TimesURL: Self-Supervised Contrastive Learning for Universal Time Series Representation Learning

Jiexi Liu[1,2], **Songcan Chen**[1,2*]

[1]College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics
[2]MIIT Key Laboratory of Pattern Analysis and Machine Intelligence
{liujiexi, s.chen}@nuaa.edu.cn

Abstract

Learning universal time series representations applicable to various types of downstream tasks is challenging but valuable in real applications. Recently, researchers have attempted to leverage the success of self-supervised contrastive learning (SSCL) in Computer Vision(CV) and Natural Language Processing(NLP) to tackle time series representation. Nevertheless, due to the special temporal characteristics, relying solely on empirical guidance from other domains may be ineffective for time series and difficult to adapt to multiple downstream tasks. To this end, we review three parts involved in SSCL including 1) designing augmentation methods for positive pairs, 2) constructing (hard) negative pairs, and 3) designing SSCL loss. For 1) and 2), we find that unsuitable positive and negative pair construction may introduce inappropriate inductive biases, which neither preserve temporal properties nor provide sufficient discriminative features. For 3), just exploring segment- or instance-level semantics information is not enough for learning universal representation. To remedy the above issues, we propose a novel self-supervised framework named TimesURL. Specifically, we first introduce a frequency-temporal-based augmentation to keep the temporal property unchanged. And then, we construct double Universums as a special kind of hard negative to guide better contrastive learning. Additionally, we introduce time reconstruction as a joint optimization objective with contrastive learning to capture both segment-level and instance-level information. As a result, TimesURL can learn high-quality universal representations and achieve state-of-the-art performance in 6 different downstream tasks, including short- and long-term forecasting, imputation, classification, anomaly detection and transfer learning.

Introduction

Time series data is ubiquitous in reality ranging from weather and economics to transportation (Wu et al. 2021; Liu et al. 2022b; Shi et al. 2015). Learning information-rich and universal time series representations for multi-type downstream tasks is a fundamental but unsolved problem. While self-supervised contrastive learning has exhibited great success in computer vision (CV), natural language processing (NLP), and recently, other types of modalities (Denton et al. 2017; Gutmann and Hyvärinen 2012; Wang and Gupta 2015; Pagliardini, Gupta, and Jaggi 2017; Chen et al. 2020), its application to time series requires tailored solutions. This is due to the high dimensionality and special temporal characteristics of time series data, as well as the need for diverse semantics information for different tasks.

To this end, we review the four main parts involved in SSCL including 1) augmentation method for positive samples designing, 2) backbone encoder, 3) (hard) negative pairs, and 4) SSCL loss for pretext tasks, and try to invest efforts to explore more effective solutions for time series feature capturing in universal representation learning. Since the backbone encoder has been extensively studied in time series encoder learning (Liu and Chen 2019; Zhou et al. 2021; Wu et al. 2023; Liu et al. 2022a), our attention is primarily directed toward the remaining three components.

First, most augmentation methods, when applied to time series data, may introduce inappropriate inductive biases as they directly borrow ideas from the fields of CV and NLP. For example, *Flipping* (Luo et al. 2023) flip the sign of the original time series that assumes the time series has symmetry between up and down directions. Nevertheless, this may ruin the temporal variations, such as trend, and peak valley, that are inherently present in the original time series. While *permutation* (Um et al. 2017) rearranges the order of segments in a time series to generate a new series, under the assumption that the underlying semantic information remains unchanged by the different orders. However, this disturbs the temporal dependencies, thereby impacting the relationships between past and future timestamp information. Consequently, since valuable semantic information of time series primarily resides in temporal variations and dependencies, such augmentations are unable to capture the appropriate features necessary for effective universal representation learning.

