Language Pre-training Guided Masking Representation Learning for Time Series Classification

Liaoyuan Tang, Zheng Wang*, Jie Wang, Guanxiong He, Zhezheng Hao, Rong Wang, Feiping Nie

School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University
127 West Youyi Road, Beilin District
Xi'an Shaanxi, 710072, P.R.China
tangly@mail.nwpu.edu.cn, zhengwangml@gmail.com, jiewang.dl@mail.nwpu.edu.cn, heguanx@mail.nwpu.edu.cn, haozhezheng@outlook.com, wangrong07@tsinghua.org.cn, feipingnie@gmail.com

Abstract

The representation learning of time series has a wide range of downstream tasks and applications in many practical scenarios. However, due to the complexity, spatiotemporality, and continuity of sequential stream data, compared with the representation learning of structural data such as images/videos, the time series self-supervised representation learning is even more challenging. Besides, the direct application of existing contrastive learning and masked autoencoder based approaches to time series representation learning encounters inherent theoretical limitations, such as ineffective augmentation and masking strategies. To this end, we propose a Language Pre-training guided Masking Representation Learning (LPMRL) for times series classification. Specifically, we first propose a novel language pre-training guided masking encoder for adaptively sampling semantic spatiotemporal patches via natural language descriptions and improving the discriminability of latent representations. Furthermore, we present the dual-information contrastive learning mechanism to explore both local and global information by meticulously designing high-quality hard negative samples of time series data samples. As a result, we also design various experiments, such as visualization of masking position and distribution and reconstruction error to verify the reasonability of proposed language guided masking technique. Last, we evaluate the performance of proposed representation learning via classification task conducted on 106 time series datasets, which demonstrates the effectiveness of proposed method.

Introduction

Time series data are being incrementally collected from multiple mechanical sensors, which is crucial for various downstream tasks, such as fault diagnosis (Chen, Liu, and Kong 2020), machine remaining useful life (RUL) prediction (Chen et al. 2020b) and classification (Eldele et al. 2023). However, due to the high-dimensionality, unstructured and temporality, large amounts of informative patterns that hidden in long sequence cannot be observed by human, severely limiting the applications of time series data. Consequently, with the advance of deep learning techniques, representation learning based approaches have been widely exploited to make time series more interpretable.

*Corresponding author.

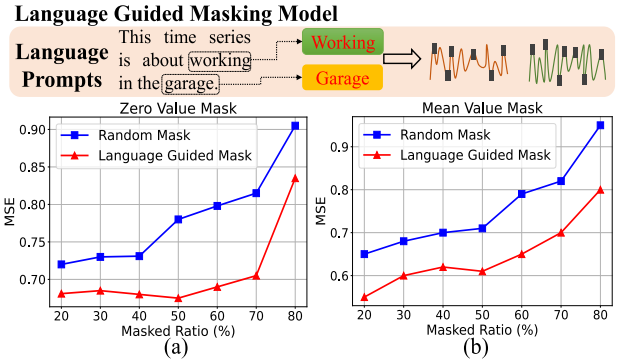


Figure 1: Randomly masking vs. proposed language guided masking. We evaluate the performance different masking strategy in terms of the metrics of mean square errors on wine time series data. Obviously, proposed language guided masking strategy outperforms randomly masking more than 0.1 reconstruction errors on both zero value mask (a) and mean value mask (b) with almost masked ratios, which demonstrates the effectiveness of proposed language guided masking model.

Representation learning aims to learn low-dimensional discriminative patterns from complex high-dimensional data which has achieved remarkable success in image/video understanding recently. Concretely, representation learning methods of images can be divided into following three parts: traditional linear dimension reduction (Nie et al. 2019) and manifold learning (Law and Jain 2006) based approaches assume that the essential structural characteristics are hidden in the low-dimensional manifold neighbouring subspace. In order to deal with large scale data, deep learning based representation learning methods have dominated the development trend of this field, such as ResNet (He et al. 2016), transformer (Vaswani et al. 2017), masked autoencoder (He et al. 2022) and so on. For further alleviating label limitation, contrastive learning (Chen et al. 2020a) has heralded a new era of remarkable achievement for self-supervised representation learning.

Despite the significant attention garnered by the above visual representation paradigms, which cannot be directly

ported to time series data. For instance, it is noting that time series data have much more complex dynamic patterns hidden in long continuous sequence, making it difficult to recognize their categories via linear transformation (Liu et al. 2024). Furthermore, the data augmentation strategies of positive and negative samples may destroy the inexplicable underlying patterns in time series, causing the major contrastive based self-supervised representation learning methods ineffective as usual (Luo et al. 2023). Last but not least, since the off-the-shelf tokens extracted from image and language patches are much more than those learned from time series patches, the underlying spatiotemporal information of time series sequence is difficult to be explored via masked autoencoder with random masking strategy. Besides, a well-trained masking sampling strategy on one dataset might not suitable for another ones (Bandara et al. 2023). Therefore, crafting sophisticated representation learning method exclusively tailored for time series data, with a focus on circumventing above issues, represents a pivotal step towards enhancing their efficacy of downstream tasks.

In this paper, we introduce a Language Pre-training Guided Masking Representation Learning (LPMRL) for time series classification. Inspired by a fact that random masking partially obscures meaningful regions, which allows easier inference of masked values by leveraging strong correlations among local regions (Shi et al. 2022). Therefore, we believe that the *masked semantic content* is crucial for effective self-supervised representation learning. Our method consists of two main mechanisms, the first is the adaptive masking mechanism driven by a pre-training language encoder, selecting semantic region as masked patches in sequences rather than random sampling strategy. The second is the dual-information contrastive learning mechanism for capturing discriminative information of representations.

