Supervised Contrastive Few-Shot Learning for High-Frequency Time Series

Xi Chen, Cheng Ge, Ming Wang, Jin Wang

Alibaba Group
{chuyu.cx, eric.gc}@alibaba-inc.com, duchen.wm@taobao.com, jet.wangj@alipay.com

Abstract

Significant progress has been made in representation learning, especially with recent success on self-supervised contrastive learning. However, for time series with less intuitive or semantic meaning, sampling bias may be inevitably encountered in unsupervised approaches. Although supervised contrastive learning has shown superior performance by leveraging label information, it may also suffer from class collapse. In this study, we consider a realistic scenario in industry with limited annotation information available. A supervised contrastive framework is developed for high-frequency time series representation and classification, wherein a novel variant of supervised contrastive loss is proposed to include multiple augmentations while induce spread within each class. Experiments on four mainstream public datasets as well as a series of sensitivity and ablation studies demonstrate that the learned representations are effective and robust compared with the direct supervised learning and self-supervised learning, notably under the minimal few-shot situation.

Introduction

In recent years, a surgence of studies in contrastive learning have led to significant advances in the field of representation learning (Oord, Li, and Vinyals 2018; Tian, Krishnan, and Isola 2020; Chen et al. 2020a). The common idea behind contrastive learning is to bring "positive pairs" (samples considered to be similar) closer together while to separate "negative pairs" (samples considered to be different) further apart in the embedding space. The learned representations are expected to be distinguishable in downstream tasks. Such approaches have made particular success in the vision domain (He et al. 2020; Henaff 2020; Chen et al. 2020b; Li et al. 2021), where various views of image data are readily available. Contrary to that, time series, as another common data type stemming from the industrial or medical field, are highly heterogeneous. Transformations are not as easily found and sometimes may lead to adverse results (Iwana and Uchida 2021; Eldele et al. 2021).

In this study, we focus on a typical subset of time series which is commonly observed in industry, i.e., high-frequency vibration generated by rotating machinery. In many engineering fields such as petroleum, steel, aerospace etc., harsh working conditions with heavy load, continued fatigue, improper installation or inadequate maintenance often lead to various failures, some of which may cause fatal breakdown with enormous maintenance cost and even safety issues. Learning useful representations for vibration monitoring and recognition is of great importance for mechanical fault diagnosis in order to ensure stable performance and economic life-cycle management (Wang et al. 2019).

Classical approaches exert massive effort on feature engineering through diverse statistical and signal processing techniques (Smith and Randall 2015; Chen et al. 2018). Although the feature extraction can be explainable, it heavily relies on domain knowledge and engineering experience which may not be able to discover all important features. With the development of deep learning, hierarchical features can be automatically learned aiming at the final targets and significantly reduce expertise intervention (Deng and Yu 2014; Shao et al. 2018). Besides, some domain-knowledge based signal analysis methods have been integrated into multi-layer neural networks for feature representation enhancement (Verstraete et al. 2017; Jin and Chen 2021; Chen et al. 2022b). However, a large number of annotated samples are required for network training whereas in practice, faulty samples are often limited and labeling can be costly and challenging.

Fortunately, recent studies on contrastive learning have shed light on the alleviation of data annotation in time series analysis. Many works focus on the proper generation of positive and negative samples under the assumption that no label is available.

A typical approach is through negative sampling. Specifically, an unsupervised time-based criterion was developed following word2vec's intuition, in which similar time series are generated from their sampled subseries (Franceschi, Dieuleveut, and Jaggi 2019). Considering the non-stationary characteristic of time series, adjacent time intervals were treated as similar pairs while intervals temporally separated were deemed to have weaker dependencies (Deldari et al. 2021). However, the non-neighborhood random sampling strategy may result in negative samples similar to the reference in case of any accidental or periodical property of time series (Arora et al. 2019; Chuang et al. 2020), which compromises the learning effectiveness. To alleviate sampling

bias , Positive-Unlabeled learning was employed to measure the loss function, wherein different weights were assigned to unlabeled samples denoting the similar probability with the reference (Tonekaboni, Eytan, and Goldenberg 2021).

Instead of sampling from the time series itself, an alternative approach is data augmentation (Chen et al. 2020a). A temporal and contextual contrastive framework was designed with two different but correlated augmentations as weak and strong. The weak piece applied a jitter-and-scale strategy whereas the strong piece used permutation-and-jitter (Eldele et al. 2021). Particularly for electrocardiogram time series, a set of signal transformation recognition tasks with pseudo labels according to the transformation type were conducted as pretext tasks representation learning (Sarkar and Etemad 2020), which includes signal-to-noise ratio, scaling factor, stretching, etc. With a drawback summation of subseries consistency, temporal consistency as well as transformation consistency, a contextual consistency strategy was developed for positive pair selection, which is compatible with diverse time series distributions, scales and even missing values (Yue et al. 2021). Considering the rarely known invariances of time series in advance, a mixup data augmentation scheme was used in which new samples were generated through convex combinations of two samples (Wickstrøm et al. 2022). In addition to data augmentation, expert features based on domain knowledge also can be utilized (Nonnenmacher et al. 2022). Compared with the agnostic data transformations, expert knowledge might be readily available along with the time series in the specific domain.

In the field of fault diagnosis using vibration time series, limited contrastive learning frameworks are put forth, which are based on self-supervised representation learning with similar pretext tasks at the instance level as to narrow the contrast loss between augmented samples. Also, diverse data augmentation methods were employed including truncation, low-pass filtering, translation, etc. (Peng et al. 2021; Ding et al. 2022).

