Self-supervised Representation Learning are Effective for Out-of-Distribution Multivariant Time Series Classification

Yuyuan Feng , **Dayuan Huang** , **Yueyang Liu** , **Zhongnan Cai** , **Yican Mo** and **Yang Lu**[*]
Xiamen University
{ehappymonkey}@outlook.com,

Abstract

Multivariant Time series classification is an important problem that has great impact on traffic, energy system and et al. In the real world however, time series data is often spatial or temporal nonstationary. i.e. the distribution changes spatially or temporally. Nowadays, it remains challenging for machine learning techniques to build models for generalization to unseen distributions. Self-supervised representation learning has been widely acknowledged in the field of computer vision to obtain robust feature selector that can be used in downstream tasks. However, due to the relatively lack of effective self-supervised representative learning methods, the field of time series classification has not yet been benefited from it. In this paper, we use a two stage separate training strategy to learn a contrastive learning based encoder and a normal decoder. Empirically, our simple method improves generalization on a time series benchmark for distribution shifts.

1 Introduction

Time series classification is one of the most challenging problems in the machine learning and statistics community [Ismail Fawaz *et al.*, 2019], [Du *et al.*, 2021]. One important nature of time series is the non-stationary property, indicating that its statistical features are changing spatially or temporally. For example, traffic or weather time series at different locations, biological time series on different persons, and time series even change with time. For years, there have been tremendous efforts for time series classification, such as hidden Markov models [Fulcher and Jones, 2014], RNN-based methods [Hewamalage *et al.*, 2021], and Transformer-based approaches [Li *et al.*, 2020].

We propose to model time series from the distribution perspective to handle its dynamically changing features; more precisely, to learn robust representations for time series that generalizes to unseen distributions. The general Out-of-Distribution/domain generalization problem has been extensively studied [Wang *et al.*, 2022], [Krueger *et al.*, 2021]., where the key is to bridge the gap between known and unknown distributions. Despite existing efforts, OOD in time series remains less studied and more challenging. Compared to image classification, the dynamic distribution of time series data keeps changing over time, containing diverse distribution information that should be harnessed for better generalization.

In this paper, we use a contrastive learning framework as encoder to learn robust features based on self-supervision and a normal decoder. For the encoder, we used self-supervised pre-training in time series by modeling Time-Frequency Consistency (TF-C) [Zhang *et al.*, 2022b]. TF-C specifies that time-based and frequency-based representations, learned from the same time series sample, should be closer to each other in the time-frequency space than representations of different time series samples. Specifically, we adopt contrastive learning in time-space to generate a time-based representation. In parallel, we propose a set of novel augmentations based on the characteristic of the frequency spectrum and produce a frequency-based embedding. TF-C is designed to be invariant to different time-series datasets, which can produce generalizable features. For the decoder, we simply use a normal linear classifier.

Ther are two stages of our framework. Firstly, we pre-train the contrasive encoder based on TF-C to learn robust feature representations across different distribution shifts. After that, we froze this encoder and train the classifier using the same training datasets but with labels to take full advantage of the supervised information. Empirically, our simple method improves generalization on a time series benchmark for distribution shifts. Theoretically, we see this improvement as a bias-variance trade-off. The end-to-end training fails to adapt domain shifts because its supervised training is completely based on biased training data that do not represent the new out-of-domain distribution. The other extreme is not to use labeling at all but only representation learning; this is also undesirable because the completely unsupervised learning will be equal to clustering, boosting the bar dramatically.