Then, the importance of hard negative sample selection has been proved in other domains (Kalantidis et al. 2020; Robinson et al. 2020), but is still underexplored in time series literature. Due to the local smoothness and Markov property, most time series segments can be considered as easy negative samples. These segments tend to exhibit semantic dissimilarity with the anchor and contribute only minor gradients, thus failing to provide useful discriminative information (Cai et al. 2020). Although the inclusion of a

*Corresponding Author

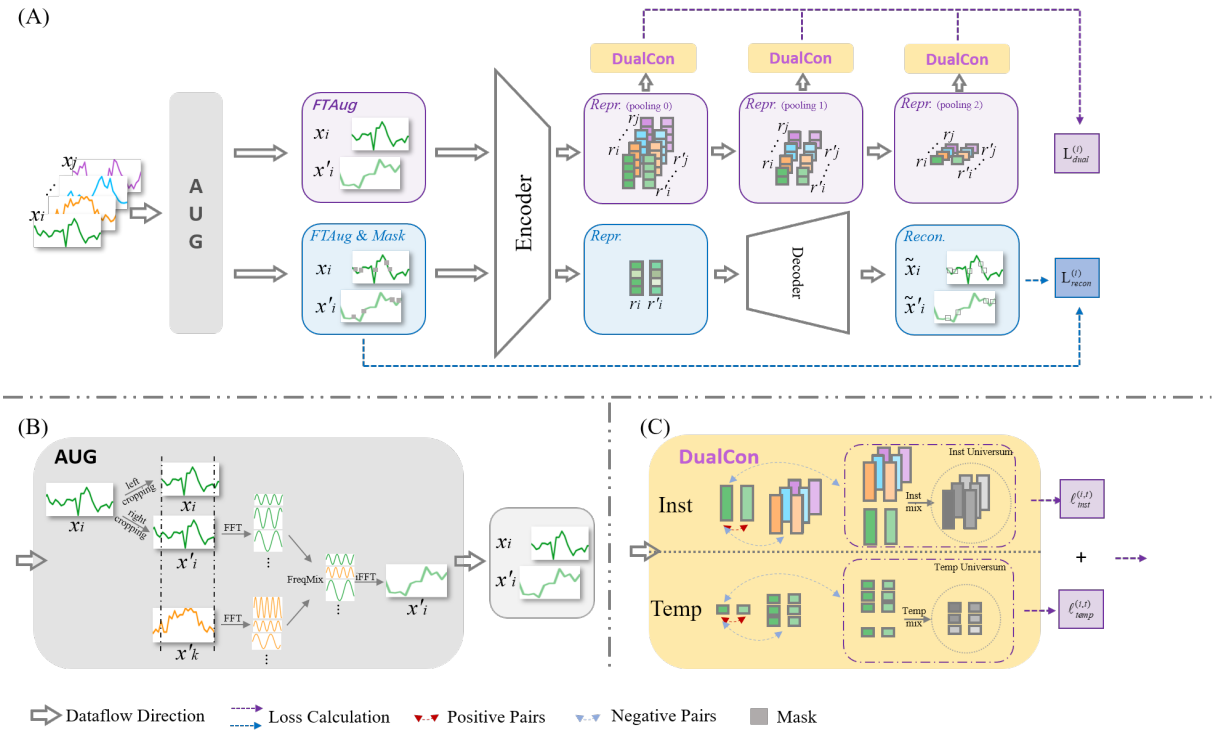


Figure 1: Overview of TimesURL approach, shown in (A), consisting of FTAug, DualCon, and Recon three components. A time series x_i is transformed into two augmented series x_i and x_i' by cropping and frequency mix. Then, the corresponding representation r_i and r_i', the colorful pieces in the rectangular box marked by Repr., are extracted by the Encoder. Within each pooling, shown in light purple rectangular boxes, the learned representations are fed into the DualCon component to synthesize the temporal- and instance-wise Universums, thereby injecting them into contrastive learning. The light blue rectangular boxes represent the reconstruction data flow. Subfigures (B) and (C) denote the specific process of FTAug and Universum synthetic.

small number of hard negative samples, which have similar but not identical semantics to the anchor, has shown to facilitate improved and expedited learning (Xu et al. 2022; Cai et al. 2020), their effectiveness is overshadowed by the abundance of easy negative samples.