We empirically show that our proposed language pre-training guided masking mechanism is able to sample more informative tokens from semantic spatiotemporal patches than random masking strategy, which is verified in Figure 1. We also conduct extensive experiments on time series classification downstream task for evaluating the effectiveness of proposed method. In summary, our contribution are:

- We propose a novel language pre-training guided masking encoder for adaptively sampling semantic spatiotemporal patches of continuous sequence by utilizing natural language descriptions generated from labels of time series data, so as to improve the discriminability of representations learned by adaptive masked antoencoder.
- We introduce the dual-information contrastive learning mechanism to explore both local and global information from long time series sequence for further improving the discriminability of learned representations in the self-supervised manner.
- We evaluate the performance of proposed representation learning method by conducting time series classification experiments on 106 benchmark time series datasets. The results shown that our method outperforms several SOTA related competitors, which verifies the effectiveness of our method.

Related Work

Neural Architectures for Time Series. In the early stages, the multi-layer perceptrons (MLP) based representation learning models directly employed fully connected network to extract low-dimensional features without consider the temporal relationships between the time series data (Ismail Fawaz et al. 2019). In order to learn temporal features in time series, recurrent neural networks and long short-term memory based approaches integrate memory cells with a gating mechanism to capture long-term dependencies in sequential data (Tang et al. 2020). With the development of CV techniques, convolutional neural networks use several convolution kernels to aggregate local information at each time step (Li et al. 2021). Moreover, a more fashionable model, named temporal convolutional networks (Lea et al. 2017), equips casual convolution operation to solve information leakage issue from past time step. Recently, transformers based approaches adapt multi-headed self-attention mechanism to dynamically computing the relationships among temporal representations themselves, which achieving notable performance in time series analysis (Wen et al. 2023).

Contrastive Learning for Time Series. Contrastive learning is a key technique in self-supervised learning that aims to learn invariant representations from different augmentations of the original data samples. Unlike self-supervised of image, the major challenge of contrastive learning for time series lies in the fact that existing augmentation techniques generally cannot be applied to sequential data. The address this issue, PCRTA (Tang et al. 2024), TS-TCC (Eldele et al. 2021) and COUTA (Xu et al. 2024) use jittering and permutation techniques to generate positive and negative samples for data augmentation. However, such strong augmentation strategies are prone to disrupting the informative patterns, causing negative samples ineffective. Therefore, ExpCLR (Nonnenmacher et al. 2022) and InfoTS (Luo et al. 2023) design implicit hard-negative mining and information-aware augmentation strategies for increasing the interpretability and stavility of negative samples generation. Recently, CDCC (Peng et al. 2024) simultaneously represents time series data in both temporal and frequency domains, leveraging contrastive learning to enhance the compactness of representations within the same domain.

Masked Autoencoder for Time Series. Masked Autoencoder (He et al. 2022) is another fashionable self-supervised representation learning technique which aims to encode the masking corrupt input into a latent space, and recovers them via a simple decoder. Due to the remarkable success of masked autoencoder in language understanding, it has gradually applied to the representation learning of time series. For example, TARNet (Chowdhury et al. 2022) focus on selecting the most informative seqments and performing data reconstruction within a single modality. TimeMAE (Cheng et al. 2023) design an end-to-end decoupled autoencoder architecture for time series representations, which can eliminate the discrepancy issues. Ti-MAE (Li et al. 2023) randomly masks embeddings, and learns a decoder to reconstruct which at point level. Simmtm (Dong et al. 2024) presents a pre-training framework of masked time series