However, the existing contrastive-learning based studies still have certain limitations to be applied in high-frequency time series analysis.

Firstly, although various negative sampling and data augmentation strategies have been proposed without annotation, the quality of generated samples cannot always guarantee as high as directly referring to label information. In practice, a small number of labeled samples are more in line with the actual situations in industry, which can be regarded as few-shot learning (Wang et al. 2020). Few studies investigate supervised contrastive learning with limited annotations. This is the major concern of this study.

Secondly, to address two key properties related to the contrastive loss as alignment and uniformity (Wang and Isola 2020), the existing supervised contrastive learning (SupCon) ensures closeness of features from positive pairs naturally by leveraging label information (Khosla et al. 2020). Nevertheless, it might suffer from class collapse in which samples of the same category shrink into the same representation. How to properly induce spread in the representation geometry is another concern.

Thirdly, some existing contrastive learning frameworks and deep models give considerations to both instance level and temporal level with various granularities, which may not be applicable to vibration signal recognition. For instance, each preprocessed vibration piece may have intensive sampling points and is relatively stationary in a short period. Therefore, only instance level is required and the model structure can be light-weighted.

To address the above issues, we propose a representation learning framework based on supervised contrastive few short learning, which can be shorted as SCFSL. To leverage limited label information, a realistic assumption is made that only a small number of annotated samples are possessed while a large group of unlabeled vibration time series are to be classified. The proposed framework is able to learn robust feature representations from limited candidates via a novel supervised contrastive approach which is able to handle multiple positive and negative pairs as well as various augmentations. A lightweight deep network is designed as a universal encoder suitable for vibration. High accuracies covering many untrained samples can be achieved in the downstream classification tasks. Results and detailed analyses on four mainstream public datasets demonstrate that the learned representation via SCFSL are general and effective.

The main contributions of this paper are summarized as follows:

- We develop SCFSL, a supervised contrastive learning framework with limited labeled samples for vibration representation. To the best of our knowledge, this is the first work that provides a supervised representation method with a combination of contrastive learning, deep learning and few-shot learning suitable for time series with high frequency.
- A modified supervised contrastive loss is developed to balance the spread problem. This modified loss is able to consider multiple augmentations in a single batch and encourage samples from different labels to be separated while control the closeness of samples in the same class.
- A combination of Fast Fourier Transformation and a deep neural network consisted of 1-d convolutional modules is designed as the contrastive learning encoder, which is lightweight and effective.
- Experiments on public datasets of multiple vibration scenarios demonstrate that the proposed method has robust performance over the end-to-end supervised deep learning as well as self-supervised learning under small sample scenarios and has the potential of broad applicability in different industrial fields.

Methodology Development

Problem Definition

Given a set of time series $\{x_1, x_2, \ldots, x_N\}$ of N instances with M categories, the objective is to learn the representation r_i of each x_i that can best describe its characteristics. The effectiveness of the representation is evaluated by the downstream task with classification accuracy. Each input x_i denotes a piece of high-frequency signal with dimension $T \times C$, where T is the intensive sequence length

and C is the measurement channel. The representation set $\{r_1, r_2, \ldots, r_N\}$ contains representation vectors $r_i \in \mathbb{R}^K$, where K is the dimension of each representation instance.

Model Architecture

The overall architecture of the representation framework is shown in Figure 1, including three main components:

Data Preprocessing. For high-frequency time series, thousands of data points are collected within one second. Segmentation with a proper window size and zero-mean are required to obtain more samples for training while each segment has sufficient and qualified data points for feature extraction. Instead of using raw time series, Fast-Fourier Transformation (FFT) is applied to acquire frequency spectrums as the encoder input. Another important step is data augmentation. Various augmentation methods can be employed to generate multiple counterparts from different views either through some general approaches or based on certain domain knowledge.

Encoder Network. The encoder contains a 1-d convolutional deep network as backbone and an MLP network as projector. The backbone consists of four convolutional blocks including a 1-d convolutional layer, a batch normalization layer and a ReLU activation layer. The parameter setting for frequency input is recommended as follows: the channel sizes and the kernel numbers of four 1-d convolutional extractors are $16, 32, 64, 128$ and $15, 9, 7, 5$ respectively. The sizes of MaxPool and AdaptivePool are 2 and 64. The output dimension of the fully-connected layer is set to 64. The projector has two linear layers (1280-d hidden layer with a batch normalization and a ReLU and 64-d output layer). Note that the MLP projection is only used during the pretraining phase.

Supervised Contrasting. Herein, we identify global pair and class-conditional pair. Global pairs are denoted as g_pos and g_neg in Figure 1, in which original samples under the same category and their augmentations constitute the label-wise positive group while other labeled pieces and their augmentations make up the negative group. This is able to fundamentally avoid sampling bias in self-supervised contrastive learning by making full use of labeled information and their augmentations. Within each class, class-conditional pairs are denoted as c_pos and c_neg in Figure 1 as to distinguish between the original sample and its augmentations. Each sample and its augmentations become positives while the original samples in the same class be negatives. This helps induce spread in the representation geometry which can prevent class collapse and increase robustness.