In summary, our contributions are as follows:

- Novel problem: We propose to tackle the domain generalization in time series problem, which is more chal-

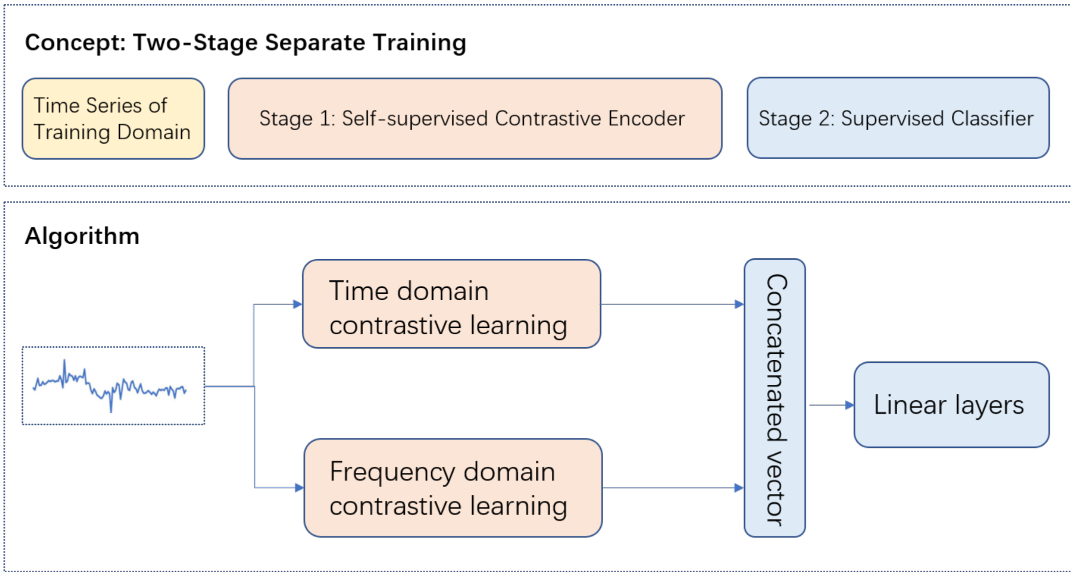

Figure 1: The framework of our two-stage separate training method. The different color of stage 1 and stage 2 means that they are trained separately and have no relation in principle.

lenging than the image classification due to the the both spatial and temporal distribution shifts in time series.

- Effective method: We use a two-stage training strategy to learn self-supervised representations while keeping the supervised information. Although the paradigm of linear probing is widely applied in computer vision, this is the first time used to tackle the out-of-distribution time series problem.
- Better performance: Empirically, our simple method improves generalization 3 points on a time series benchmark for distribution shifts.

2 Related Work

2.1 Time series classification

Time series classification is a challenging problem. Researches mainly focus on temporal relation modeling via specially-designed methods, RNN-based networks [Hewamalage *et al.*, 2021], or Transformer architecture [Li *et al.*, 2020]. To our best knowledge, there is only one recent work [Du *et al.*, 2021] that studied time series from the distribution level. However, AdaRNN is a two-stage non-differential method that is tailored for RNN.

2.2 Domain Generalization

Domain / OOD generalization [Wang *et al.*, 2022] typically assumes the availablality of domain labels for training. Specifically, [Matsuura and Harada, 2019] also studied DG without domain labels by clustering with the style features for images, which is not applied to time series and is not end-to-end trainable. Disentanglement [Peng *et al.*, 2019], [Zhang *et al.*, 2022a] tries to disentangle the domain and label information, but they also assume access to domain information. In summary, the methodology of using representation learning to tackle domain generalization problem remains undiscovered.

2.3 Self supervised learning for time series

Although there are studies on self-supervised representation learning for time series [Rebjock *et al.*, 2021], [Sarkar and Etemad, 2020] and self-supervised pre-training for images [Chen *et al.*, 2020a], [Chen *et al.*, 2020b], all previous work has been focus on fine-tuning to adapt to downstream tasks. It seems to be an undiscovered area to take full advantage of representation learning for domain generalization. [Shi *et al.*, 2021] developed the only model to date that is explicitly designed for self-supervised time series pre-training. The model captures the local and global temporal pattern, but it is not convincing why the designed pretext task can capture generalizable representations. Although several studies applied transfer learning in the context of time series [Rebjock *et al.*, 2021], [Sarkar and Etemad, 2020], there is no foundation yet of which conceptual properties are most suitable for pre-training on time series and why.