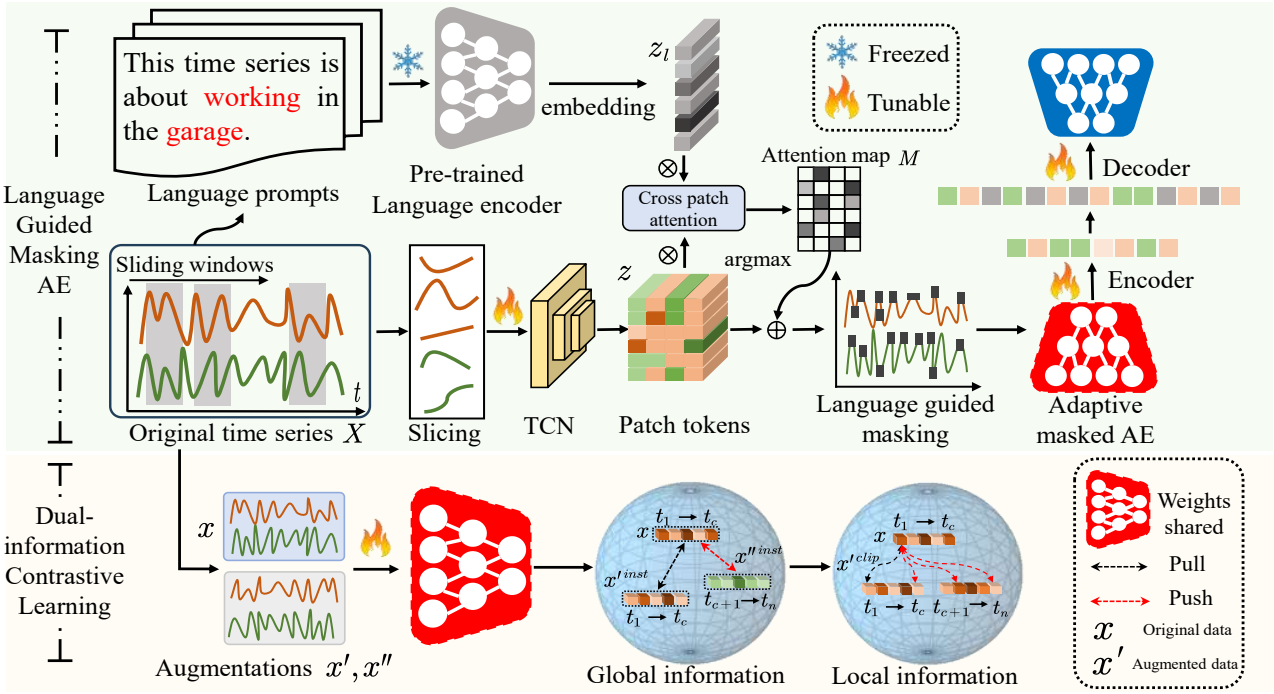


Figure 2: An illustration of the proposed Language Pre-training guided Masking Representation Learning model (LPMRL), consisting of language guided masking autoencoder module for adaptively masking patches in time series, and dual-information contrastive learning module for exploring both local and global information of learned representations.

modeling that recovers masked values according to the weighted aggregation of multiple neighbor within in same manifold. Recently, TimesURL (Liu and Chen 2024) proposes a masked autoencoder module to preserve important temporal variation information in learned representations. However, most masked autoencoder methods randomly mask the patches in sequences, resulting in meaningless regions to be masked, and weaken model's ability to learn semantic regions.

Proposed Method

The Model Overview

The illustration of proposed Language Pre-training guided Masking Representation Learning model is shown in Figure 2, which mainly consists of two modules: (i) language guided masking autoencoder module; and (ii) dual-information contrastive learning module. The former one firstly generates slice (patch) by using a fixed sliding window on original long sequential data samples, then trains a TCN (Lea et al. 2017) model for extracting patch token of each slice of time series. Subsequently, according to the label embeddings learned from pre-trained language encoder, it calculates the relationships between time series patch tokens and label embedddings, which is able to guide the selection of important patches with more semantic information. Finally, all the normal and masked time series patches are transported to an adaptive masked autoencoder for learning their latent representations. For further improving the discriminability of learned representations, we also design dual-information contrastive learning to extracting both local and global information from instance and clip data.

In the following subsections, we will provide an elaborate description of the proposed framework.

Language Guided Masking Autoencoder

This subsection will introduce the inherent mechanism of the core idea in our paper, named language guided masking autoencoder. Firstly, we will present how to generate the language label embedding which is the key indicator to guide the mask location in time series sequence. Then, we will introduce how to use language tokens to mask patches in sequences and how to enhance the discriminative power of learned representations.

Language embeddings generation. We use the label information of time series as the language text description for each data sample. Concretely, as shown in Figure 2, we directly use the text description "*This time series is about working in the garage.*" as an input, wherein the keywords "*working*" and "*garage*" are strongly associated with the class label information, the remaining words are considered as hard prompts (Liu et al. 2023a). According to the literature (Khattak et al. 2023), the pre-trained language encoder shown in Figure 2 is a CLIP text encoder that generates feature representations for text description by tokenizing the words and projecting them to word embeddings z_l^i:

$$z_l^i = \text{TextProj}(w_k^i), \quad z_l^i \in \mathbb{R}^d \tag{1}$$

where d is the dimension of each learned embedding z_l, and w_k^i denotes the output of k-th transformer layer of i-th word encoding, $\mathrm{TextProj}(\cdot)$ is able to project w_k^i into a common latent embedding space. Besides, we also choose the predicted soft labels with high confidence as the pseudo-labels for reducing the interference from incorrect label information (Zhang et al. 2021).

Language guided masking antoencoder. As shown in Figure 2, we first use a sliding windows of length l to clip the original long sequential time series data sample $\mathbf{X}$ to n subsequences. For each subsequence of $\mathbf{X}$, the i-th slice can be denoted as $x_{(j,j+l-1)}^i \in \mathbb{R}^l$ where j is the starting point timestamp. Then, we transport each subsequence into a TCN model to calculate the patches tokens of all subsequence, which can be denoted as $\mathbf{Z} \in \mathbb{R}^{n \times d}$. In order to calculate the similarity between patch token z_i and word embedding z_l^i, we set the number of neurons in last layer of TCN model to d as well. The main idea of cross patch attention shown in Figure 2 is to calculate the correlation of each word embedding with the patch tokens at each position. Therefore, the possibility of a semantic part to appearing in each position of patch token can be formulated as an attention map:

$$\mathbf{M} = \mathbf{Z}_l \otimes \mathbf{Z} := \mathrm{softmax}(\mathbf{Z}_l^T \mathbf{W}_l^T \mathbf{W} \mathbf{Z}), \tag{2}$$

where $\mathbf{M} \in \mathbb{R}^{c \times n}$ denotes the attention map of each time series data, c is the number of word embeddings, and $\mathbf{W}_l$, $\mathbf{W}$ are the projection matrixes.

After obtaining the attention map via Eq.(2), where each attention map value $m_{i,j}$ indicates the possibility of the corresponding i-th language word embedding related to j-th time series patch position. Then, we can obtain the semantic patches position p according to attention map as follow:

$$p = \arg\max_j (m_j), \quad j = 1, ..., n \tag{3}$$

where $m_j = \sum_{i=1}^{c} m_{i,j}$ is the sum of column in the attention map $\mathbf{M}$. Therefore, the patch token in p-th position contains more semantic information that strongly related to the language word embeddings, which should be masked for encouraging model to learn more class-aware information rather than meaningless background redundancy, so as to improve the discriminability of learned representations.

