an open problem [53]. Here, we aim to study the choice of suitable augmentations for our contrastive learning problem.

We define the weak augmentation as the one that applies limited change on the shape of the original signal as shown in the second row of Fig. 9. Examples of weak augmentations include scaling and time shifting. On the other hand, strong augmentation makes strong perturbations on the signal shape with keeping some of its temporal information, such as permutation. This is shown in the third row of Fig. 9. Permutation includes splitting the signal into $M$ chunks and shuffling their order. Notably, some augmentations such as jittering (i.e., adding random noise) can be considered as both strong and weak augmentations depending on the added noise level. To justify the selection of our proposed augmentations, we provide a systematic study on the impact of applying several different data augmentations for each view in TS-TCC (Fig. 1). The results are provided in Table VII.

We first apply weak augmentation only, i.e., scaling, to one view without applying any augmentation to the other view, and the accuracy is only 46.12% as shown in Table VII. However, by composing jitter to scaling, we can observe a large performance improvement. Meanwhile, we also apply strong augmentation only and keep the other view without augmentations. The accuracy, in this case, is 62.07%, which is much higher than applying only weak augmentation. Besides, we notice that adding a jitter to the permutation can further improve the accuracy to 72.35%.

We also test to apply another weak augmentation that does not highly affect the signal characteristics along with the strong augmentation. We can find that using time shift as a weak augmentation achieves relatively good performance, i.e., an accuracy of 71.57%. Again, it can be found that adding jitter to time shift further improves the accuracy to 74.69%. Similar results are achieved when replacing time-shift with scaling. In addition, applying only weak or strong augmentations to both views results in poorer performance as shown in Table VI. At this point, we find that applying weak augmentation for one view and strong augmentations for other views achieves the best performance, i.e., an accuracy of 77.58%. As an explanation, weak augmentations apply limited changes to the shape of the original signal as shown in Fig. 9, which helps the model to perform well on the test data.

![Image](image-url)
Via observing the different characteristics of three different samples in Fig. 9 in terms of signal magnitude and sampling rates, we conclude that the choice of proper parameters for augmentations (e.g., jitter and scaling ratio) will vary from one dataset to another. Consequently, the range of parameter choices will be highly dependent on the characteristics of each time-series data. Therefore, we propose to normalize the signals as a preprocessing step in our framework to improve parameter selection. For example, the value of added jitter after normalizing the signals in weak augmentation should be less than the ones added in the strong augmentation. In the experiments, we observe that the best practice is to normalize the data between 0 and 1 and set the weak jitter to be in the range [0, 0.1] and the strong jitter in the range [0.1, 1]. Similarly, a scaling ratio of 2 would be sufficient for weak augmentation in any time-series signals.

Similarly, for strong augmentation, it is important to properly select the number of chunks $M$, where the value of $M$ in time-series data with longer sequences should be greater than its value in those with shorter sequences. We find that selecting 40% of the total feature size achieves the best performance for the three time-series datasets (see Section V-E).

VI. CONCLUSION

In this paper, we proposed two representation learning frameworks for time-series data. The first, i.e., TS-TCC, is for self-supervised learning from unlabeled data, while the second, i.e., CA-TCC is for semi-supervised learning when few labeled samples are available. In particular, we propose time-series-specific weak and strong augmentations and provide a systematic study of the choice of these augmentations. We use these augmentations to learn transformation-invariant representations through our proposed temporal and contextual contrasting modules in TS-TCC. By training a linear classifier with few labels on top of the learned representation by TS-TCC, it achieved comparable performance to the fully-supervised training. In addition, TS-TCC showed noticeable improvement under different labeling budgets, where 10% of the labeled data could achieve close performance to the supervised training with full labeled data. We extended TS-TCC to the semi-supervised settings and proposed CA-TCC, which benefits from the pseudo labels generated by the fine-tuned TS-TCC model to train a class-aware supervised contrastive loss. CA-TCC was able to further improve this performance with only 1% of labeled data when testing on different datasets. In addition, both variants were able to improve the transferability of the representations in real-world transfer learning scenarios.

