TRLS: A TIME SERIES REPRESENTATION LEARNING FRAMEWORK VIA SPECTROGRAM FOR MEDICAL SIGNAL PROCESSING

Luyuan Xie[1,2], *Cong Li*[1,2], *Xin Zhang*[1,2], *Shengfang Zhai*[1,2], *Yuejian Fang*[1,2], *Qingni Shen*[1,2], *Zhonghai Wu*[1,2,*]

[1]School of Software and Microelectronics, Peking University, Beijing, China
[2]National Engineering Research Center for Software Engineering, Peking University;

ABSTRACT

Representation learning frameworks in unlabeled time series have been proposed for medical signal processing. Despite the numerous excellent progresses have been made in previous works, we observe the representation extracted for the time series still does not generalize well. In this paper, we present a **T**ime series (medical signal) **R**epresentation **L**earning framework via **S**pectrogram (TRLS) to get more informative representations. We transform the input time-domain medical signals into spectrograms and design a time-frequency encoder named Time Frequency RNN (TFRNN) to capture more robust multi-scale representations from the augmented spectrograms. Our TRLS takes spectrogram as input with two types of different data augmentations and maximizes the similarity between positive ones, which effectively circumvents the problem of designing negative samples. Our evaluation of four real-world medical signal datasets focusing on medical signal classification shows that TRLS is superior to the existing frameworks. We will open-source our code when the paper is accepted.

Index Terms— Medical signal, time series, representation learning, spectrogram, Time Frequency RNN

1. INTRODUCTION

Medical signal is a type of time series which plays a crucial role in medical fields such as ECG prediction . With the development of IoT and wearable devices, it is more convenient to collect time series [1]. However, the annotation of time series is greatly limited due to the high requirement for professional knowledge, which make learning representations from unlabeled time series a significant and meaningful challenge.

Due to this challenge, self-supervised learning methods [2–4] are introduced to learn effective data representations from unlabeled data. As a kind of self-supervised representation learning method, contrastive learning performs well in many downstream tasks [5–10], and particularly in time series analysis field [1, 11–13] as well.

This work was supported by the National Key R&D Program of China under Grant No.2022YFB2703301.“*” denotes the corresponding authors.

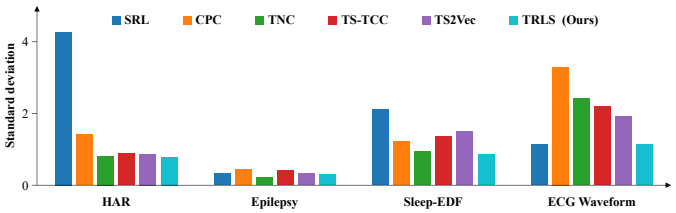


Fig. 1: The accuracy standard deviation of the time series (medical signal) contrastive learning frameworks under 5-fold cross validation for four datasets. The experimental results show that the existing framework performs significantly different when training data changing.

Despite the excellent progress made in previous works, the representation extracted for the time series still does not generate well. Fig. 1 shows an example of previous frameworks by using different training set. Due to their inability to generate sufficiently robust representations, they have significant differences in accuracy when trained on different subsets of the same dataset.

To tackle this problem, we use the spectrogram as input to enhance the representations with more frequency domain information. The frequency part would capture medical signals from a different perspective, which will compensate for the information missing from the time domain, such as high-frequency details. In this work, we show ability of frequency domain information in medical signal analysis. Moreover, to better encode the time-frequency representation, we proposed a time-frequency encoder network to encode the spectrogram into representations. This network handles spectrogram as a series not only in temporal axis but also in frequency axis. The temporal axis serial encoding would capture the spectrum similarity and variation along time, while the frequency axis serial encoding would capture the temporal similarity and variation along different frequencies. This time-frequency encoder network would capture representations with more frequency information for downstream tasks.

Specifically, we propose a **T**ime series (medical signal) **R**epresentation **L**earning framework via **S**pectrogram (TRLS) based on time-frequency analysis. First, we convert the time series into a spectrogram through short-time Fourier transform (STFT). TRLS can train a more robust encoder by max-

imizing the similarity between positive samples in the training stage and it avoids the problem of constructing fragile negative samples. Besides, we design a new encoder called Time-Frequency RNN (TFRNN) to extract multi-scale representations of the augmented spectrogram. We use RNN on both temporal and frequency axises, so the network can easily capture the similarity and variation of different time slots and frequencies. In summary, the main contributions of this work are as follows.