Subsequently, we construct an adaptive masked autoencoder to map the masked time series sequence into latent representations and then reconstructs the full instance from the latent representations. The masking strategy is designed according to above Eq. (3), and the loss function based on Mean Squared Error (MSE) metric is applied to calculate the errors of original data and reconstructed data in each timestamp. To be specific, as suggested in MAE, the loss function of proposed adaptive masked autoencoder is

$$\mathcal{L}_{recon} = \frac{1}{|\mathfrak{B}|} \sum_i \| \mathbf{m}_i \odot (\widetilde{x}_i - x_i) \|_2^2, \tag{4}$$

where $\mathbf{m}_i \in \{0, 1\} \in \mathbb{R}^n$ denotes the mask of i-th sample, where if $j \neq p$, the x_j is observed, and $m_{i,j} = 1$; otherwise x_j is missing, and $m_{i,j} = 0$. Besides, $\widetilde{x}_i$ indicates the reconstruction samples, and $|\mathfrak{B}|$ denotes the number of samples in each batch.

Dual-information Contrastive Learning

Inspired by existing contrastive learning methods for image/video (Kalantidis et al. 2020), we can conclude a fact that hard negative samples play an more important role in contrastive learning, however, hard negative time series samples are difficult to be explored. A previous technique to select hard negative sample is designed according to timestamps, namely, the fragments with small time intervals are mutually hard negative samples, while those with large time intervals are mutually simple negative samples (Hu et al. 2022). However, the content evolution rule in time series is unknown, thus it is inaccurate to judge the similarity of fragments solely based on timestamps.

To address this issue, inspired by TimesURL (Liu and Chen 2024), we define a hard negative samples generation strategy at both the clip-level and instance-level for exploring the local and global information, respectively. Given the i-th time series data in t-th timestamp that can be denoted as $x_{i,t}$, and we can obtain two augmentation of $x_{i,t}$ as $x'_{i,t}$ and $x''_{i,t}$, respectively. Then, the synthetic hard negative samples of augmentations for i-th time series at timestamp t can be calculated as

$$\begin{aligned} x_{i,t}^{\prime\text{clip}} &= \alpha \cdot x'_{i,t} + \beta \cdot x'_{i,t'}, \\ x_{i,t}^{\prime\prime\text{clip}} &= \alpha \cdot x''_{i,t} + \beta \cdot x''_{i,t}, \end{aligned} \tag{5}$$

where $t' \neq t$ is a different timestamp of another clip. Analogously, the instance-level hard negative samples of augmentations in all timestamps can be formulated as

$$\begin{aligned} x_i^{\prime\text{inst}} &= \alpha \cdot x'_i + \beta \cdot x'_j, \\ x_i^{\prime\prime\text{inst}} &= \alpha \cdot x''_i + \beta \cdot x''_j, \end{aligned} \tag{6}$$

where $j \neq i$ denotes the another time series sample in batch $\mathfrak{B}$. Besides, the parameters α and β are used to control the quality of hard negative sample generation, if the β increases, the quality of hard negative samples decreases, vice versa. Last, according to the hard negative sample generation in Eq.(5) and Eq.(6), we can present the loss function of dual-information contrastive learning for extracting both local and global information in i-th time series data as follow:

$$\mathcal{L}_{i,t}^{\text{local}} = -\log \frac{\exp(x'_{i,t}, x''_{i,t})}{\exp(x'_{i,t}, x''_{i,t}) + \sum_{q_{i,t'} \in \mathbb{N}_i} \exp(x'_{i,t}, q_{i,t'})}, \tag{7}$$

$$\mathcal{L}_i^{\text{global}} = -\log \frac{\exp(x'_i, x''_i)}{\exp(x'_i, x''_i) + \sum_{q_j \in \mathbb{N}_j} \exp(x'_i, q_j)}, \tag{8}$$

where $\mathbb{N}_i$ and $\mathbb{N}_j$ denote the set $\{x'_{i,t'}, x''_{i,t'}\} \bigcup \{x_{i,t'}^{\prime\text{clip}}, x_{i,t'}^{\prime\prime\text{clip}}\}$ and $\{x'_j, x''_j\} \bigcup \{x_j^{\prime\text{inst}}, x_j^{\prime\prime\text{inst}}\}$, respectively. Therefore, in order to capture both local and global information in time series data, the total loss function dual-information contrastive learning is

$$\mathcal{L}_{cons} = \frac{1}{|\mathfrak{B}|} \sum_i \sum_t (\mathcal{L}_{i,t}^{\text{local}} + \mathcal{L}_i^{\text{global}}). \tag{9}$$

In a word, the overall loss function of proposed language guided masking autoencoder is

$$\mathcal{L}_{total} = \mathcal{L}_{recon} + \lambda \mathcal{L}_{cons}, \tag{10}$$

where parameter λ is used to balance two loss.

Experimental Results

In this section, to demonstrate the effectiveness of LPMRL, we first perform a comparison with the baseline algorithm on 106 datasets from the UCR time series dataset. Second, we perform ablation experiments on the model to evaluate the contribution of each module within the model. Third, visualization experiments are conducted to validate the classification performance of LPMRL, focusing on its representational dimension. Finally, we assess the robustness of LPMRL through sensitivity tests.