REFERENCES

[1] A. Gharehbaghi and M. Lindén, “A deep machine learning method for classifying cyclic time series of biological signals using time-growing neural network,” IEEE Trans. Neural Netw. Learn. Syst., vol. 29, no. 9, pp. 4102–4115, Sep. 2018.

[2] T. Ching et al., “Opportunities and obstacles for deep learning in biology and medicine,” J. Roy. Soc. Interface, vol. 15, 2018, Art. no. 20170387.

[3] M. Noroozi and P. Favaro, “Unsupervised learning of visual representations by solving jigsaw puzzles,” in Proc. Eur. Conf. Comput. Vis., B. Leibe, J. Matas, N. Sebe, and M. Welling, eds., Cham, Switzerland: Springer, 2016, pp. 69–84.

[4] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, “A simple framework for contrastive learning of visual representations,” in Proc. 37th Int. Conf. Mach. Learn., 2020, pp. 1597–1607.

[5] S. Gidaris, P. Singh, and N. Komodakis, “Unsupervised representation learning by predicting image rotations,” in Proc. Int. Conf. Learn. Representations, 2018.

[6] R. D. Hjelm et al., “Learning deep representations by mutual information estimation and maximization,” in Proc. Int. Conf. Learn. Representations, 2019.

[7] 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., 2020, pp. 9726–9735.

[8] J.-Y. Franceschi, A. Dieuleveut, and M. Jaggi, “Unsupervised scalable representation learning for multivariate time series,” in Proc. Adv. Neural Inf. Process. Syst., 2019.

[9] M. N. Mohsenvand, R. M. Izadi, and P. Maes, “Contrastive representation learning for electroencephalogram classification,” in Proc. Mach. Learn. Health NeuralIPS Workshop, E. Alsantzer, M. B. A. McDermott, F. Falck, S. K. Sarkar, S. Roy, and S. L. Hyland, eds., 2020, pp. 238–253.

[10] Y. Cheng, H. Goh, K. Dognonou, O. Tuzel, and E. Azemi, “Subject-aware contrastive learning for biosignals,” 2020, arXiv: 2007.04871.

[11] E. Eldele et al., “Time-series representation learning via temporal and contextual contrasting,” in Proc. 13th Int. Joint Conf. Artif. Intell., 2021, pp. 2352–2359.

[12] G.-J. Qi and J. Luo, “Small data challenges in big data era: A survey of recent progress on unsupervised and semi-supervised methods,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 44, no. 4, pp. 2168–2187, Apr. 2022.

[13] L. Jing and Y. Tian, “Self-supervised visual feature learning with deep neural networks: A survey,” IEEE Trans. on Pattern Anal. Mach. Intell., vol. 43, no. 11, pp. 4037–4058, Nov. 2021.

[14] 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, 2021, Art. no. 2.

[15] R. Zhang, P. Isola, and A. A. Efros, “Colorful image colorization,” in Proc. Eur. Conf. Comput. Vis., B. Leibe, J. Matas, N. Sebe, and M. Welling, eds., Cham, Switzerland: Springer, 2016, pp. 649–666.

[16] I. Misra and L. van der Maaten, “Self-supervised learning of pretext-invariant representations,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2020, pp. 6706–6716.

[17] X. Zhai, A. Oliver, A. Kolesnikov, and L. Beyer, “S4L: Self-supervised semi-supervised learning,” in Proc. IEEE/CVF Int. Conf. Comput. Vis., 2020, pp. 1476–1485.

[18] N. Srivastava, E. Mansimov, and R. Salakhutdinov, “Unsupervised learning of video representations using LSTMs,” in Proc. 32nd Int. Conf. Int. Conf. Mach. Learn., 2015, pp. 843–852.

[19] D. Wei, Joseph J. Lim, A. Zisserman, and W. T. Freeman, “Learning and using the arrow of time,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2018, pp. 8052–8060.

[20] X. Liu, J. van de Weijer, and A. D. Bagdanov, “Exploiting unlabeled data in CNNs by self-supervised learning to rank,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 41, no. 8, pp. 1862–1878, Aug. 2019.