Last but not least, only using information at segment- or instance-level alone is not enough for learning a universal representation. Prior research has generally classified the aforementioned tasks into two categories (Yue et al. 2022). The first category includes forecasting, anomaly detection, and imputation that rely more on fine-grained information captured in segment level (Yue et al. 2022; Woo et al. 2022; Luo et al. 2023) as these tasks require inferring specific timestamps or sub-sequences. While the second category consists of classification and clustering that prioritize instance-level information, i.e. coarse-grained information (Eldele et al. 2021, 2022; Liu and wei Liu 2022), aiming to infer the target across the entire series. Therefore, when confronted with a task-agnostic pre-training model that lacks prior knowledge or awareness of specific tasks during the pre-training phase, both segment- and instance-level information become indispensable for achieving effective universal time series representation learning.

To address these challenges, in this paper, we propose a novel self-supervised framework termed **TimesURL** to learn universal representations capable of effectively supporting various downstream tasks. We first conduct instance-wise and temporal contrastive learning to incorporate temporal variations and sample diversity. Specifically, to maintain the temporal variations and dependencies, we design a new frequency-temporal-based augmentation method called FTAug which is a combination of cropping in the time domain and frequency mixing in the frequency domain. Moreover, inspired by the concept of learning through contradiction, we elaborately design double Universums as hard negative samples. It is a kind of anchor-specific mixup in the embedding space that mixup the specific positive sample (anchor) each time with a negative sample. Our designed double Universums are generated on instance-wise and temporal dimensions respectively, serving as special high-quality hard negative samples that boost the performance of contrastive learning. Additionally, we observe that contrastive learning alone is limited to capturing only one level of information. Therefore, in our paper, we jointly optimize contrastive learning and time reconstruction to capture and leverage information at both segment- and instance levels.

Benefitting from the aforementioned designs, TimesURL consistently achieves state-of-the-art (SOTA) results across a broad range of downstream tasks, thereby demonstrating its ability to learn universal and high-quality representations

for time series data. The contributions of this work can be summarized as follows:

- We revisit the existing contrastive learning framework for time series representation and propose TimesURL, a novel framework that can capture both segment- and instance-level information for universal representation with an additional time reconstruction module.
- We introduce a new frequency-temporal-based augmentation method and inject novel double Universums into contrastive learning to remedy the positive and negative pairs construction problems.
- We evaluate the performance of representation learned by TimesURL via 6 benchmark time series tasks with about 15 baselines. The consistent SOTA performance proves the universality of representation.

Related Work

Unsupervised Representation Learning for Time Series. Representation learning for time series has been well-studied for years (Chung et al. 2015; Krishnan, Shalit, and Sontag 2017; Bayer et al. 2020). However, there is still a dearth of research focusing on the more challenging aspect of unsupervised representation learning. SPIRAL (Lei et al. 2019) bridges the gap between time series data and static clustering algorithms by learning a feature representation that effectively preserves the pairwise similarities inherent in the raw time series data. TimeNet (Malhotra et al. 2017) is a recurrent neural network that trains the encoder-decoder pair to minimize the reconstruction error from its learned representations. DTCR (Ma et al. 2019) integrates the temporal reconstruction and K-means objective into the seq2seq model to learn cluster-specific temporal representations. ROCKET (Dempster, Petitjean, and Webb 2020) is a classification method with small computational expense and fast speed that transforms time series using random convolutional kernels and uses the transformed features to train a linear classifier. Therefore, numerous previous studies concentrate on developing encoder-decoder architectures to minimize reconstruction errors for unsupervised time series representation learning. Some (Lei et al. 2019; Ma et al. 2019) have attempted to leverage the inherent correlations present in time series data, but have fallen short of fully realizing time series data potential.