Datasets description. Our approach was rigorously evaluated using the well-established UCR time series archive (Dau et al. 2019), a benchmark dataset in real-world scenarios. In selecting the datasets, we adhered to the methodology employed in the time series SSC work (Liu et al. 2023b), ensuring consistency in our evaluation. Out of the 128 available UCR datasets, we carefully chose 106 datasets that are representative and challenging. For our experiments, class labels were assigned to each instance, allowing us to effectively measure the model's classification performance. We randomly selected a certain ratio of the 106 datasets in which multiple algorithms were compared. Ranking by accuracy yields fair and generalizable results. Table 1 presents 10 representative UCR datasets, which cover various application scenarios and data characteristics, effectively demonstrating the applicability of our method. Notably, the UCR datasets are univariate, and due to the constraints of the text guidance, we did not select multivariate datasets in our experiments. However, the LPMRL method is applicable to multivariate datasets.

Baselines. Our method is compared with five SOTA time series classification algorithm including DTW, T-Loss (Franceschi, Dieuleveut, and Jaggi 2019), TS-TCC (Eldele et al. 2021), TST (Zerveas et al. 2021), TNC (Tonekaboni, Eytan, and Goldenberg 2021) and TS2Vec (Yue et al. 2022). Dynamic Time Warping (DTW) is an algorithm used to measure the similarity between two time series, even if they have different lengths or are shifted in time. T-Loss is a method for unsupervised time series representation learning. It learns representations of time series through a self-supervised task, which is suitable for representation learning. Time-series representation learning via temporal and contextual contrasting (TS-TCC) is a time series representation learning method that leverages both temporal and contextual contrasting. It combines temporal and contextual information, improving the representation learning of time series data. TST applies a Transformer-based framework to multivariate time series representation learning which can effectively capture long-term dependencies and interactions in multivariate time series data. TNC is an unsupervised representation learning method for time series, using temporal neighborhood coding. TNC emphasizes the temporal relationships within the data, making it robust for unsupervised learning tasks. TS2Vec aims to achieve universal representation of time series through a self-supervised learning framework. By creating multi-level temporal contrastive tasks, TS2Vec learns representations that can generalize across different time series tasks.

Dataset	# Type	# Classes	# Length	# Samples
Adiac	Image	37	176	390
CricketX	Motion	12	300	390
ElectricDevices	Device	7	96	8926
FiftyWords	Image	50	270	450
Phoneme	Sensor	39	1024	214
ShapesAll	Image	60	512	600
WordSynonyms	Image	25	270	267
Crop	Image	24	46	7200
PigArtPressure	Hemodynamics	52	2000	104
PowerCons	Power	2	144	180

Table 1: Datesets description.

Experimental setups. We adhered to the experimental setup established in TS2Vec, utilizing an SVM classifier with an RBF kernel to train on the learned representations, which allowed us to rigorously assess both the robustness and accuracy of our approach. To maintain consistency across methods, we configured the representation dimension for all classification techniques, except for DTW, to 320. Additionally, the maximum epoch was set at 500, the learning rate was configured to 1e-4, and the batch size was set to 64. These experiments were meticulously conducted in a controlled environment using PyTorch 1.10, running on two high-performance NVIDIA GeForce RTX A6000 GPUs, ensuring that our results were both reliable and reproducible.

Results Analysis

The results of the evaluation experiments for time series classification are shown in Table 2. LPMRL achieves the best classification performance on 106 UCR time series datasets with different percentages of randomly selected datasets. It achieved an average accuracy of 85.1% on all 106 datasets, and ranked the highest accuracy on 60 of the 106 datasets, surpassing the previous SOTA model.

In the baseline method, DTW is computed by the alignment method, but much effective information is missing. Although TST can effectively capture the complex relationships of time series, the method cannot be trained in large numbers for datasets with small data sizes, whereas our method performs multiple effective enhancements and still maintains excellent accuracy on datasets with small sample sizes. TS-TCC, although it also uses the contrastive learning approach, it only considers contrasts on temporal information. Our method uses the time-frequency transform to obtain information in both time and frequency domains simultaneously, which makes the model representation more powerful, and thus better results are obtained.

TNC achieves better results by capturing both local consistency and global feature representation. Whereas, our approach captures the important information of the time series locally better by the cross-attention mechanism through the linguistic level guidance, and enhances the model representation capability through the mask autoencoder (MAE). TS2Vec achieves a generalized representation of time series through a self-supervised learning framework, but it

Dataset Ratio	10%			20%			30%			40%		
	Avg.ACC	Win	P-value	Avg.ACC	Win	P-value	Avg.ACC	Win	P-value	Avg.ACC	Win	P-value
DTW	0.828 ±0.125	2	9.9E-03	0.776 ±0.155	2	4.2E-02	0.757 ±0.176	2	7.9E-02	0.773 ±0.179	4	1.2E-02
TST	0.771 ±0.160	1	1.9E-01	0.699 ±0.190	1	9.7E-01	0.683 ±0.225	1	6.8E-01	0.703 ±0.221	2	9.2E-01
TS-TCC	0.871 ±0.113	2	1.0E-04	0.816 ±0.150	5	2.7E-03	0.816 ±0.173	6	7.5E-04	0.819 ±0.178	8	1.1E-04
TNC	0.861 ±0.113	3	1.5E-04	0.806 ±0.140	3	3.2E-03	0.799 ±0.178	5	4.2E-03	0.806 ±0.181	6	5.5E-04
T-Loss	0.895 ±0.096	2	1.2E-04	0.835 ±0.136	5	2.7E-04	0.825 ±0.159	5	1.3E-04	0.824 ±0.167	5	2.4E-05
TS2Vec	0.892 ±0.087	4	4.4E-05	0.846 ±0.128	6	5.3E-05	0.839 ±0.153	10	2.0E-05	0.837 ±0.163	13	4.0E-06
LPMRL	**0.912 ±0.087**	**8**	–	**0.867 ±0.126**	**12**	–	**0.846 ±0.160**	**17**	–	**0.848 ±0.153**	**23**	–