Modified Supervised Contrastive Loss

The family of contrastive loss generates from self-supervised learning involving single positive and negative pairs in the early phase, e.g., traditional contrastive loss (Chopra, Hadsell, and LeCun 2005) and Triplet loss (Schroff, Kalenichenko, and Philbin 2015) originally used for face recognition as well as its derivatives (Che et al. 2017; Rambhatla, Che, and Liu 2022) used for time series analysis. As many studies have shown that the contrastive performance can be improved with increasing number of negatives (Chen et al. 2020a; He et al. 2020; Henaff 2020), recent trend of contrastive training is to include multiple positive and negative pairs in one batch and the InfoNCE loss as well as its derivatives are widely applied (Oord, Li, and Vinyals 2018; Tian, Krishnan, and Isola 2020).

Considering the adaptation for supervised domain, a supervised contrastive loss (SupCon) was developed to leverage label information (Khosla et al. 2020). Nevertheless, it may suffer from class collapse, wherein representations are difficult to distinguish fine-grained details within each class, sometimes leading to poor transfer ability and robustness. Modifications to SupCon have shown empirical promise through a comprehensive study in transfer capability exploration (Islam et al. 2021). Also, an insightful research into the proper degree of spread as well as the permutation invariance was conducted, in which a modified supervised contrastive loss was proved effective by adding a weighted class-conditional InfoNCE (Chen et al. 2022a). However, the first term of the modified loss does not take augmentations into consideration and the overall formulation cannot handle samples with multiple augmentations.

Herein, we extend previous researches by developing a modified supervised contrastive loss which can address various augmentations. Let $r_i = Encoder(f(x_i)) \in \mathbb{R}^K$ be the i^{th} representation anchor of the preprocessed data $f(x_i)$ in a multi-viewed batch of data B. r_i^p represents its positive counterparts including potential augmentations whereas r_i^{pa} denotes its augmentations specifically. r_i^a represents all the rest sample representations where $a \in A(i) \equiv B \backslash i$. Then, the similarity calculations between r_i and r_i^p. r_i^{pa}, r_i^a are expressed respectively as

$$z_i^p = Sim(r_i, r_i^p)/\tau \tag{1}$$

$$z_i^a = Sim(r_i, r_i^a)/\tau \tag{2}$$

$$z_i^{pa} = Sim(r_i, r_i^{pa})/\tau \tag{3}$$

where τ is a scalar temperature parameter for tuning how concentrated the features are in the representation space. $Sim(\cdot)$ denotes the similarity calculation which can be inner product, Euclidean distance or cosine similarity.

Let B_{aug} be the batch of representations with multiple augmentations. For a weight factor $\alpha \in [0, 1]$, the overall loss function is expressed as follows:

$$\mathcal{L}^{sup} = \frac{1}{|B_{aug}|} \sum_{i \in B_{aug}} \alpha \ell_i^{sup} + \frac{1}{|B|} \sum_{i \in B} (1 - \alpha) \ell_i^{aug} \tag{4}$$

The first term is the supervised contrastive loss addressing multiple number of positives:

$$\ell_i^{sup} = -\frac{1}{|P(i)|} \sum_{p \in P(i)} \log \frac{e^{z_i^p}}{\sum_{a \in A(i)} e^{z_i^a}} \tag{5}$$

where $P(i) \equiv \{p \in A_{aug}(i) : y_p = y_i\}$, $A_{aug}(i) \equiv B_{aug} \backslash i$. $|P(i)|$ is the cardinality.

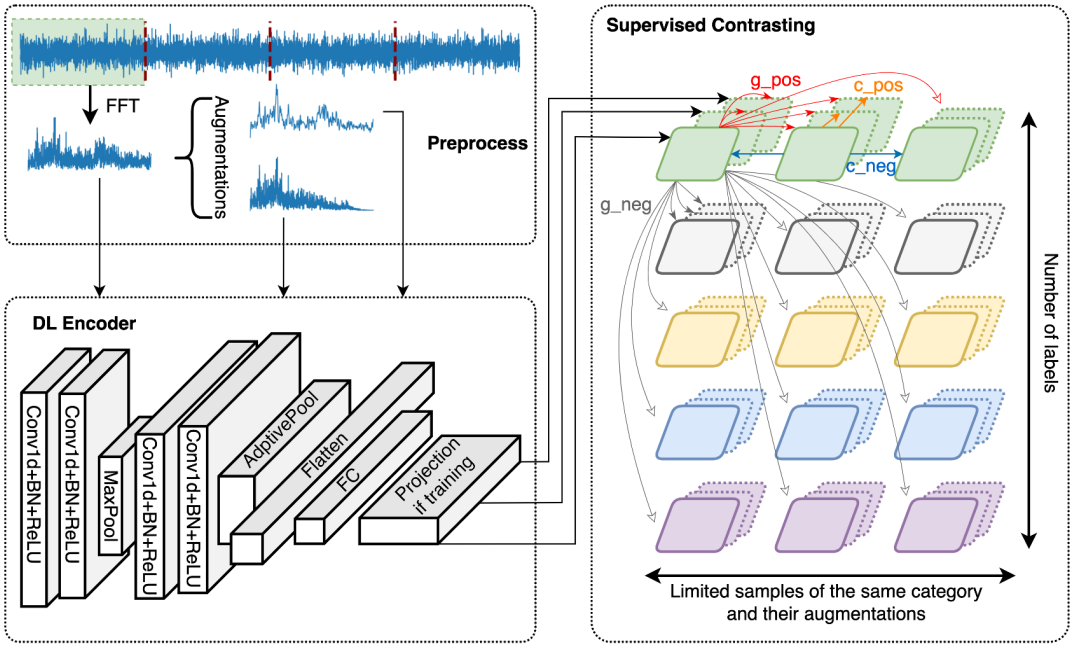

Figure 1: The proposed architecture of SCFSL.