Time-Series Contrastive Learning. Self-supervised contrastive learning intends to learn invariant representations from different augmented views of data. It is another type of representation learning method for unannotated data over designed pretext tasks. TS-TCC (Eldele et al. 2021) focuses on designing a challenging pretext task for robust representation learning from time series data. It tackles this by designing a tough cross-view prediction task against perturbations introduced by different timestamps and augmentations. TNC (Tonekaboni, Eytan, and Goldenberg 2021) discusses the choice of positive and negative pair construction by a novel neighborhood-based method for nonstationary multivariate time series with sample weight adjustment. InfoTS (Luo et al. 2023) highlights the importance of selecting appropriate augmentations and designs an automatically selecting augmentation method with meta-learning to prevent introducing prefabricated knowledge. TS2Vec(Yue et al. 2022) is a unified framework that learns contextual representations for arbitrary sub-series at various semantic levels. CoST(Woo et al. 2022) contributes to pretext task design by leveraging inductive biases in the model architecture. It specifically focuses on learning disentangled seasonal and trend representations and incorporates a novel frequency domain contrastive loss to encourage discriminative seasonal representations. However, they are prone to be affected by improper prior assumptions, an overabundance of easy negative samples, and a lack of sufficient information for downstream tasks. These limitations arise from inappropriate augmentation methods, the lack of hard negative samples, and the neglect of leveraging both segment- and instance-level information. In this paper, we address all these problems in a unified framework for universal representation learning for time series.

Proposed TimesURL Framework

In this section, we make an elaborate description of the newly designed frame, TimesURL. We first formulate the representation learning problem and subsequently delve into the implementation of the key components including contrastive learning and time reconstruction. Particularly, within the contrastive learning component, we emphasize our designed augmentation and double Universum synthesizing methods.

Problem Formulation. Similar to most time series representation learning methods, our goal is to learn a nonlinear embedding function f_θ, such that each instance x_i in time series set $\mathcal{X} = \{x_1, x_2, \ldots, x_N\}$ can map to the best described representation r_i. Each input time series instance is $x_i \in \mathbb{R}^{T\times F}$, where T is the time series length and F is the feature dimension. The representation for the i-th time series is $r_i = \{r_{i,1}, r_{i,2}, \ldots, r_{i,T}\}$, in which $r_{i,t} \in \mathbb{R}^K$ is the representation vector at time t, where K is the dimension of representation vector. Since our model is a two-step progress, we then use the learned representation to accomplish the downstream tasks.

Method Introduction. As shown in Figure 1, we first generate augmentation sets $\mathcal{X}'$ and $\mathcal{X}'_M$ through FTAug for original series $\mathcal{X}$ and masked series $\mathcal{X}_M$, respectively. Then we get two pairs of original and augmentation series sets, the first pair $(\mathcal{X}, \mathcal{X}')$ is for contrastive learning, while the second pair $(\mathcal{X}_M, \mathcal{X}'_M)$ is for time reconstruction. After that, we map the above sets with f_θ to achieve corresponding representations. We encourage $\mathcal{R}$ and $\mathcal{R}'$ to have transformation consistency and design a reconstruction method to precisely recover the original dataset $\mathcal{X}$ using both $\mathcal{R}_M$ and $\mathcal{R}'_M$.

The effectiveness of the model above is guaranteed by 1) using a suitable augmentation method for positive pair construction, 2) having a certain amount of hard negative samples for model generalization, and 3) optimizing the encoder f_θ by contrastive learning and time reconstruction losses jointly for capturing both levels of information. We will then discuss the three parts in the following subsections.

FTAug Method

A key component of contrastive learning is to choose appropriate augmentations that can impose some priors to construct feasible positive samples so that encoders can be trained to learn robust and discriminative representations (Chen et al. 2020; Grill et al. 2020; Yue et al. 2022). Most augmentation strategies are task-dependent (Luo et al. 2023) and may introduce strong assumptions of data distribution. More seriously, they may perturb the temporal relationship and semantic consistency that is crucial for tasks like forecasting. Therefore, we choose the contextual consistency strategy (Yue et al. 2022), which treats the representations at the same timestamp in two augmented contexts as positive pairs. Our FTAug combines the advantages in both frequency and temporal domains that generate the augmented contexts by frequency mixing and random cropping.

Frequency mixing. Frequency mixing is used to produce a new context view by replacing a certain rate of the frequency components in one training instance x_i calculated by Fast Fourier Transform (FFT) operation with the same frequency components of another random training instance x_k in the same batch (Chen et al. 2023). Then we use the inverse FFT to convert back to get a new time domain time series. Exchanging frequency components between samples will not introduce unexpected noise or artificial periodicities and can offer more reliable augmentations for preserving the semantic characteristics of the data.