3 Methods

In this section, we present the architecture of the two stage separate training strategy, self-supervised contrastive encoder F, linear classifier and implementation details.

3.1 Contrastive Encoder

Time-based Contrasive Encoder: For a given multivariant time series x_i, an data augmentation set X_i^T is established through a time-based augmentation bank, which includes jittering, scaling, time-shifts, and neighborhood segments, all well-established in contrastive learning [Kiyasseh *et al.*, 2021]. For each x_i and augmented sample $\widetilde{\boldsymbol{x}}_i^{\mathrm{T}} \in \mathcal{X}_i^{\mathrm{T}}$,

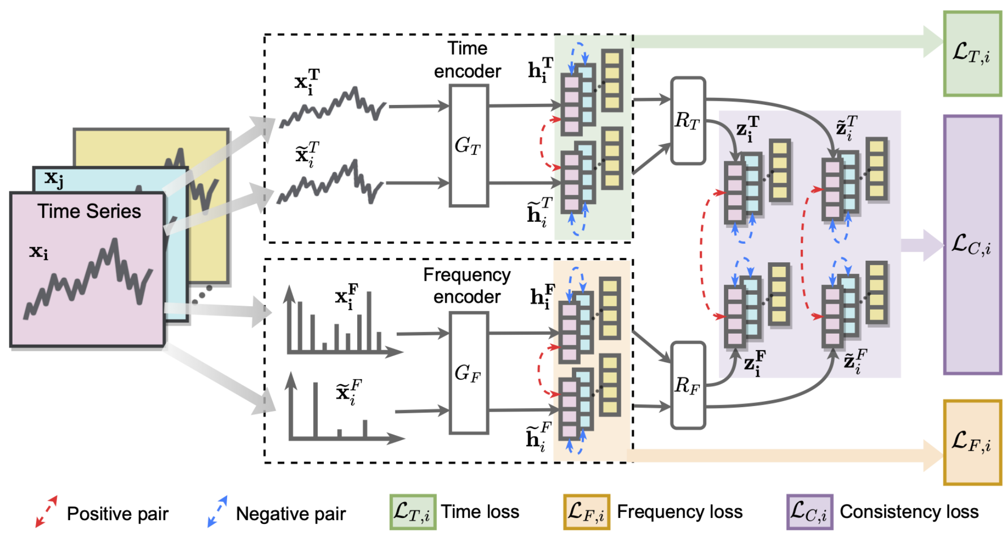

Figure 2: Borrowed idea of TF-C approach. The TF-C property is realized by promoting the alignment of time- and frequency-based representations in the latent time-frequency space, providing a vehicle for transferring F to a target dataset not seen before.

based on temporal characteristics, we send them both into the time encoder G_T, then we will have $\boldsymbol{h}_i^{\mathrm{T}} = G_{\mathrm{T}}\left(\boldsymbol{x}_i^{\mathrm{T}}\right)$ and $\widetilde{\boldsymbol{h}}_{\boldsymbol{i}}^{\mathrm{T}} = G_{\mathrm{T}}\left(\widetilde{\boldsymbol{x}}_i^{\mathrm{T}}\right)$. As $\widetilde{\boldsymbol{x}}_i^{\mathrm{T}}$ is generated based on $\boldsymbol{x}_i^{\mathrm{T}}$, after passing through G_{T}, we assume the embedding of $\boldsymbol{x}_i^{\mathrm{T}}$ is close to the embedding of $\widetilde{\boldsymbol{x}}_i^{\mathrm{T}}$ but far away from the embedding of $\boldsymbol{x}_j^{\mathrm{T}}$ and $\widetilde{\boldsymbol{x}}_j^{\mathrm{T}}$ that are derived from another sample $\boldsymbol{x}_j^{\mathrm{T}} \in \mathcal{D}^{\text{pret}}$ [Chen *et al.*, 2020a]. In specific, we select the positive pair as $\left(\boldsymbol{x}_i^{\mathrm{T}}, \widetilde{\boldsymbol{x}}_i^{\mathrm{T}}\right)$ and negative pairs as $\left(\boldsymbol{x}_i^{\mathrm{T}}, \boldsymbol{x}_j^{\mathrm{T}}\right)$ and $\left(\boldsymbol{x}_i^{\mathrm{T}}, \widetilde{\boldsymbol{x}}_j^{\mathrm{T}}\right)$.