The second term is a class-conditional version of the InfoNCE loss enabling multiple number of augmentations:

$$\ell_i^{aug} = -\frac{1}{|P_{aug}(i)|} \sum_{pa \in P_{aug}(i)} \log \frac{e^{z_i^{pa}}}{\sum_{p \in P(i)} e^{z_i^p}} \quad (6)$$

where $P_{aug}(i) \equiv \{pa \in P(i) \backslash i\}$. $|P_{aug}(i)|$ is the cardinality. Inspired by (Khosla et al. 2020), the summation over positives is located outside of the $log(\cdot)$ operation.

The first part of the loss function treats the original representations and their augmentations equally. This guarantees all positives in the augmented multi-viewed batch contribute to the numerator for any anchor, encouraging the encoder to learn closely aligned representations to all entries from the same class.

The second term focuses on augmentations within each class. Positives in the numerator only contains the anchor's augmentations while the denominator consist of all representations from the same class, intuitively encouraging augmentations derived from the same anchor to be close while different original representations in the same class to be spread apart.

Downstream Classification

After supervised contrastive training, an embedded feature space is expected to be learned in which representations of the same class are clustered close together whereas those of different labels stay far apart. For downstream tasks, two kinds of methods are employed intuitively. The first category is based on similarity measurement. For a test sample after encoding, we calculate its similarity with all learned representations. The label corresponding to the sample with the highest similarity is recorded. In this study, dot product is used as the similarity measurement in both contrastive learning and downstream classification with L2 normalization performed ahead in the feature dimension. The second category is based on classical machine learning models including Logistic Regression (LR), Support Vector Machine (SVM) with a grid search for hyper-parameters optimization. The maximum value among the three methods are taken as the final result.

Experiments and Analysis

Datasets and Experimental Setup

One of typical applications for high-frequency time series is fault diagnosis of rotating machinery. Four mainstream public datasets are utilized with a brief introduction as follows:

- **MFPT** This dataset is provided by Society of Machinery Failure Prevention Technology (MFPT 2018). Two kinds of bearing damage with 48.828 kHz sampling frequency under various working loads are considered. Normal bearing data sampled at 97.656 kHz are used with down-sampling.
- **PU** The bearing dataset is from Paderborn University (PU 2016), including up to 32 sets of current and vibration signals with heath state, artificial damage as well as real damage under four working conditions. Herein, real damaged bearings and fault modes under the working condition of N15_M07_F10 are utilized.

Datasets	Frequency	Sample No.	Sample Length	Categories
MFPT	48.8kHz	1137	2048	15
PU	16kHz	1625	2048	13
CWRU	12kHz	650	2048	10
SEU	12kHz	10220	2048	20

Table 1: Datasets overview.

- **CWRU** This is one of the most widely used datasets provided by Case Western Reserve University Bearing Data Center containing bearings with single-point damage under four motor loads (CWRU 2015). Data collected from the drive end with 12 kHz sampling frequency is used.
- **SEU** This dataset has both gearbox and bearing vibrations provided by Southeast University (SEU 2018). Both were generated on drivetrain dynamic simulator under two working conditions with rotating speed–load configuration setup. The second row of each vibration is used here for classification.

Time series segmentation and zero-mean is conducted firstly and an overview of the prepared datasets is shown in Table 1. Considering training uncertainties, all experiments are repeated ten times on a MacBook Pro with 2.4GHz processor and 16GB RAM. The average values and the standard deviations are recorded.

It should be noted that in addition to the supervised contrasting module, the encoder network also has a decisive influence on the final effect. Therefore, we conduct experiments in the end-to-end supervised approach by attaching a Softmax for direct classification with cross-entropy as the loss function and prove that the combination of FFT and the encoder network is able to achieve sound performance on all selected datasets. Detailed fault modes and working conditions are consistent with (Zhao et al. 2020) for fair comparison.

Few-Shot Learning

In practice, acquiring enough labeled data can be costly and challenging. Herein, we mainly consider the few-shot learning scenario and make the small sample assumption as follows:

Suppose for each class, there are only a limited number of annotated samples available as N_{sample} for training while a large number of unlabeled data are to be classified. Firstly, a representation model are learned via pretraining on the limited training set. Secondly, downstream classification tasks are conducted based on representations of both training and testing samples. In terms of the datasets preparation, each dataset is evenly split into half-half. A small number of samples in the training half are assumed available while the entire testing half are used for evaluation.

Experiments are conducted under two situations where $N_{sample} = 4$ and $N_{sample} = 8$. For each batch, the number of samples with the same label is denoted as N_{view}. To maintain the same number of batches in two situations, N_{view} is set to 2 and 4, respectively. The influence of different N_{sample} and N_{view} values on final performance is further conducted in Sensitivity Analysis. Also, we employ two other deep methods for comparison. The first is direct supervised learning (DSL) based on the encoder network using cross-entropy while the second is self-supervised learning (SSL) based on TS2Vec (Yue et al. 2021) which has been shown to outperform many counterparts including T-Loss (Franceschi, Dieuleveut, and Jaggi 2019), TS-TCC (Eldele et al. 2021), TST (Zerveas et al. 2021) and TNC (Tonekaboni, Eytan, and Goldenberg 2021), etc.