Dataset Ratio	50%			60%			80%			100%		
	Avg.ACC	Win	P-value	Avg.ACC	Win	P-value	Avg.ACC	Win	P-value	Avg.ACC	Win	P-value
DTW	0.781 ±0.171	4	1.3E-03	0.752 ±0.174	4	2.2E-02	0.753 ±0.179	6	7.4E-03	0.747 ±0.172	6	8.6E-01
TST	0.699 ±0.217	2	9.7E-01	0.671 ±0.213	2	2.8E-01	0.666 ±0.221	4	1.5E-01	0.661 ±0.211	5	4.4E-05
TS-TCC	0.814 ±0.169	9	1.0E-05	0.791 ±0.169	12	7.5E-05	0.790 ±0.178	14	5.6E-06	0.782 ±0.181	15	7.6E-02
TNC	0.813 ±0.175	7	2.3E-05	0.787 ±0.174	8	2.1E-04	0.788 ±0.181	12	1.0E-05	0.780 ±0.179	13	9.1E-02
T-Loss	0.831 ±0.160	7	2.7E-07	0.808 ±0.159	8	1.5E-06	0.810 ±0.167	13	2.3E-06	0.816 ±0.160	17	6.1E-05
TS2Vec	0.845 ±0.156	16	1.6E-07	0.823 ±0.156	19	5.1E-06	0.825 ±0.163	20	2.8E-08	0.833 ±0.156	24	4.4E-07
LPMRL	**0.862 ±0.126**	**29**	–	**0.844 ±0.142**	**32**	–	**0.846 ±0.153**	**45**	–	**0.851 ±0.148**	**60**	–

Table 2: Time series classification results in different ratio on 106 UCR multivariate datasets. Win represents the number of datasets for which the current algorithm has reached the best accuracy. The best is in **bold**.

lacks sufficient instance information in the sample dimension. Since LPMRL uses enhancements on the sample dimension, it can better capture instance-level information, resulting in better performance. Additionally, we evaluate the significance of the test classification accuracy using Wilcoxon signed rank test (Demšar 2006). The results show that LPMRL significantly outperforms all other baseline classification algorithms.

Ablation Study

In this subsection we discuss the impact of each module on LPMRL as a whole and perform ablation experiments. We first selected 12 UCR time series datasets and then calculated the accuracy using different sub-models: (1)w/o Language Guided: we removed the cross-attention mechanism and used the self-attention mechanism to find adaptive masks; (2)w/o Contrastive Learning: we removed the contrast learning module to get the representations using only the mask encoder; (3)w/o Mask Autoencoder: we remove the mask autoencoder; (4)Random Mask: we replace the mask matrix with a random mask matrix.

As shown in Table 3, both adaptive masks and mask autocompilers can effectively improve the classification performance of the model. In particular, the language-guided adaptive coding is due to the fact that the token obtained through linguistic modality can more effectively extract the segments containing rich information under the influence of the cross-attention mechanism, which effectively enhances the performance of the mask self-encoder and improves the model's classification ability.

Method	Avg.ACC	Win	P-value
w/o Language Guided	0.851	2	1.2E-04
w/o Contrastive Learning	0.737	0	5.6E-03
w/o Mask Autoencoder	0.674	0	1.1E-02
Random Mask	0.832	0	3.9E-04
LPMRL	0.882	10	–

Table 3: Ablation study performances of proposed method and its degradation models on 12 UCR time series datasets.

Visualization Analysis

To visually demonstrate the effectiveness of LPMRL classification, we performed t-SNE visualization on multiple datasets. Figure 3 shows the visualizations on 2 UCR time series datasets, where the original data become separated after the LPMRL representation is extracted. This illustrates the effectiveness of the LPMRL model. Figure 4 highlights

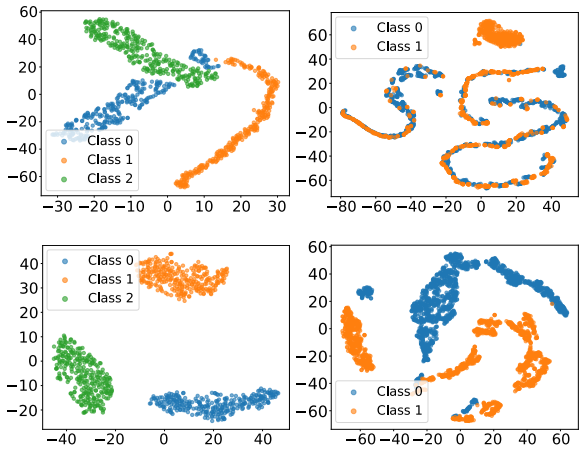


Figure 3: The t-SNE visualization on two UCR time series datasets, i.e., CBF (left column) and Freezer (right column).

the distinction between randomly generated masks and those produced by the LPMRL model. The upper section of the figure depicts masks created through language-guided masking, whereas the lower section displays masks generated randomly. Despite having the same 50% masking ratio, the LPMRL-generated masks concentrate more on the mid- and high-frequency regions of the time domain, where information is denser. By focusing on these information-rich segments, the LPMRL model produces higher-quality masks, thereby enhancing its representation performance.