(1) We propose TRLS, a novel time series representation learning framework that uses spectrograms to generate more robust representations.

(2) TRLS incorporates a new encoder, TFRNN, that extracts multi-scale representations of the augmented spectrogram to learn more robust representations of time and frequency.

(3) We evaluate TRLS on four public datasets focusing on medical signal classification and show that it outperforms state-of-the-art frameworks.

2. METHODS

TRLS Framework. The aim of our proposed TRLS is to learn more effective representations R_i ($i \in K$, K denotes the number of representations) for downstream tasks. As depicted in Fig. 2(a), TRLS framwork operates on time series T by using STFT and two different data augmentations to generate the spectrogram in two different augmented views, denoted as V and V'. The online network takes the first augmented view V as input and produces multi-scale representations R_i, as well as multi-scale projections Pj_i and Pd_i. In contrast, the target network uses the second augmented view V' and outputs R'_i and multi-scale projections Pj'_i. Notably, the predictor is only applied to the online network, resulting in an asymmetric architecture between the online and target pipelines. Additionally, a mean squared error (MSE) is defined to evaluate the normalized multi-scale predictions against the target multi-scale projections.

$$L = \frac{1}{K}\sum_{i=1}^{K}\left\|\overline{Pd}_i - \overline{Pj}'_i\right\|_2^2 = \frac{1}{K}\sum_{i=1}^{K} 2 - 2\times\frac{\langle Pd_i, Pj'_i\rangle}{\|Pd_i\|_2\times\left\|Pj'_i\right\|_2} \quad (1)$$

In order to symmetrize the loss function L defined in Eq.1, we compute a new term L' by seprately feeding V' to the online network and V to the target network.

$$L_{total} = L + L' \quad (2)$$

L_{total} is used to maximize the similarity between positive samples for training the encoder. This means that the TRLS does not require negative samples in the training stage, and effectively avoids the problems caused by the inappropriate negative samples. By performing an adam optimization step, we minimize L_{total} for updating the online network's parameters β at each training step. The target network is depicted by the stop gradient in Fig. 2(a). Target network's parameters δ update by:

$$\delta \leftarrow \tau\delta + (1-\tau)\beta \quad (3)$$

τ is the moving average decay.

Data Augmentation. Data augmentation is a crucial component of contrastive learning methods [2, 14] and has been shown to significantly influence the performance of trained encoders in recent research [1, 2, 15]. Therefore, it is imperative to design appropriate data augmentation strategies for our proposed framework. Traditional contrastive learning methods often use two (random) variants of the same augmentation, which can negatively impact the robustness of learned representations. To avoid this issue, we follow the approach in [1] and employ two different data augmentation methods. Existing time series contrastive learning methods only rely on time-domain data augmentation without considering augmentation based on spectrogram, so we incorporate image augmentation techniques on the spectrogram, such as ColorJitter and GaussianBlur. In TRLS, we apply ColorJitter and RandomHorizontalFlip for the first augmentations and Gaussian Blur and RandomHorizontalFlip for the second augmentations[16, 17].

The Encoder for the Spetcrogram. Temporal CNN (TCN) [18] is used in the time series contrastive learning frameworks frequently. It adopts 1D convolution that can well extract time series time domain features. Actually, apart from time series domain information, spectrogram also includes frequency domain information, which results in that TCN may lose frequency information when extracting time dimension information of spectrogram. To cope with this problem, we design a Time and Frequency RNN (TFRNN) to learn the time-frequency characteristics of the spectrogram. As shown in Fig. 2(c), the Time-Frequency RNN block (TFRblock) applies different RNN for the time and frequency of the spectrogram accordingly Fig. 2(b). For any given feature map $X \in t \times f$ (t denotes the time dimension, and f denotes the frequency dimension), TFRblocak generates $Y \in t \times f$.

Based on TFRblock , we propose TFRNN (Fig. 2(d)) as the encoder of the TRLS. After the input spectrogram passes through three layers of TFRNN, the high-dimensional mapping $f \rightarrow f_h$ of the frequency domain dimension is performed with the 1D convolution (kernel size = 1) to obtain the feature $H \in t \times f_h$. Then, the pyramid pooling [19] is adopted and max pooling is used for multi-scale downsampling in the time dimension of feature Y_{out} to get multi-scale features D_i:

$$D_i = \begin{cases} \text{Maxpooling}(D_{i-1}) & 2 \le i \le K \\ D_1 & i = 1 \end{cases} \quad (4)$$

where K refers to the number of downsampling times. The global average pooling (GAP) of time dimension is performed on the multi-scale features D_i to obtain multi-scale representations R_i.