Due to the very limited number of training samples, the training set is taken as a whole without further splitting into train and validation. The stop criterion of the training process is that, if any of the following two circumstances occurs more than five times, training is stopped. (1) The loss decreases less than 0.0001. (2) The loss starts to increase. Results on all the four datasets are shown in Table 2 and we can observe that:

- In the left situation where each class has only 4 labeled samples, our proposed method SCL exhibits significant generality and stability over the end-to-end DSL as well as SSL with higher accuracies and lower standard deviations on all datasets. Although DSL can reach 100% during training, over-fitting is more likely to occur due to the very limited training set. This is a great advantage of SCL over traditional DSL under the few-shot scenario. SSL also shows superior performance compared with DSL without referring to any label information, but provided that a small number of label is available, SCL stands out.
- By comparing the two situations, increasing the number of training samples directly improves the performance by all three methods, especially for DSL on some of the datasets. It can be expected that with more annotated samples included, the effects will be further improved. Related experiments can be found in Sensitivity Analysis.

Enhancement with Augmentations

Different from image augmentation, in which many physically-meaningful techniques can be applied, e.g., random cropping, resize, color distortion, Gaussian blur, horizontal flip, etc. Time series augmentation is less intuitive and should be carefully treated in case the inherent property is compromised. Vibration signals generated by rotating machinery is often multi-mode aliasing and noisy. It is difficult to foresee the potential effective modes so as to control the scale of certain augmentation. For example, too much Gaussian noise may overwhelm the signal itself while too much denoising might erase certain useful frequency components. Cropping and masking can be effective in many general time series with low frequency but in our case using frequency spectrum from vibration signals, unpredictable problems might be induced due to semantic meaning changing, which can be even worse by reversing or time warping. Herein, we evaluate four augmentation methods with different parameters and one domain-knowledge based method using wavelet decomposition and reconstruction for sub-signals at different frequency bands. A piece of

Datasets	Number of samples per label = 4, $N_{view} = 2$				Number of samples per label = 8, $N_{view} = 4$			
	Train/Test	DSL	SSL	SCL	Train/Test	DSL	SSL	SCL
MFPT	15x4/569	0.169±0.042	0.737±0.024	**0.938±0.008**	15x8/569	0.954±0.014	0.766±0.024	**0.978±0.007**
PU	13x4/813	0.150±0.090	0.778±0.007	**0.929±0.019**	13x8/813	0.913±0.027	0.805±0.006	**0.965±0.003**
CWRU	10x4/325	0.137±0.072	0.922±0.033	**0.997±0.002**	10x8/325	0.319±0.175	0.957±0.014	**1.0±0.0**
SEU	20x4/5110	0.319±0.227	0.896±0.013	**0.988±0.003**	20x8/5110	0.410±0.229	0.912±0.011	**1.0±0.0**

Table 2: Classification accuracies comparison between DSL, SSL and SCL on four datasets.

Datasets	Number of samples per label = 4 x 3 Each with 2 augmentations, N_{view} = 6			
	Train/Test	DSL	SSL	SCL
MFPT	15x12/569	0.935±0.009	0.716±0.015	**0.947±0.006**
PU	13x12/813	0.862±0.017	0.745±0.029	**0.960±0.010**
CWRU	10x12/325	0.996±0.002	0.948±0.016	**0.999±0.001**
SEU	20x12/5110	0.991±0.001	0.851±0.012	**0.994±0.002**

Table 3: Classification accuracies with multiple data augmentations on four datasets.

time series, its FFT as well as various examples of each augmentation are shown in Figure 2 and the single augmentation results are detailed in Table 4.

We select the top 2 augmentation methods to enhance the few-shot learning performance in which the number of samples per label is increased from 4 to 12, and N_{view} becomes 6. Results are presented in Table 3. Compared with the left part of Table 2, improvements can be observed on all datasets, particularly for PU, demonstrating the effectiveness of the enhancement with multiple augmentations. Nevertheless, it should be acknowledged that the accuracies are still lower compared with the situation including more real labeled samples as shown in the right part of Table 2. Theoretically, each sample can be augmented many times, but our experiments empirically reveal that more augmentations do not always guarantee better results in different tasks as shown in Figure 5.

A visualization of the learned representations of the training set under varying weight factor α is shown in Figure 3. We can observe that a smaller α induces more spread within class, which is consistent with the loss function in Equation 4 . Quantitative analysis to balance class collapse and over-uniform is conducted in the following section.

Sensitivity Analysis

Since the classification accuracies of CWRU and SEU datasets are already high, sensitivity analysis is performed on MFPT and PU datasets to investigate the influence of four factors including the temperature τ for tuning the representation concentration, the weight factor α in the loss function to control spread, the number of samples per label N_{sample} as well as the number of samples from the same class per batch N_{view}. Note that the experiments on N_{view} are conducted under the situation where $N_{sample} = 16$.