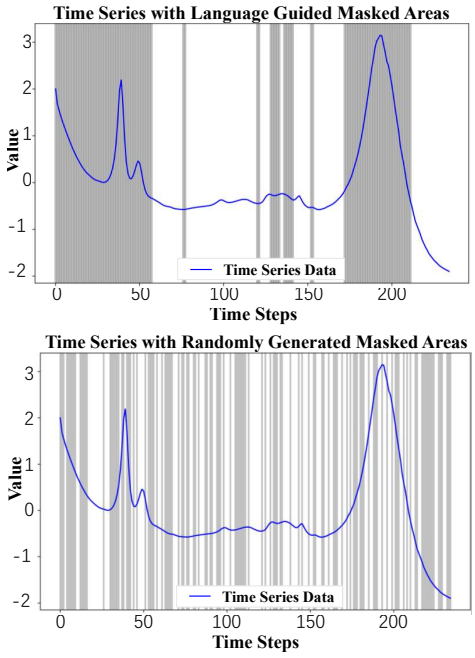
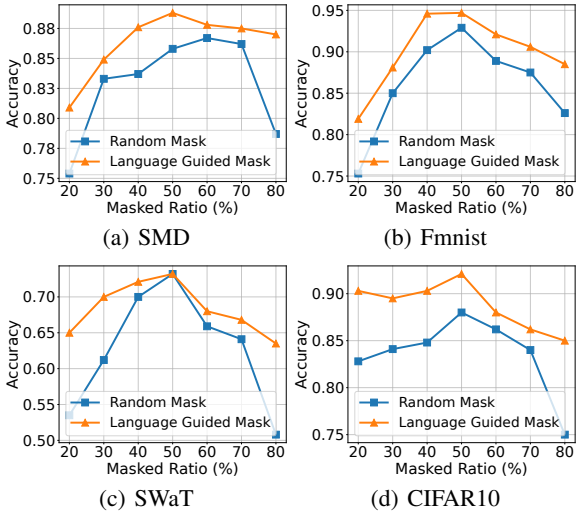


Figure 4: Comparative visualization of time series masking: Language-Guided vs. Random Generation on UCR wine dataset, with shaded areas indicating masked regions.

Sensitivity Study

In this subsection, we conduct a parameter sensitivity analysis on the LPMRL model to examine how different parameter settings impact its classification performance. We adjusted the mask ratio incrementally by 0.1 and measured the model's classification accuracy at each scale. Figure 5 illustrates the variation in classification accuracy between the random mask model and the LPMRL model across different datasets as the mask ratio increases. The results indicate that, due to the language-guided masks, the LPMRL model is significantly better at identifying and focusing on the information-rich segments, especially when the mask ratio is low, resulting in higher accuracy. As the mask ratio increases, LPMRL consistently targets the most informative regions, thereby enhancing its robustness and performance.



(a) SMD (b) Fmnist

(c) SWaT (d) CIFAR10

Figure 5: Sensitivity study of classification accuracy for Language-Guided vs. Random Masking at various mask ratios on UCR time series datasets.

Conclusion

In this paper, we proposed a novel approach to time series representation learning termed LPMRL. Our method utilizes language pre-training to guide an adaptive masking encoder, enhancing the discriminability of learned representations. This is achieved by adaptively sampling semantic spatiotemporal patches, using natural language descriptions derived from time series labels. Additionally, we incorporated a dual-information contrastive learning mechanism that captures both local and global information from long sequences in a self-supervised manner. Extensive experiments on 106 benchmark datasets confirmed that our approach consistently outperforms several state-of-the-art methods, validating the effectiveness of the proposed framework. Our following work will focus on investigating the relationship between similar segments of temporal information appearing on different time slices.

Acknowledgments

We thank Professor Eamonn Keogh and all the people who have contributed to the UCR time series classification archive. This work was supported by the National Natural Science Foundation of China under Grant 62406250 and the Fundamental Research Funds for the Central Universities, in part by the Key Laboratory of the Ministry of Education on Application of Artificial Intelligence in Equipment, Rocket Force University of Engineering.

References

Bandara, W. G. C.; Patel, N.; Gholami, A.; Nikkhah, M.; Agrawal, M.; and Patel, V. M. 2023. Adamae: Adaptive masking for efficient spatiotemporal learning with masked autoencoders. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14507–14517.

Chen, G.; Liu, M.; and Kong, Z. 2020. Temporal-logic-based semantic fault diagnosis with time-series data from industrial internet of things. *IEEE Transactions on Industrial Electronics*, 68(5): 4393–4403.

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, Z.; Wu, M.; Zhao, R.; Guretno, F.; Yan, R.; and Li, X. 2020b. Machine remaining useful life prediction via an attention-based deep learning approach. *IEEE Transactions on Industrial Electronics*, 68(3): 2521–2531.

Cheng, M.; Liu, Q.; Liu, Z.; Zhang, H.; Zhang, R.; and Chen, E. 2023. Timemae: Self-supervised representations of time series with decoupled masked autoencoders. *arXiv preprint arXiv:2303.00320*.

Chowdhury, R. R.; Zhang, X.; Shang, J.; Gupta, R. K.; and Hong, D. 2022. Tarnet: Task-aware reconstruction for time-series transformer. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 212–220.

Dau, H. A.; Bagnall, A.; Kamgar, K.; Yeh, C.-C. M.; Zhu, Y.; Gharghabi, S.; Ratanamahatana, C. A.; and Keogh, E. 2019. The UCR time series archive. *IEEE/CAA Journal of Automatica Sinica*, 6(6): 1293–1305.

Demšar, J. 2006. Statistical comparisons of classifiers over multiple data sets. *The Journal of Machine learning research*, 7: 1–30.