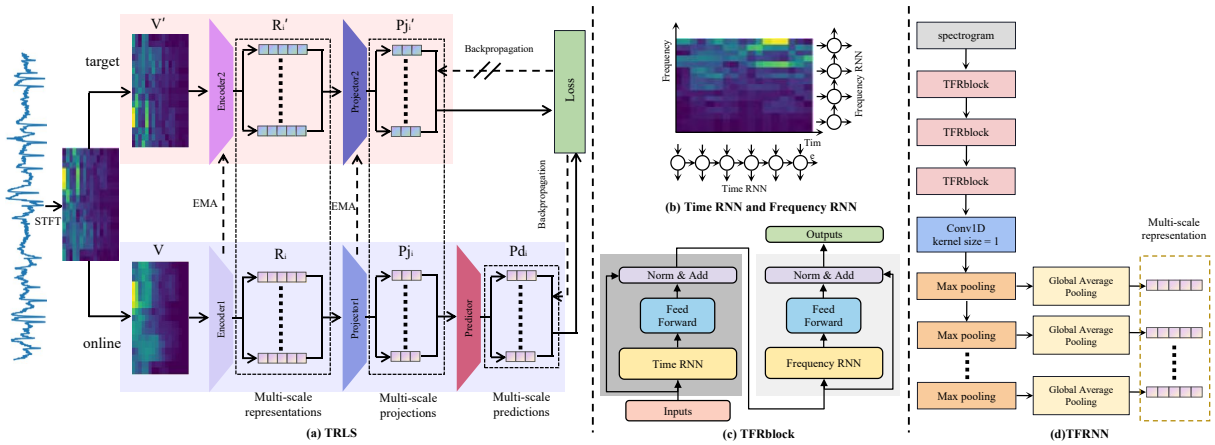


Fig. 2: Overall architecture of proposed TRLS framework.

	HAR		Epilepsy		Sleep-EDF		ECG Waveform	
	ACC	MF1	ACC	MF1	ACC	MF1	ACC	MF1
Supervised	92.44±1.09	90.32±0.88	97.33±1.09	95.62±0.59	**85.80±1.32**	74.76±0.40	84.67±2.37	67.36±1.24
SRL	64.70±4.27	62.37±1.37	87.62±0.33	83.47±0.69	77.31±2.12	67.73±0.57	75.21±1.14	53.44±1.32
CPC	86.43±1.41	83.27±1.66	96.61±0.43	94.44±0.76	83.10±1.22	73.31±0.73	69.11±3.30	50.22±1.78
TNC	89.12±0.81	88.67±0.49	95.44±0.21	95.21±0.43	82.97±0.94	71.34±0.91	78.01±2.42	60.32±1.93
TS-TCC	91.89±0.89	89.91±0.44	97.65±0.41	95.74±0.29	83.31±1.36	72.47±0.46	76.33±2.20	62.21±1.19
TS2Vec	90.44±0.87	88.42±0.24	97.67±0.32	96.01±0.47	83.07±1.49	71.29±0.73	78.41±1.92	63.09±1.44
TRLS (ours)	**93.61±0.73**	**91.23±0.27**	**97.92±0.22**	**96.02±0.31**	85.40±0.82	**76.76±0.41**	**88.73±1.51**	**68.83±0.32**

Table 1: Comparisons between our proposed TRLS framework against baselines using linear classifier evaluation experiment.

K	ECG Waveform	
	ACC	MF1
1	85.29±2.57	63.94±1.75
3	86.97±0.74	66.61±0.87
5	**88.73±1.51**	68.83±0.32
7	87.92±1.62	69.07±0.49
9	88.02±1.57	69.14±0.26
11	87.44±1.88	**69.19±0.33**

Table 2: Effect of K on robustness.

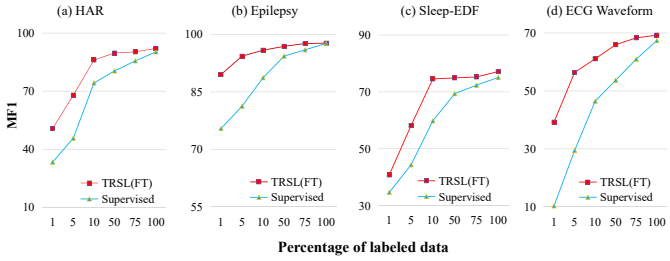


Fig. 3: Comparisons between supervised training vs TRLS fine-tuning for different few-labeled data scenarios in terms of MF1.