Frequency-based Contrastive Encoder: We generate the frequency spectrum x_i^F from a time series sample x_i^T through a transform operator (e.g., Fourier Transformation [Brigham and Morrow, 1967]). The frequency information in time series is universal and plays a key role in classic signal processing [Soklaski *et al.*, 2022], but it is rarely investigated in self-supervised contrastive representation learning for time series. In this section, we develop augmentation method to perturb x_i^F based on characteristics of frequency spectra and show how to generate frequency-based representations. We hope this augmentation will improve the robustness in representation learning.

Similar to the time-based contrastive encoder, We utilize a frequency encoder G_{F} to map the frequency spectrum $\left(e.g., \boldsymbol{x}_i^{\mathrm{F}}\right)$ to a frequency-based embedding (e.g., $\boldsymbol{h}_i^{\mathrm{F}} = G_{\mathrm{F}}\left(\boldsymbol{x}_i^{\mathrm{F}}\right)$). We assume the frequency encoder G_{F} can learn similar embedding for the original frequency spectrum x_i^{F} and a slightly perturbed frequency spectrum $\widetilde{\boldsymbol{x}}_i^{\mathrm{F}} \in \mathcal{X}_i^{\mathrm{F}}$. Thus, we set the positive pair as $\left(\boldsymbol{x}_i^{\mathrm{F}}, \widetilde{\boldsymbol{x}}_i^{\mathrm{F}}\right)$ and the negative pairs as $\left(\boldsymbol{x}_i^{\mathrm{F}}, \boldsymbol{x}_j^{\mathrm{F}}\right)$ and $\left(\boldsymbol{x}_i^{\mathrm{F}}, \widetilde{\boldsymbol{x}}_j^{\mathrm{F}}\right)$.

From the time and frequency encoders above, we can now calculate two contrasive loss for sample x_i as:

$$\mathcal{L}_{\mathrm{T},i} = d\left(\boldsymbol{h}_i^{\mathrm{T}}, \widetilde{\boldsymbol{h}}_i^{\mathrm{T}}, \mathcal{D}^{\text{pret}}\right) = -\log \frac{\exp\left(\operatorname{sim}\left(\boldsymbol{h}_i^{\mathrm{T}}, \widetilde{\boldsymbol{h}}_i^{\mathrm{T}}\right)/\tau\right)}{\sum_{\boldsymbol{x}_j \in \mathcal{D}^{\text{pret}}} \mathbb{1}_{i \neq j} \exp\left(\operatorname{sim}\left(\boldsymbol{h}_i^{\mathrm{T}}, G_{\mathrm{T}}\left(\boldsymbol{x}_j\right)\right)/\tau\right)}, \tag{1}$$

$$\mathcal{L}_{\mathrm{F},i} = d\left(\boldsymbol{h}_i^{\mathrm{F}}, \widetilde{\boldsymbol{h}}_i^{\mathrm{F}}, \mathcal{D}^{\text{pret}}\right) = -\log \frac{\exp\left(\operatorname{sim}\left(\boldsymbol{h}_i^{\mathrm{F}}, \widetilde{\boldsymbol{h}}_i^{\mathrm{F}}\right)/\tau\right)}{\sum_{\boldsymbol{x}_j \in \mathcal{D}^{\text{pret}}} \mathbb{1}_{i \neq j} \exp\left(\operatorname{sim}\left(\boldsymbol{h}_i^{\mathrm{F}}, G_{\mathrm{F}}\left(\boldsymbol{x}_j\right)\right)/\tau\right)} \tag{2}$$