Dong, J.; Wu, H.; Zhang, H.; Zhang, L.; Wang, J.; and Long, M. 2024. Simmtm: A simple pre-training framework for masked time-series modeling. *Advances in Neural Information Processing Systems*, 36.

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*.

Eldele, E.; Ragab, M.; Chen, Z.; Wu, M.; Kwoh, C.-K.; Li, X.; and Guan, C. 2023. Self-supervised contrastive representation learning for semi-supervised time-series classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Franceschi, J.-Y.; Dieuleveut, A.; and Jaggi, M. 2019. Unsupervised scalable representation learning for multivariate time series. *Advances in neural information processing systems*, 32.

He, K.; Chen, X.; Xie, S.; Li, Y.; Dollár, P.; and Girshick, R. 2022. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 16000–16009.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Hu, D.; Wang, Z.; Nie, F.; Wang, R.; and Li, X. 2022. Self-Supervised Learning for Heterogeneous Audiovisual Scene Analysis. *IEEE Transactions on Multimedia*, 25: 3534–3545.

Ismail Fawaz, H.; Forestier, G.; Weber, J.; Idoumghar, L.; and Muller, P.-A. 2019. Deep learning for time series classification: a review. *Data mining and knowledge discovery*, 33(4): 917–963.

Kalantidis, Y.; Sariyildiz, M. B.; Pion, N.; Weinzaepfel, P.; and Larlus, D. 2020. Hard negative mixing for contrastive learning. *Advances in neural information processing systems*, 33: 21798–21809.

Khattak, M. U.; Rasheed, H.; Maaz, M.; Khan, S.; and Khan, F. S. 2023. Maple: Multi-modal prompt learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19113–19122.

Law, M. H.; and Jain, A. K. 2006. Incremental nonlinear dimensionality reduction by manifold learning. *IEEE transactions on pattern analysis and machine intelligence*, 28(3): 377–391.

Lea, C.; Flynn, M. D.; Vidal, R.; Reiter, A.; and Hager, G. D. 2017. Temporal convolutional networks for action segmentation and detection. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 156–165.

Li, G.; Choi, B.; Xu, J.; Bhowmick, S. S.; Chun, K.-P.; and Wong, G. L.-H. 2021. Shapenet: A shapelet-neural network approach for multivariate time series classification. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, 8375–8383.

Li, Z.; Rao, Z.; Pan, L.; Wang, P.; and Xu, Z. 2023. Ti-mae: Self-supervised masked time series autoencoders. *arXiv preprint arXiv:2301.08871*.

Liu, J.; and Chen, S. 2024. Timesurl: Self-supervised contrastive learning for universal time series representation learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 13918–13926.

Liu, P.; Yuan, W.; Fu, J.; Jiang, Z.; Hayashi, H.; and Neubig, G. 2023a. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9): 1–35.

Liu, Z.; Ma, Q.; Ma, P.; and Wang, L. 2023b. Temporal-frequency co-training for time series semi-supervised learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 8923–8931.

Liu, Z.; Pei, W.; Lan, D.; and Ma, Q. 2024. Diffusion language-shapelets for semi-supervised time-series classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 14079–14087.

Luo, D.; Cheng, W.; Wang, Y.; Xu, D.; Ni, J.; Yu, W.; Zhang, X.; Liu, Y.; Chen, Y.; Chen, H.; et al. 2023. Time series contrastive learning with information-aware augmentations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 4534–4542.

Nie, F.; Wang, Z.; Wang, R.; Wang, Z.; and Li, X. 2019. Towards Robust Discriminative Projections Learning via Non-Greedy $\ell_{2,1}$-Norm MinMax. *IEEE transactions on pattern analysis and machine intelligence*, 43(6): 2086–2100.

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.

Peng, F.; Luo, J.; Lu, X.; Wang, S.; and Li, F. 2024. Cross-Domain Contrastive Learning for Time Series Clustering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 8921–8929.

Shi, Y.; Siddharth, N.; Torr, P.; and Kosiorek, A. R. 2022. Adversarial masking for self-supervised learning. In *International Conference on Machine Learning*, 20026–20040. PMLR.

Tang, L.; Wang, Z.; He, G.; Wang, R.; and Nie, F. 2024. Perturbation Guiding Contrastive Representation Learning for Time Series Anomaly Detection. In *Proc. 33rd Int. Joint Conf. Artif. Intell*, 4955–4963.

Tang, X.; Yao, H.; Sun, Y.; Aggarwal, C.; Mitra, P.; and Wang, S. 2020. Joint modeling of local and global temporal dynamics for multivariate time series forecasting with missing values. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, 5956–5963.

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

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L. u.; and Polosukhin, I. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems*, volume 30.

Wen, Q.; Zhou, T.; Zhang, C.; Chen, W.; Ma, Z.; Yan, J.; and Sun, L. 2023. Transformers in time series: a survey. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, 6778–6786.

Xu, H.; Wang, Y.; Jian, S.; Liao, Q.; Wang, Y.; and Pang, G. 2024. Calibrated one-class classification for unsupervised time series anomaly detection. *IEEE Transactions on Knowledge and Data Engineering*.

Yue, Z.; Wang, Y.; Duan, J.; Yang, T.; Huang, C.; Tong, Y.; and Xu, B. 2022. Ts2vec: Towards universal representation of time series. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 8980–8987.

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.

Zhang, B.; Wang, Y.; Hou, W.; Wu, H.; Wang, J.; Okumura, M.; and Shinozaki, T. 2021. Flexmatch: Boosting semi-supervised learning with curriculum pseudo labeling. *Advances in Neural Information Processing Systems*, 34: 18408–18419.