3. EXPERIMENT

3.1. Dataset and Setup

To verify the effectiveness of our framework, we compare our framework with the current optimal time series representation learning framework on four public medical signal datasets. Specifically, they are the UCI HAR, the Epileptic Seizure Recognition, the Sleep-EDF and the ECG Waveform.

We divide four datasets into training set, valid set and test set according to 6:2:2 ratio. For all experiments, we conduct 5-fold cross-validation, and report the mean and standard deviation. In the pre-training stage, the epoch is set to 50 and the batch size is 32. The epoch and batch size are 100 and 128 in the downstream tasks. We use Adam optimizer with a learning rate of 3e-4, weight decay of 1e-4, $\beta 1 = 0.9$, and $\beta 2 = 0.99$. We design two different augmentations. Particularly, our target network's moving average decay $\tau = 0.7$. We also set RNN's dropout to 0.2 and the number of multi-scale representations K to 5. When STFT is performed on input data, the window size of FFT for short time series sets (HAR, Epilepsy) is 32, while that for long time series sets (Sleep-EDF, ECG Waveform) is 128. Moreover, we build our framework using PyTorch 1.10 and train it on a NVIDIA GeForce RTX 2080 Ti GPU.

3.2. Experiment Results

Comparisons with Baseline Approaches. We evaluate the performance of our framework using the standard linear benchmarking evaluation scheme in Table 1. Our results show that TRLS achieves the best performance in three datasets. For Sleep-EDF, its performance is similar to supervised learning. Additionally, TRLS exhibits the smallest variance, indicating that the representations generated by TRLS are robust enough to maintain stable performance across different training data. For investigating the effectiveness of our TRLS under semi-supervised settings, we train the model with 1%, 5%, 10%, 50%, and 75% of randomly selected instances from the training data. TRLS finetune (i.e., red curves in Fig. 3) means that a few labeled samples are employed to finetune the pre-trained encoder. With only 10% data by TRLS fine-tuning, TRLS can achieve supervised training performance

dropout	Sleep-EDF		SNR(db)	Sleep-EDF	
	ACC	MF1		ACC	MF1
0	**85.40±0.82**	**76.76±0.41**	-	**85.40±0.82**	**76.76±0.41**
0.1	85.14±0.62	75.3±0.42	10	84.21±0.84	75..01±1.54
0.2	85.11±0.75	74.6±0.54	5	83.39±1.11	73.65±0.37
0.3	84.52±0.96	74.24±0.98	1	82.02±1.74	72.01±1.33
0.4	83.61±1.08	73.58±0.46	0.9	81.20±0.38	71.12±0.32
0.5	82.87±0.83	72.36±0.58	0.7	81.85±0.43	71.67±0.55
0.6	81.19±1.33	71.12±0.97	0.5	81.71±0.19	71.53±0.83
0.7	79.27±0.70	70.02±1.41	0.3	81.54±0.38	71.37±0.62
0.8	76.53±1.24	66.41±0.50	0.1	81.30±0.41	71.22±0.61
0.9	66.53±2.51	58.44±2.06	0.01	81.26±0.77	71.08±0.07

Table 3: TRLS performance in time series with different sparsity and SNR.

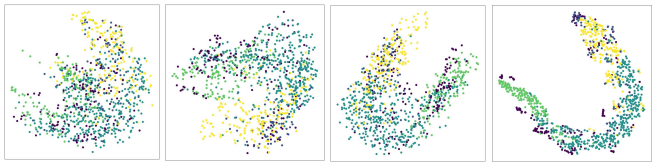
	Framework	Encoder	HAR	Epilepsy	Sleep-EDF	ECG Waveform
Supervised	TS-TCC	TFR	**91.83±1.23**	**97.65±0.35**	**83.7±0.27**	**83.21±0.44**
		TCN	90.14±2.49	96.66±0.24	83.41±1.44	82.43±1.32
	TNC	TFR	**92.52±1.33**	**96.37±1.22**	**85.2±1.32**	**85.33±0.27**
		TCN/RNN	92.03±2.48	94.81±0.28	83.72±0.74	84.81±0.28
Linear evaluation	TS-TCC	TFR	**91.89±0.89**	**97.65±0.41**	**83.31±1.36**	**76.33±2.20**
		TCN	90.37±0.34	97.23±0.10	83.00±0.71	74.81±1.10
	TNC	TFR	**89.12±0.81**	**95.44±0.21**	**82.97±0.94**	**78.01±2.42**
		TCN/RNN	88.32±0.12	93.22±0.42	81.33±0.33	77.79±0.84

Table 4: The results of different encoders in the TS-TCC or TNC with supervised and linear classifier evaluation.