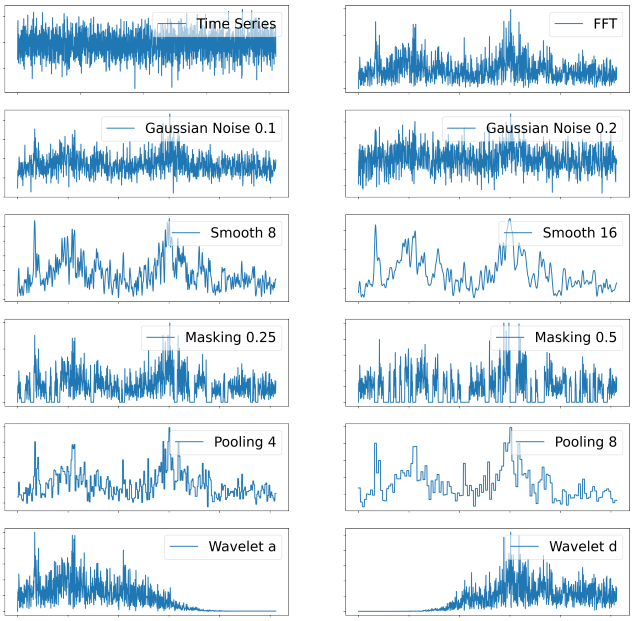

Figure 2: Various augmentation methods on MFPT dataset. The first row are the piece of time series and its frequency spectrum. The second row includes two Gaussian noise added with scale=0.1 and 0.2. The third row uses smoothing by a hanning window with size=8 and 16. Random masking ranges between size 8-16 with 0.25 and 0.5 probabilities as well as Maximum pooling with size=4 and 8 are plotted in the next two rows. The last row depicts the 1^{st} level subsignals after wavelet decomposition (db8) and reconstruction.

We can find from Figure 4 that the sensitivity analysis on τ and α is able to provide the optimum degree of concentration and spread for each task. With increasing annotated N_{sample} available, the performance improves. Promisingly, even under the extreme situation where $N_{sample} = 2$, accuracies still keep relatively high as shown in the left-bottom picture. Lastly, classification results are less sensitive to N_{view}.

Furthermore, sensitivities considering the number of augmentations per sample under different weight factors are shown in Figure 5. It is observed that for both datasets, increasing augmentations may compromise the performance.

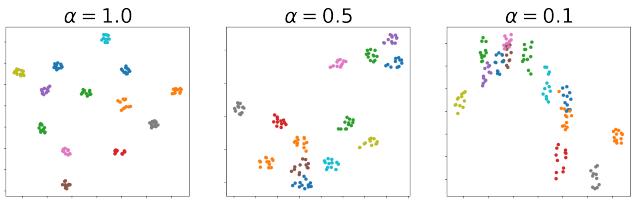

Figure 3: Spread balance on learned representations with different α.

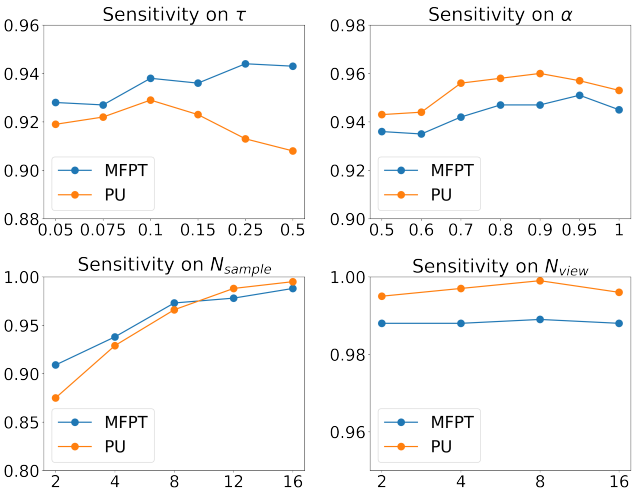

Figure 4: Sensitivity analysis on four influencing factors.

Ablation Study

To verify the efficacy of the proposed components in SCFSL, a comparison between the full method and its four variants on MFPT dataset is presented in Table 4, in which (1) w/o Projection Layer removes the MLP projector, (2) w/o Spread Loss removes the second spread term from the loss function, (3) w/o Augmentation performs contrastive learning based on the original samples without any augmentation, (4) w/o FFT uses the high-frequency time series directly as the encoder input. The results show that all the components of SCFSL take effect.

We also conduct comparisons among various augmentation methods in which **Smooth 16** and **Pooling 4** are selected for performance enhancement as discussed in the above section.

Further, to evaluate our designed backbone, we replace the encoder network with three SOTA models for time series classification developed in recent years, including FCN (Wang, Yan, and Oates 2017), XResNet1d18 (He et al. 2019) and InceptionTime (Ismail Fawaz et al. 2020). Compared with their model sizes, our proposed backbone is relatively lightweight with only **4.66M**. Besides, the classification accuracy decreases significantly in all replaced cases, demonstrating our network design is effective in addressing high-frequency time series.

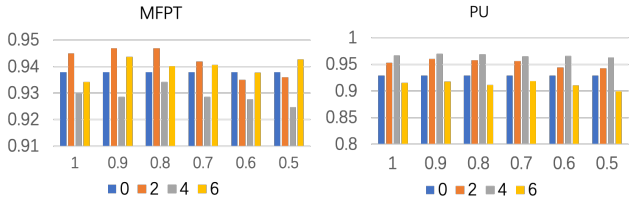

Figure 5: Performance against the number of augmentations per sample [0,2,4,6] under different weight factors [0.5-1].