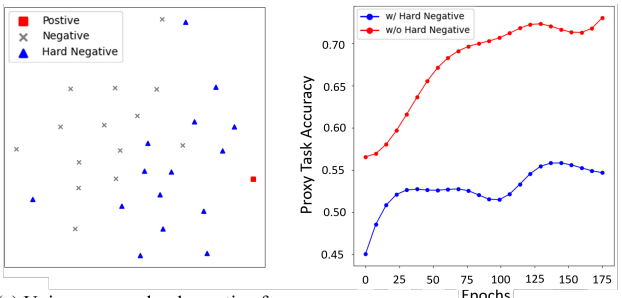
Random cropping. Random cropping is the key step for contextual consistency strategy. For each instance x_i, we randomly sample two overlapping time segments $[a_1, b_1]$, $[a_2, b_2]$ such that $0 < a_1 \leq a_2 \leq b_1 \leq b_2 \leq T$. The contrastive learning and time reconstruction further optimize the representation in the overlapping segment $[a_2, b_1]$.

Ultimately, the proposed FTAug is helpful for various kinds of tasks since it can keep the important temporal relationship and semantic consistency for time series. Here, the FTAug is only applied in the training process.

Double Universum Learning

As revealed by recent studies (Kalantidis et al. 2020; Robinson et al. 2020; Cai et al. 2020), hard negative samples play an important role in contrastive learning but have never been explored in the time series domain. Moreover, due to the local smoothness and the Markov property in time series, most negative samples are easy that are insufficient for capturing temporal-wise information since they fundamentally lack the learning signals required to drive contrastive learning. As a real example of ERing dataset in UEA archive (Bagnall et al. 2018) in Figure 2, for each positive anchor (red square), the corresponding negative samples (gray marks) contain many easy negatives and few hard ones, i.e. many of the negatives are too far to contribute to the contrastive loss.

Our double Universums are Mixup Induced Universums (Han and Chen 2023; Vapnik 2006; Chapelle et al. 2007) in both instance- and temporal-wise, which is anchor-specific mixing in the embedding space that mixes the specific positive feature (anchor) with the negative features for unannotated datasets.



(a) Universums as hard negative for each positive (anchor) sample

(b) Accuracy on the proxy task

Figure 2: Properties of Universums on ERing dataset in UEA archive

Let i be the index of the input time series sample and t be the timestamp. $r_{i,t}$ and $r'_{i,t}$ denote the representations for the same timestamp t but from two augmentation of x_i. The synthetic temporal-wise Universums for the i-th time series at timestamp t can be formulated as

$$\begin{aligned} r^{\text{temp}}_{i,t} &= \lambda_1 \cdot r_{i,t} + (1-\lambda_1) \cdot r_{i,t'}, \\ r'^{\text{temp}}_{i,t} &= \lambda_1 \cdot r'_{i,t} + (1-\lambda_1) \cdot r'_{i,t'}, \end{aligned} \tag{1}$$

in which t' is randomly chosen from Ω, the set of timestamps within the overlap of the two subseries, and $t' \neq t$. While the instance-wise Universums indexed with (i, t) are similar be formulated as

$$\begin{aligned} r^{\text{inst}}_{i,t} &= \lambda_2 \cdot r_{i,t} + (1-\lambda_2) \cdot r_{j,t}, \\ r'^{\text{inst}}_{i,t} &= \lambda_2 \cdot r'_{i,t} + (1-\lambda_2) \cdot r'_{j,t}, \end{aligned} \tag{2}$$

where j indicates any other instance except i in batch $\mathcal{B}$. Here, $\lambda_1, \lambda_2 \in (0, 0.5]$ are randomly chosen mixing coefficients for the anchor, and $\lambda_1, \lambda_2 \leq 0.5$ guarantees that the anchor's contribution is always smaller than negative samples.