3.2 Time-Frequency Consistency

To measure the distance between the temporal and frequency embeddings, we map h_i^T from time space and h_i^F from frequency space to a joint time-frequency space through projectors R_T and R_F, respectively. In specific, for every input sample xi, we have four embeddings, which are $z_i^{\mathrm{T}} = R_{\mathrm{T}}\left(\boldsymbol{h}_i^{\mathrm{T}}\right), \widetilde{z}_i^{\mathrm{T}} = R_{\mathrm{T}}\left(\widetilde{\boldsymbol{h}}_i^{\mathrm{T}}\right), z_i^{\mathrm{F}} = R_{\mathrm{F}}\left(\boldsymbol{h}_i^{\mathrm{F}}\right)$, and $\widetilde{z}_i^{\mathrm{F}} = R_{\mathrm{F}}\left(\widetilde{\boldsymbol{h}}_i^{\mathrm{F}}\right)$. The first two embeddings are generated based on temporal characteristics and the latter two embeddings are produced based on the properties of frequency spectrum. After that, we use $S_i^{\mathrm{TF}} = d\left(z_i^{\mathrm{T}}, z_i^{\mathrm{F}}, \mathcal{D}^{\text{pret}}\right)$ to define the distance between z_i^T and z_i^F. So far, we can get a consistency loss $L_{C,i}$ that measures the distance between a time-based embedding and a frequency-based embedding:

$$L_{C,i} = S_i^{TF} \tag{3}$$

Dataset	Subjects	Sensors	Classes	Samples
EMG	36	1	7	33903472

Table 1: Information on EMG dataset.

3.3 Construction of loss function

Self-supervised loss: The overall loss function in pre-training has three terms. First, the time-based contrastive loss LT urges the model to learn embeddings invariant to temporal augmentations. Second, the frequencybased contrastive loss LF promotes learning of embeddings invariant to frequency spectrum-based augmentations. Third, the consistency loss LC guides the model to retain the consistency between time-based and frequency-based embeddings. In summary, the self-supervised loss is defined as:

$$L_{TF-C,i} = \lambda(L_{T,i} + L_{F,i}) + (1 - \lambda)L_{C,i} \tag{4}$$

where λ controls the relative importance of the contrastive and consistency losses. We calculate the total loss by summing $L_{TF-C,i}$ across all pre-training samples.

Supervised Classification loss: We design a normal linear classifier for the supervised classification task, and the loss is defined as cross-entropy. During classification, we concatenate the projected time and frequency embeddings obtained through the encoder:

$$L_C = crossentropy(Encoder[x_i], y) \tag{5}$$

3.4 A Normal Classifier and Separate Training

In summary, our framework is composed of two blocks: one TF-C encoder and one linear classifier. During self-supervision training, we only update the encoder through time-frequency contrasive learning to obtain label-agnostic universial representations using L_{TF-C}. It should be noted that the two blocks are trained separately. During supervised training, we froze the pretrained encoder and update the classifier to establish the relationship between the representations and labels.

4 Experiments

4.1 Dataset

Electromyography (EMG) is a typical time-series data that is based on bioelectric signals. We use EMG for gestures Data Set (Lobov et al., 2018) that contains raw EMG data recorded by MYO Thalmic bracelet. The bracelet is equipped with eight sensors equally spaced around the forearm that simultaneously acquire myographic signals. Data of 36 subjects are collected while they performed series of static hand gestures and the number of instances is 40000-50000 recordings in each column. It contains 7 classes and we select 6 common classes for our experiments. We randomly divide 36 subjects into four domains (0, 1, 2, 3) without overlapping and each domain contains data of 9 persons.