		Avg. Accuracy
SCFSL		**0.951**
w/o Projection Layer		0.931 (-2.1%)
w/o Spread Loss		0.945 (-0.7%)
w/o Augmentations		0.938 (-1.4%)
w/o FFT		0.517 (-45.6%)
Single Augmentation		
+ Gaussian 0.1		0.937 (-1.5%)
+ Gaussian 0.2		0.932 (-2.0%)
+ Smooth 8		0.937 (-1.5%)
+ Smooth 16		0.943 (-0.8%)
+ Masking 0.25		0.940 (-1.2%)
+ Masking 0.5		0.938 (-1.4%)
+ Pooling 4		0.944 (-0.7%)
+ Pooling 8		0.942 (-0.9%)
+ Wavelet a		0.938 (-1.4%)
+ Wavelet d		0.918 (-3.5%)
Encoder Network	Size	
→ FCN	9.48M	0.617 (-35.1%)
→ XResNet1d18	21.58M	0.638 (-32.9%)
→ InceptionTime	13.86M	0.729 (-23.3%)

Table 4: Ablation results on MFPT dataset.

Conclusion

This paper proposes a supervised representation learning framework addressing high-frequency time series classification, namely SCFSL, to leverage limited annotation information. A novel supervised contrastive loss is developed which is able to include multiple augmentations and induce spread within each class. The evaluation of the learned representations on four public fault diagnostic tasks under small sample scenarios demonstrates the superior performance in both accuracy and stability of SCFSL over the direct supervised learning as well as self-supervised learning. Various data augmentation methods and parameter sensitivities are investigated which further improve the final performance. Ablation study proves the indispensability of the proposed components.

Last but not least, although this paper focuses on high-frequency time series, the proposed supervised contrastive learning framework can be easily applied for more general time series representation at the instance level by replacing the encoder with suitable backbones.

References

Arora, S.; Khandeparkar, H.; Khodak, M.; Plevrakis, O.; and Saunshi, N. 2019. A theoretical analysis of contrastive unsupervised representation learning. *arXiv preprint arXiv:1902.09229*.

Che, Z.; He, X.; Xu, K.; and Liu, Y. 2017. DECADE: a deep metric learning model for multivariate time series. In *KDD workshop on mining and learning from time series*. sn.

Chen, M.; Fu, D. Y.; Narayan, A.; Zhang, M.; Song, Z.; Fatahalian, K.; and Ré, C. 2022a. Perfectly Balanced: Improving Transfer and Robustness of Supervised Contrastive Learning. In *International Conference on Machine Learning*, 3090–3122. PMLR.

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. 2020a. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, 1597–1607. PMLR.

Chen, T.; Kornblith, S.; Swersky, K.; Norouzi, M.; and Hinton, G. E. 2020b. Big self-supervised models are strong semi-supervised learners. *Advances in neural information processing systems*, 33: 22243–22255.

Chen, X.; Ge, C.; Wang, M.; and Wang, J. 2022b. An Integration of Spectrum Analysis and Attention-based Network for Condition Monitoring of Vibration Components. In *2022 IEEE International Conference on Prognostics and Health Management (ICPHM)*, 108–113. IEEE.

Chen, X.; Kopsaftopoulos, F.; Wu, Q.; Ren, H.; and Chang, F.-K. 2018. Flight state identification of a self-sensing wing via an improved feature selection method and machine learning approaches. *Sensors*, 18(5): 1379.

Chopra, S.; Hadsell, R.; and LeCun, Y. 2005. Learning a similarity metric discriminatively, with application to face verification. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, 539–546. IEEE.

Chuang, C.-Y.; Robinson, J.; Yen-Chen, L.; Torralba, A.; and Jegelka, S. 2020. Debiased contrastive learning. *arXiv preprint arXiv:2007.00224*.

CWRU. 2015. Case Western Reserve University (CWRU) Bearing Data Center. [EB/OL]. https://csegroups.case.edu/bearingdatacenter/ Accessed on December 10, 2021.

Deldari, S.; Smith, D. V.; Xue, H.; and Salim, F. D. 2021. Time Series Change Point Detection with Self-Supervised Contrastive Predictive Coding. In *Proceedings of the Web Conference 2021*, 3124–3135.

Deng, L.; and Yu, D. 2014. Deep learning: methods and applications. *Foundations and trends in signal processing*, 7(3–4): 197–387.

Ding, Y.; Zhuang, J.; Ding, P.; and Jia, M. 2022. Self-supervised pretraining via contrast learning for intelligent incipient fault detection of bearings. *Reliability Engineering & System Safety*, 218: 108126.

Eldele, E.; Ragab, M.; Chen, Z.; Wu, M.; Kwoh, C. K.; Li, X.; and Guan, C. 2021. Time-series representation learning via temporal and contextual contrasting. *arXiv preprint arXiv:2106.14112*.

Franceschi, J.-Y.; Dieuleveut, A.; and Jaggi, M. 2019. Unsupervised scalable representation learning for multivariate time series. *arXiv preprint arXiv:1901.10738*.

He, K.; Fan, H.; Wu, Y.; Xie, S.; and Girshick, R. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9729–9738.

He, T.; Zhang, Z.; Zhang, H.; Zhang, Z.; Xie, J.; and Li, M. 2019. Bag of tricks for image classification with convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 558–567.

Henaff, O. 2020. Data-efficient image recognition with contrastive predictive coding. In *International Conference on Machine Learning*, 4182–4192. PMLR.

Islam, A.; Chen, C.-F. R.; Panda, R.; Karlinsky, L.; Radke, R.; and Feris, R. 2021. A broad study on the transferability of visual representations with contrastive learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 8845–8855.

Ismail Fawaz, H.; Lucas, B.; Forestier, G.; Pelletier, C.; Schmidt, D. F.; Weber, J.; Webb, G. I.; Idoumghar, L.; Muller, P.-A.; and Petitjean, F. 2020. Inceptiontime: Finding alexnet for time series classification. *Data Mining and Knowledge Discovery*, 34(6): 1936–1962.