	Sleep-EDF		ECG Waveform	
	ACC	MF1	ACC	MF1
TRLS	**85.40±0.82**	**76.76±0.41**	**88.73±1.51**	**68.83±0.32**
w/o spetcrogram	83.43±0.85	73.11±0.53	79.11±1.13	62.33±0.82
w/o mutli-scale representations	83.80±1.72	74.79±0.41	85.29±2.57	63.94±1.75
w/o ColorJitter	84.70±0.46	75.44±0.64	86.20±1.27	66.91±0.83
w/o RandomHorizontalFlip	83.93±0.64	74.96±0.57	85.61±1.39	64.23±0.96
w/o GaussianBlur	84.92±1.44	75.82±0.55	86.54±0.73	67.18±0.74
w/ Negative samples	84.77±0.91	75.21±0.59	87.93±0.82	66.92±0.99
Encoder				
LSTM	**85.40±0.82**	**76.76±0.41**	**88.73±1.51**	**68.83±0.32**
→ GRU	85.21±0.87	76.03±0.62	87.47±0.89	67.73±0.51
→ RNN	84.03±0.99	76.76±0.41	86.09±0.71	65.21±0.85
→ TCN	83.01±1.41	72.93±1.31	82.12±2.33	61.31±1.71
Augmentation				
Color Jitter / Gaussian Blur	**85.40±0.82**	**76.76±0.41**	**88.73±1.51**	**68.83±0.32**
Color Jitter / Color Jitter	84.32±1.47	75.47±0.93	87.63±1.22	67.46±0.51
JS / permutation	79.71±0.62	72.93±1.21	79.94±1.46	62.03±2.12
JS / JS	77.39±2.02	66.41±2.43	79.64±3.11	60.82±1.97
permutation / permutation	72.62±3.11	63.18±2.93	71.87±2.45	56.33±3.11
JS / Color Jitter	81.66±0.98	72.93±1.21	83.61±1.49	63.94±1.09

Table 5: Ablation study of each component in TRLS and data augmenation performed with linear classifier evaluation experiment. Red and blue represent different types data augmentations on the spectrogram and time domain, respectively.

close to 100% with the four datasets. This demonstrates the effectiveness of TRLS under semi-supervised settings.



(a) TNC (b) TS-TCC (c) TS2Vec (d) TRLS

Fig. 4: TSNE map of the time series representations from different frameworks in the Sleep-EDF.

Representations Visualization. In Fig. 4, TSNE map implies that representations generated by TRLS framework possess superior discriminability compared to other frameworks. This observation is highly beneficial for downstream tasks.

Quantitative Robustness Experiments. A series of quantitative experiments on TRLS to show the robustness of it in Table 2 and Table 3. We adjust K that is the number of multi-scale representations and then observe the results in ECG Waveform. To ensure the balance between accuracy and MF1, our framework chooses K = 5. In the end, we evaluate TRLS's performance on Sleep-EDF by applying dropout and adding Gaussian noise with different Signal to Noise Ratio (SNR). The experimental results of Table 3 demonstrate that TRLS exhibits good robustness to data sparsity and noise.

The Effectiveness of TFRblock. We compare TFRblock with TCN or RNN in TS-TCC and TNC's downstream classification tasks. The results of our experiments on four datasets are presented in Table 4. Our TFRblock outperforms the TCN and RNN in both supervised and linear evaluation tasks, respectively. This implies that our TRFblock is a better feature extraction module. It can achieve good performance not only on spectrogram data but also on time domain data as input in contrastive learning frameworks.

Ablation Study. Table 5 shows ablation study of TRLS. First, whether to use the spetcrogram has the greatest impact on TRLS. Second, the results of modifying LSTM in TFRblock to GRU, RNN and TCN shows that both RNN and TCN significantly decrease compared to LSTM, while GRU just decreases slightly. Last, we conduct different types of data augmentation experiments on the spectrogram and time domain. The spectrogram augmentation directly applies data augmentation to the spectrogram, while the time domain augmentation involves transforming the time domain data into a spectrogram after time domain data augmentation. The results indicate that different spetcrogram augmentations effectively improve TRLS performance.