EMG data is affected by many factors since it comes from bioelectric signals. EMG data are scene and device-dependent, which means the same person may generate different data when performing the same activity with the same device at a different time (i.e., distribution shift across time (Wilson et al., 2020; Purushotham et al., 2016)) or with the different devices at the same time. Thus, the EMG benchmark is challenging.

Algorithm 1 Separate Training for Domain Generalization

Input: A set of time series sample X, one out-of-distribution time series sample x_{OOD}
Parameter: Initialized list of hyper parameters
Output: The classification result of the out-of-distribution time series x_{OOD}

Stage one: Self-supervised TF-C Contrastive Training
for $x_i, x_j (i \neq j)$ in X **do**
  Produce time based augmentation x, x^T
  Produce frequency based augmentation x, x^F
  Pass the time/frequency encoder respectively and get z^T, z^F
  Calculate the contrastive loss L_{TF-C}
  Update the encoder
end for
Stage two: Supervised Classifier Training
for $x_i, x_j (i \neq j)$ in X **do**
  Produce time based augmentation x, x^T
  Produce frequency based augmentation x, x^F
  Pass the time/frequency encoder respectively and get z^T, z^F without gradient descent
  Pass the classifier and calculate classification loss L_C
  Update the classifier
end for
Test: Out-of distribution classification
Put x_{OOD} through encoder and classifier and get y_{OOD}
return y_{OOD}

4.2 Preprocessing

We will introduce how we preprocess data and the final dimension of data for experiments here. For EMG dataset, we set the window size 200 and the step size 100, which means there exist 50 prevcents overlaps between two adjacent samples. We normalize each sample with $\tilde{x} = \frac{x - minX}{maxX - minX}$, where X contains all x. The final dimension is $8 \times 1 \times 200$.

4.3 Result

Time series OOD algorithms are currently less studied and there are only two recent strong approaches for comparison: GILE (Qian et al., 2021) and AdaRNN (Du et al., 2021).We further compare with 7 general OOD methods from DomainBed (Gulrajani & Lopez-Paz, 2021). Table 1 shows that with the same experimental settiings, our method achieves the best average accuracy performance and is 3.2 % better than the second-best method. And Table 2 gives more details on various classification performance evaluation metrics.

Target	0	1	2	3	AVG
ERM	62.6	69.9	67.9	69.3	67.4
DANN	62.9	70.0	66.5	68.2	66.9
CORAL	66.4	74.6	71.4	74.2	71.7
Mixup	60.7	69.9	70.5	68.2	67.3
GroupDRO	67.6	77.4	73.7	72.5	72.8
RSC	70.1	74.6	72.4	71.9	72.2
ANDMask	66.5	69.1	71.4	68.9	69.0
AdaRNN	68.8	81.1	75.3	78.1	75.8
Diversify	71.7	**82.4**	76.9	77.3	77.1
Ours	**80.1**	80.2	**79.2**	**81.7**	**80.3**

Table 2: Results on EMG dataset. "Target" 0 - 4 denotes unseen test distribution that is only for testing.

Target	0	1	2	3	AVG
Accuracy	80.1	80.2	79.2	81.7	67.4
Precision	87.6	82.5	86.4	82.7	66.9
Recall	88.0	83.1	87.6	83.4	71.7
F1 Score	87.7	82.7	86.5	82.6	67.3
AUROC	98.4	97.4	98.2	98.1	98.4
AUPRC	94.5	92.0	93.1	93.3	72.2

Table 3: More details on various classification performance evaluation metrics.