Iwana, B. K.; and Uchida, S. 2021. An empirical survey of data augmentation for time series classification with neural networks. *Plos one*, 16(7): e0254841.

Jin, C.-c.; and Chen, X. 2021. An end-to-end framework combining time–frequency expert knowledge and modified transformer networks for vibration signal classification. *Expert Systems with Applications*, 171: 114570.

Khosla, P.; Teterwak, P.; Wang, C.; Sarna, A.; Tian, Y.; Isola, P.; Maschinot, A.; Liu, C.; and Krishnan, D. 2020. Supervised contrastive learning. *arXiv preprint arXiv:2004.11362*.

Li, C.; Yang, J.; Zhang, P.; Gao, M.; Xiao, B.; Dai, X.; Yuan, L.; and Gao, J. 2021. Efficient self-supervised vision transformers for representation learning. *arXiv preprint arXiv:2106.09785*.

MFPT. 2018. Society For Machinery Failure Prevention Technology. [EB/OL]. https://mfpt.org/fault-data-sets/ Accessed on December 10, 2021.

Nonnenmacher, M. T.; Oldenburg, L.; Steinwart, I.; and Reeb, D. 2022. Utilizing Expert Features for Contrastive Learning of Time-Series Representations. In *International Conference on Machine Learning*, 16969–16989. PMLR.

Oord, A. v. d.; Li, Y.; and Vinyals, O. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Peng, T.; Shen, C.; Sun, S.; and Wang, D. 2021. Fault Feature Extractor based on Bootstrap Your Own Latent and Data Augmentation Algorithm for Unlabeled Vibration Signals. *IEEE Transactions on Industrial Electronics*.

PU. 2016. KAt-DataCenter, Chair of Design and Drive Technology, Paderborn University. [EB/OL]. https://mb.uni-

paderborn.de/kat/forschung/datacenter/bearing-datacenter/ Accessed on December 10, 2021.

Rambhatla, S.; Che, Z.; and Liu, Y. 2022. I-SEA: Importance Sampling and Expected Alignment-based Deep Distance Metric Learning for Time Series Analysis and Embedding. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(7): 8045–8053.

Sarkar, P.; and Etemad, A. 2020. Self-supervised ECG representation learning for emotion recognition. *IEEE Transactions on Affective Computing*.

Schroff, F.; Kalenichenko, D.; and Philbin, J. 2015. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 815–823.

SEU. 2018. Southeast University Gearbox Datasets. [EB/OL]. https://github.com/cathysiyu/Mechanical-datasets/ Accessed on December 10, 2021.

Shao, H.; Jiang, H.; Lin, Y.; and Li, X. 2018. A novel method for intelligent fault diagnosis of rolling bearings using ensemble deep auto-encoders. *Mechanical Systems and Signal Processing*, 102: 278–297.

Smith, W. A.; and Randall, R. B. 2015. Rolling element bearing diagnostics using the Case Western Reserve University data: A benchmark study. *Mechanical systems and signal processing*, 64: 100–131.

Tian, Y.; Krishnan, D.; and Isola, P. 2020. Contrastive multiview coding. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*, 776–794. Springer.

Tonekaboni, S.; Eytan, D.; and Goldenberg, A. 2021. Unsupervised representation learning for time series with temporal neighborhood coding. *arXiv preprint arXiv:2106.00750*.

Verstraete, D.; Ferrada, A.; Droguett, E. L.; Meruane, V.; and Modarres, M. 2017. Deep learning enabled fault diagnosis using time-frequency image analysis of rolling element bearings. *Shock and Vibration*, 2017.

Wang, T.; Han, Q.; Chu, F.; and Feng, Z. 2019. Vibration based condition monitoring and fault diagnosis of wind turbine planetary gearbox: A review. *Mechanical Systems and Signal Processing*, 126: 662–685.

Wang, T.; and Isola, P. 2020. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International Conference on Machine Learning*, 9929–9939. PMLR.

Wang, Y.; Yao, Q.; Kwok, J. T.; and Ni, L. M. 2020. Generalizing from a few examples: A survey on few-shot learning. *ACM Computing Surveys (CSUR)*, 53(3): 1–34.

Wang, Z.; Yan, W.; and Oates, T. 2017. Time series classification from scratch with deep neural networks: A strong baseline. In *2017 International joint conference on neural networks (IJCNN)*, 1578–1585. IEEE.

Wickstrøm, K.; Kampffmeyer, M.; Mikalsen, K. Ø.; and Jenssen, R. 2022. Mixing up contrastive learning: Self-supervised representation learning for time series. *Pattern Recognition Letters*, 155: 54–61.

Yue, Z.; Wang, Y.; Duan, J.; Yang, T.; Huang, C.; Tong, Y.; and Xu, B. 2021. TS2Vec: Towards Universal Representation of Time Series. *arXiv preprint arXiv:2106.10466*.

Zerveas, G.; Jayaraman, S.; Patel, D.; Bhamidipaty, A.; and Eickhoff, C. 2021. A transformer-based framework for multivariate time series representation learning. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, 2114–2124.

Zhao, Z.; Li, T.; Wu, J.; Sun, C.; Wang, S.; Yan, R.; and Chen, X. 2020. Deep learning algorithms for rotating machinery intelligent diagnosis: An open source benchmark study. *ISA transactions*, 107: 224–255.