[21] A. van den Oord, Y. Li, and O. Vinyals, “Representation learning with contrastive predictive coding,” 2018, arXiv: 1807.03748.

[22] J.-B. Grill et al., “Bootstrap your own latent a new approach to self-supervised learning,” in Proc. 34th Int. Conf. Neural Inf. Process. Syst., Red Hook, NY, USA, 2020, pp. 21271–21284.

[23] X. Chen and K. He, “Exploring simple siamese representation learning,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2021, pp. 15750–15758.

[24] K. Sohn et al., “FixMatch: Simplifying semi-supervised learning with consistency and confidence,” in Proc. Adv. Neural Inf. Process. Syst., H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, eds., 2020, pp. 596–608.

[25] A. Saeed, T. Ozcelebi, and J. Luijkken, “Multi-task self-supervised learning for human activity detection,” in Proc. ACM Interact. Mob. Wearable Ubiquitous Technol., vol. 3, no. 2, Jun. 2019, Art. no. 61.

[26] P. Sarkar and A. Etemad, “Self-supervised ECG representation learning for emotion recognition,” IEEE Trans. Affect. Comput., vol. 13, no. 3, pp. 1541–1554, Third Quarter 2022.

[27] A. Saeed, V. Ungureanu, and B. Gfeller, “Sense and learn: Self-supervision for omnipresent sensors,” Mach. Learn. Appl., vol. 6, 2021, Art. no. 100152.
[28] K. Aggarwal, S. Joty, L. Fernandez-Luque, and J. Srivastava, “Adversarial unsupervised representation learning for activity time-series,” in Proc. AAAI Conf. Artif. Intell., 2019.

[29] H. J. Banville, O. Chehab, A. Hyvärinen, D.-A. Engemann, and A. Gramfort, “Uncovering the structure of clinical EEG signals with self-supervised learning,” J. Neural Eng., vol. 18, no. 4, 2021, Art. no. 46020.

[30] Z. Yue et al., “TS2Vec: Towards universal representation of time series,” in Proc. AAAI Conf. Artif. Intell., 2022, pp. 8900–8908.

[31] L. Yang and S. Hong, “Unsupervised time-series representation learning with iterative bilinear temporal-spectral fusion,” in Proc. 39th Int. Conf. Mach. Learn., 2022, pp. 2503–2505.

[32] X. Zhang, Z. Zhao, T. Tsiligkaridis, and M. Zotik, “Self-supervised contrastive pre-training for time series via time frequency consistency,” in Proc. Adv. Neural Inf. Process. Syst., 2022, pp. 3988–4003.

[33] 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. IEEE Int. Conf. Comput. Commun., 2021, pp. 1–10.

[34] Q. Zhu, Z. Chen, and Y. C. Soh, “A novel semisupervised deep learning method for human activity recognition,” IEEE Trans. Ind. Infrom., vol. 15, no. 7, pp. 3821–3830, Jul. 2019.

[35] M. Rezagholi-rah and Md A. Haidar, “REG-GAN: Semi-supervised learning based on generative adversarial networks for regression,” in Proc. IEEE Int. Conf. Acoust. Speech Signal Process., 2019, pp. 2806–2810.

[36] K. Maruss and K. Buza, “SUCCESS: A new approach for semi-supervised time-series,” in Proc. Artif. Intell. Soft Comput., 2013, pp. 437–447.

[37] H. Xing, Z. Xiao, D. Zhan, S. Luo, P. Dai, and K. Li, “SelfMatch: Robust semisupervised time-series classification with self-distillation,” Int. J. Intell. Syst., vol. 37, no. 11, pp. 8532–8610, 2022.

[38] H. Fan, F. Zhang, R. Wang, X. Huang, and Z. Li, “Semi-supervised time series classification by temporal relation prediction,” in Proc. IEEE Int. Conf. Acoust. Speech Signal Process., 2021, pp. 3545–3549.