5 Discussion

5.1 Problem Setting of time series domain generalization

Just as the fields of computer vision and natural language processing, the field of time series not only contain domain distribution shifts, but more challenging due to both spatial and temporal variances. For instance, data collected by sensors of three persons may belong to two different distributions due to their dissimilarities. Data collected in different locations, using different sensors, and different characteristics (such as car with different brands, battery with different chemistry) undoubtedly would cause dissimilarities. This can be termed as spatial distribution shift. Moreover, there are even temporal distribution shifts in temporal data. For example, when a bank leverages a model to predict whether a person will be a "defaulted borrower", features like "annual incoming", "profession type", and "marital status" are considered. However, due to the temporal change of the society, how these feature values indicate the prediction output should change accordingly following some trends that could be predicted somehow in a range of time. Those shifts widely exist in time series, as suggested by [Zhang *et al.*, 2021] [Ragab *et al.*, 2022] Time series DG is a promising yet extremely challenging area where the goal is to learn models under spatially and temporally changing data distributions and generalize to unseen data distributions following the trends of the change.

5.2 Limits

Borrowed idea of TF-C: One of the most critical part of our framework, the TF-C self supervised encoder, is a borrowed idea from [Zhang *et al.*, 2022b]. Actually, unlike CV and NLP, supervised learning still takes the dominant of time series analysis. The reason we choose TF-C is this is the first work to develop frequency based contrastive augmentation to leverage rich spectral information and explore time-frequency consistency in time series and its good performance on fine-tuning datasets [Zhang *et al.*, 2022b]. What we did is to use it as part of our framework and try to tackle a more challenging domain generalization problem setting, where target data and labels are completely unreachable during training. Our results well explains the intuition that self-supervised learning are more inclined to acquire general features other than supervised learning, where the data distribution often comes with bias. Actually, we believe that other self-supervised learning techniques, such as masked time modeling will also work well in domain generalization under our separate training framework.

Lack of Experiments: As mentioned before, time series domain generalization is a promising yet extremely challenging and less discovered area that in great need to collect valuable and challenging datasets and establish benchmarks, in both spatial and temporal scenarios. Due to time and resource constraints, collecting and processing raw data from Internet is time and resource costing. So in this paper, we only use a public processed benchmark by [Lu *et al.*, 2023]. In future work, we will try to do experiments on more diverse datasets with both spatial and temporal distribution shifts, as well as proposing more powerful methods with novelty to tackle this challenging problem.

6 Conclusion

We proposed a two-stage separate training framework to learn generalized representation for time series classification. We take good advantage of the feature universality of self-supervised representation learning in stage one while keeping the information brought by supervised labels in stage two. Empirically, our simple method improves generalization on one time series classification benchmark for distribution shifts. Theoretically, our method accords to the robustness of self-supervised learning when facing data distribution variances.

Acknowledgments

We gratefully acknowledge support by **prof.Yang Lu** and the teaching faculty of **Deep Learning, 2023 Fall** for their thoughtful lectures and assignments, which provides us with solid theoretical and hands-on foundation in this field to do scientific research.
This work is supported by **prof.Zhihong Zhang**'s research group, school of Software, Xiamen University.

References

[Brigham and Morrow, 1967] E. O. Brigham and R. E. Morrow. The fast fourier transform. *IEEE Spectrum*, 4(12):63–70, 1967.

[Chen *et al.*, 2020a] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations, 2020.

[Chen *et al.*, 2020b] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey Hinton. Big self-supervised models are strong semi-supervised learners, 2020.

[Du *et al.*, 2021] Yuntao Du, Jindong Wang, Wenjie Feng, Sinno Pan, Tao Qin, Renjun Xu, and Chongjun Wang. Adarnn: Adaptive learning and forecasting of time series, 2021.

[Fulcher and Jones, 2014] Ben D. Fulcher and Nick S. Jones. Highly comparative feature-based time-series classification. *IEEE Transactions on Knowledge and Data Engineering*, 26(12):3026–3037, December 2014.