4. CONCLUSION

This paper proposes a novel framework called TRLS for better learning appropriate representations in medical signal processing. TRLS focuses on problems that are not concerned in the current mainstream frameworks: complexity and sparsity of time series, more robust encoders and how to construct negative samples. It adopts spectrogram, TFRNN, and training without negative samples to solve the above problems respectively. The experiment results show that TRLS is prior to all the existing ones on all the used datasets. In the future, TRLS can be applied to a wider range of time series tasks, not merely medical signal.

5. REFERENCES

[1] Emadeldeen Eldele, Mohamed Ragab, Zhenghua Chen, Min Wu, Chee Keong Kwoh, Xiaoli Li, and Cuntai Guan, "Time-series representation learning via temporal and contextual contrasting," *arXiv preprint arXiv:2106.14112*, 2021.

[2] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton, "A simple framework for contrastive learning of visual representations," in *International conference on machine learning*. PMLR, 2020, pp. 1597–1607.

[3] Mehdi Noroozi and Paolo Favaro, "Unsupervised learning of visual representations by solving jigsaw puzzles," in *European conference on computer vision*. Springer, 2016, pp. 69–84.

[4] Spyros Gidaris, Praveer Singh, and Nikos Komodakis, "Unsupervised representation learning by predicting image rotations," *arXiv preprint arXiv:1803.07728*, 2018.

[5] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton, "A simple framework for contrastive learning of visual representations," in *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*. 2020, vol. 119 of *Proceedings of Machine Learning Research*, pp. 1597–1607, PMLR.

[6] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 9729–9738.

[7] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al., "Bootstrap your own latent-a new approach to self-supervised learning," *Advances in neural information processing systems*, vol. 33, pp. 21271–21284, 2020.

[8] Xinlei Chen and Kaiming He, "Exploring simple siamese representation learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 15750–15758.

[9] Xinlei Chen, Saining Xie, and Kaiming He, "An empirical study of training self-supervised vision transformers," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 9640–9649.

[10] Chongjian Ge, Youwei Liang, Yibing Song, Jianbo Jiao, Jue Wang, and Ping Luo, "Revitalizing cnn attention via transformers in self-supervised visual representation learning," *Advances in Neural Information Processing Systems*, vol. 34, pp. 4193–4206, 2021.

[11] Travers Ching, Daniel S Himmelstein, Brett K Beaulieu-Jones, Alexandr A Kalinin, Brian T Do, Gregory P Way, Enrico Ferrero, Paul-Michael Agapow, Michael Zietz, Michael M Hoffman, et al., "Opportunities and obstacles for deep learning in biology and medicine," *Journal of The Royal Society Interface*, vol. 15, no. 141, pp. 20170387, 2018.

[12] Jean-Yves Franceschi, Aymeric Dieuleveut, and Martin Jaggi, "Unsupervised scalable representation learning for multivariate time series," *Advances in neural information processing systems*, vol. 32, 2019.

[13] Sana Tonekaboni, Danny Eytan, and Anna Goldenberg, "Unsupervised representation learning for time series with temporal neighborhood coding," *arXiv preprint arXiv:2106.00750*, 2021.

[14] Yuanhao Zhai, Tianyu Luan, David Doermann, and Junsong Yuan, "Towards generic image manipulation detection with weakly-supervised self-consistency learning," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2023, pp. 22390–22400.

[15] Ling Yang and Shenda Hong, "Unsupervised time-series representation learning with iterative bilinear temporal-spectral fusion," in *International Conference on Machine Learning*. PMLR, 2022, pp. 25038–25054.

[16] Tianyu Luan, Yuanhao Zhai, Jingjing Meng, Zhong Li, Zhang Chen, Yi Xu, and Junsong Yuan, "High fidelity 3d hand shape reconstruction via scalable graph frequency decomposition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2023, pp. 16795–16804.

[17] Tianyu Luan, Yali Wang, Junhao Zhang, Zhe Wang, Zhipeng Zhou, and Yu Qiao, "Pc-hmr: Pose calibration for 3d human mesh recovery from 2d images/videos," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021, vol. 35, pp. 2269–2276.

[18] Shaojie Bai, J Zico Kolter, and Vladlen Koltun, "An empirical evaluation of generic convolutional and recurrent networks for sequence modeling," *arXiv preprint arXiv:1803.01271*, 2018.

[19] Luyuan Xie, Cong Li, Zirui Wang, Xin Zhang, Boyan Chen, Qingni Shen, and Zhonghai Wu, "Shisrcnet: Super-resolution and classification network for low-resolution breast cancer histopathology image," *arXiv preprint arXiv:2306.14119*, 2023.