[39] Z. Wang, W. Yan, and T. Oates, “Time series classification from scratch with deep neural networks: A strong baseline,” in Proc. Int. Joint Conf. Neural Netw., 2017, pp. 1578–1585.

[40] A. Vaswani et al., “Attention is all you need,” in Proc. Adv. Neural Inf. Process. Syst., I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., Curran Associates, Inc., 2017, pp. 6000–6010.

[41] Q. Wang et al., “Learning deep transformer models for machine translation,” in Proc. 57th Annu. Meet. Assoc. Comput. Linguistics, Florence, Italy, 2019, pp. 1810–1822.

[42] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” in Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Hum. Lang. Technol., Minneapolis, Minnesota, 2019, pp. 4171–4186.

[43] P. Khosla et al., “Supervised contrastive learning,” in Advances Neural Information Processing Systems, H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, Eds., Red Hook, NY, USA: Curran Associates, Inc., 2020, pp. 18661–18673.

[44] D. Anguita, A. Ghio, L. Oneto, X. Parra, and J. L. Reyes-Ortiz, “A public domain dataset for human activity recognition using smartphones,” in Proc. Euro. Symp. Artif. Neural Netw., 2013, pp. 437–442.

[45] A. L. Goldberger et al., “PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals,” Circulation, vol. 101, no. 23, pp. 215–220, 2000.

[46] A. Supratak, H. Dong, C. Wu, and Y. Guo, “DeepSleepNet: A model for automatic sleep stage scoring based on raw single-channel EEG,” IEEE Trans. Neural Syst. rehabilit. Eng., vol. 25, no. 11, pp. 1998–2008, Nov. 2017.

[47] E. Eldele et al., “An attention-based deep learning approach for sleep stage classification with single-channel EEG,” IEEE Trans. Neural Syst. Rehabil. Eng., vol. 29, pp. 809–818, 2021.

[48] R. S. Andrzejrajk, K. Léhér, F. Moomann, C. Rieke, P. David, and C. E. Elzer, “Indications of nonlinear deterministic and finite-dimensional structures in time series of brain electrical activity: Dependence on recording region and brain state,” Phys. Rev. E, vol. 64, Nov. 2001, Art. no. 061907.

[49] C. Lessmeier, J. K. Kimotho, D. Zimmer, and W. Sextro, “Condition monitoring of bearing damage in electromagnetic drive systems by using motor current signals of electric motors: A benchmark data set for data-driven classification,” in Proc. Eur. Conf. Prognostics Health Manage. Soc., 2016.

[50] M. Ragab et al., “Adversarial multiple-target domain adaptation for fault classification,” IEEE Trans. Instrum. Meas., vol. 70, 2021, Art. no. 3500211.

[51] A. Tarvainen and H. Valpola, “Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results,” in Proc. 31st Int. Conf. Neural Inf. Process. Syst., Red Hook, NY, USA, 2017, pp. 119–1204.

[52] J. Li, R. Socher, and S. C. H. Hoi, “DivideMix: Learning with noisy labels as semi-supervised learning,” in Proc. Int. Conf. Learn. Representations, 2020.

[53] Q. Wen et al., “Time series data augmentation for deep learning: A survey,” in Proc. 13th Int. Joint Conf. Artif. Intell., 2021, pp. 4653–4660.

**Emadeldeen Eldele** received the BSc and MSc degrees in computer engineering from the University of Electronic Science and Technology of China (UESTC), Chengdu, China, in 2012 and 2018, respectively, and the PhD degree from the School of Computer Science and Engineering, Nanyang Technological University (NTU), Singapore, in 2023. Currently, he is a scientist with the Centre for Frontier AI Research (CFAR), Agency for Science, Technology and Research (A*STAR), Singapore. His research interests include improving the robustness of deep learning models against challenges, e.g., domain shift, label scarcity, and adversarial attacks.

**Mohamed Ragab** received the PhD degree in computer science and engineering from Nanyang Technological University (NTU), Singapore, in 2022. He is a research scientist with both the Institute for Infocomm Research and the Center for Frontier AI Research (CFAR), A*STAR. His research emphasis is on resilient and transferable AI solutions for dynamic predictive maintenance scenarios. His research interests include self-supervised learning, transfer learning, and robustness for time series applications.