[Hewamalage *et al.*, 2021] Hansika Hewamalage, Christoph Bergmeir, and Kasun Bandara. Recurrent neural networks for time series forecasting: Current status and future directions. *International Journal of Forecasting*, 37(1):388–427, January 2021.

[Ismail Fawaz *et al.*, 2019] Hassan Ismail Fawaz, Germain Forestier, Jonathan Weber, Lhassane Idoumghar, and Pierre-Alain Muller. Deep learning for time series classification: a review. *Data Mining and Knowledge Discovery*, 33(4):917–963, March 2019.

[Kiyasseh *et al.*, 2021] Dani Kiyasseh, Tingting Zhu, and David A. Clifton. Clocs: Contrastive learning of cardiac signals across space, time, and patients, 2021.

[Krueger *et al.*, 2021] David Krueger, Ethan Caballero, Joern-Henrik Jacobsen, Amy Zhang, Jonathan Binas, Dinghuai Zhang, Remi Le Priol, and Aaron Courville. Out-of-distribution generalization via risk extrapolation (rex), 2021.

[Li *et al.*, 2020] Shiyang Li, Xiaoyong Jin, Yao Xuan, Xiyou Zhou, Wenhu Chen, Yu-Xiang Wang, and Xifeng Yan. Enhancing the locality and breaking the memory bottleneck of transformer on time series forecasting, 2020.

[Lu *et al.*, 2023] Wang Lu, Jindong Wang, Xinwei Sun, Yiqiang Chen, and Xing Xie. Out-of-distribution representation learning for time series classification, 2023.

[Matsuura and Harada, 2019] Toshihiko Matsuura and Tatsuya Harada. Domain generalization using a mixture of multiple latent domains, 2019.

[Peng *et al.*, 2019] Xingchao Peng, Zijun Huang, Ximeng Sun, and Kate Saenko. Domain agnostic learning with disentangled representations, 2019.

[Ragab *et al.*, 2022] Mohamed Ragab, Zhenghua Chen, Wenyu Zhang, Emadeldeen Eldele, Min Wu, Chee-Keong Kwoh, and Xiaoli Li. Conditional contrastive domain generalization for fault diagnosis. *IEEE Transactions on Instrumentation and Measurement*, 71:1–12, 2022.

[Rebjock *et al.*, 2021] Quentin Rebjock, Barış Kurt, Tim Januschowski, and Laurent Callot. Online false discovery rate control for anomaly detection in time series, 2021.

[Sarkar and Etemad, 2020] Pritam Sarkar and Ali Etemad. Self-supervised learning for ecg-based emotion recognition. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, May 2020.

[Shi *et al.*, 2021] Pengxiang Shi, Wenwen Ye, and Zheng Qin. Self-supervised pre-training for time series classification. *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8, 2021.

[Soklaski *et al.*, 2022] Ryan Soklaski, Michael Yee, and Theodoros Tsiligkaridis. Fourier-based augmentations for improved robustness and uncertainty calibration, 2022.

[Wang *et al.*, 2022] Jindong Wang, Cuiling Lan, Chang Liu, Yidong Ouyang, Tao Qin, Wang Lu, Yiqiang Chen, Wenjun Zeng, and Philip S. Yu. Generalizing to unseen domains: A survey on domain generalization, 2022.

[Zhang *et al.*, 2021] Wenyu Zhang, Mohamed Ragab, and Ramon Sagarna. Robust domain-free domain generalization with class-aware alignment, 2021.

[Zhang *et al.*, 2022a] Hanlin Zhang, Yi-Fan Zhang, Weiyang Liu, Adrian Weller, Bernhard Schölkopf, and Eric P. Xing. Towards principled disentanglement for domain generalization, 2022.

[Zhang *et al.*, 2022b] Xiang Zhang, Ziyuan Zhao, Theodoros Tsiligkaridis, and Marinka Zitnik. Self-supervised contrastive pre-training for time series via time-frequency consistency, 2022.