**Zhenghua Chen** (Senior Member, IEEE) received the BEng degree in mechatronics engineering from the University of Electronic Science and Technology of China (UESTC), Chengdu, China, in 2011, and the PhD degree in electrical and electronic engineering from Nanyang Technological University (NTU), Singapore, in 2017. Currently, he is a senior scientist and lab head with the Institute for Infocomm Research, and an early career investigator with Centre for Frontier AI Research (CFAR), Agency for Science, Technology and Research (A*STAR). He has won several competitive awards, such as First Place Winner for CVPR 2021 UG2+ Challenge, A*STAR Career Development Award, First Runner-Up Award for Grand Challenge at IEEE VCP 2020, Finalist Academic Paper Award at IEEE ICPHM 2020, etc. He serves as an associate editor for Elsevier Neurocomputing and IEEE Transactions on Instrumentation and Measurement. He is currently the vice chair of IEEE Sensors Council Singapore Chapter. His research interests include smart sensing, data analytics, machine learning, transfer learning and related applications.
Min Wu (Senior Member, IEEE) received the BE degree in computer science from the University of Science and Technology of China (USTC), in 2006, and the PhD degree in computer science from Nanyang Technological University (NTU), Singapore, in 2011. He is currently a principal scientist with Institute for Infocomm Research (I2R), Agency for Science, Technology and Research (A*STAR), Singapore. He received the best paper awards in IEEE ICIEA 2022, IEEE SmartCity 2022, InCoB 2016 and DASFAA 2015. He also won the CVPR UG2+ challenge in 2021 and the IICAI competition on repeated buyers prediction in 2015. His current research interests include machine learning, data mining, and bioinformatics.

Chee-Keong Kwoh received the bachelor’s degree in electrical engineering and master’s degree in industrial system engineering from the National University of Singapore, Singapore, in 1987 and 1991, respectively, and the PhD degree from the Imperial College of Science, Technology, and Medicine, University of London, in 1995. He is a researcher with the field of Data Analytics and Data mining. He has been with the School of Computer Engineering, Nanyang Technological University (NTU). His research interests include development of computational methods for analyzing and interpreting data, particularly data related to protein interactions, and drug-target interactions. He has published many articles in these areas in scientific journals and has been cited frequently by other researchers in the field. He is a member of the association for Medical and Bio-Informatics, Imperial College Alumni Association of Singapore. He was conferred the Public Service Medal by the president of Singapore in 2008.

Xiaoli Li (Senior Member, IEEE) is currently a senior principal scientist and department head with the Institute for Infocomm Research, A*STAR, Singapore. He also holds adjunct professor position with Nanyang Technological University. His research interests include AI, machine learning, data mining, and bioinformatics. He has been serving as general chairs/area chairs/senior PC members in AI and data mining related conferences. He is currently serving as editor-in-chief of World Scientific Annual Review of Artificial Intelligence, and associate editors of IEEE Transactions on Artificial Intelligence, Knowledge and Information Systems and Machine Learning with Applications. He has published more than 300 high quality papers and won 9 best paper/benchmark competition awards.

Cuntai Guan (Fellow IEEE) received the PhD degree from Southeast University, China, in 1993. He is a president’s chair professor with the School of Computer Science and Engineering, director with the Artificial Intelligence Research Institute, director with the Centre for Brain-Computing Research, and co-director of S-Lab for Advanced Intelligence, Nanyang Technological University, Singapore. His research interests include brain-computer interfaces, machine learning, medical signal and image processing, and neural and cognitive rehabilitation. He is a recipient of the Annual BCI Research Award (first prize), the King Salman Award for Disability Research, the IES Prestigious Engineering Achievement Award, the Achiever of the Year (Research) Award, and the Finalist of President Technology Award. He is also an elected Fellow of the US National Academy of Inventors (NAI), the Academy of Engineering Singapore (SAEng), and the American Institute for Medical and Biological Engineering (AIMBE).