As in Figure 2(a), most Universum (blue triangles) are much closer to the anchor and thus can be seen as hard negative samples. Moreover, we utilize a proxy task to indicate the difficulty of hard negatives (Kalantidis et al. 2020), i.e. Universums. The proxy task performance is shown in Figure 2(b), i.e. the percentage of anchors where the positive sample is ranked overall negatives across training our TimesURL with and without Universums on ERing dataset. Despite the drop in proxy task performance of TimesURL, however, further performance gains are observed for linear classification from 0.896 (without Universums) to 0.985 (with Universums), which means that the additional Universum makes the proxy task harder to solve but can further improve the model performance in the downstream task. Therefore, Univerums in TimesURL can be seen as high-quality negatives. To sum up, our Universums can be treated as a kind of high-quality hard negative samples.

By mixing with the anchor sample, the possibility of the universum data falling into target regions in the data space is minimized, thereby ensuring the hard negativity of Universum. Moreover, the double Universum set contains all other negative samples that are beneficial to learning discriminative sample information to increase model capability.

Contrastive Learning for Segment-level Information

We use a straightforward way to inject the double Universums into contrastive learning as additional hard negative samples in temporal- and instance-wise contrastive loss, respectively. The two losses for the i-th time series at timestamp t can be formulated as

$$\ell_{\text{temp}}^{(i,t)} = -\log \frac{\exp(r_{i,t} \cdot r'_{i,t})}{\exp\left(r_{i,t} \cdot r'_{i,t}\right) + \sum_{z_{i,t'} \in \mathbb{N}_i} \exp\left(r_{i,t} \cdot z_{i,t'}\right)} \tag{3}$$

$$\ell_{\text{inst}}^{(i,t)} = -\log \frac{\exp(r_{i,t} \cdot r'_{i,t})}{\exp(r_{i,t} \cdot r'_{i,t}) + \sum_{z_{j,t} \in \mathbb{N}_j} \exp\left(r_{i,t} \cdot z_{j,t}\right)} \tag{4}$$

where in Eq.(3) and (4), $\mathbb{N}_i \triangleq \mathbb{Z}_i \cup \mathbb{Z}'_i \cup \mathbb{U}_i \cup \mathbb{U}'_i$, and $\mathbb{N}_j \triangleq \mathbb{Z}_j \cup \mathbb{Z}'_j \cup \mathbb{U}_j \cup \mathbb{U}'_j$, in which $\mathbb{Z}_i \cup \mathbb{Z}'_i = \{r_{i,t'}, r'_{i,t'} | t' \in \Omega \backslash t\}$ and $\mathbb{Z}_j \cup \mathbb{Z}'_j = \{r_{j,t}, r'_{j,t} | j \in \{1, \ldots, |\mathcal{B}|\} \backslash i\}$ are original negative samples, while $\mathbb{U}_i \cup \mathbb{U}'_i = \{r_{i,t'}^{\text{temp}}, r_{i,t'}^{\prime\text{temp}} | t' \in \Omega\}$ and $\mathbb{U}_j \cup \mathbb{U}'_j = \{r_{j,t}^{\text{inst}}, r_{j,t}^{\prime\text{inst}} | j \in \{1, \ldots, |\mathcal{B}|\}$ are proposed double Universums as Eq.(1),(2), where $|\mathcal{B}|$ denotes the batch size.

The two losses are complementary to each other to capture both instance-specific characteristics and the temporal variation. We use hierarchical contrastive loss (Yue et al. 2022) for multi-scale information learning by using max pooling on the learned representations along the time axis in Eq.(3) and (4). Here, we have to mention that important temporal variation information, such as trend and seasonal are lost after several max pooling operations, therefore contrasting at top levels cannot actually capture sufficient instance-level information for downstream tasks.

$$\mathcal{L}_{\text{dual}} = \frac{1}{|\mathcal{B}|T} \sum_i \sum_t \left(\ell_{\text{temp}}^{(i,t)} + \ell_{\text{inst}}^{(i,t)}\right) \tag{5}$$

Time Reconstruction for Instance-level Information

The masked autoencoding technique in self-supervised learning has been proven to perform well in various domains, such as BERT-based pre-training model(Kenton and Toutanova 2019) in NLP as well as MAE (He et al. 2022) in CV. The main idea of such methods is to reconstruct the original signal given its partial observation.

Motivated by the masked autoencoding technique, we design a reconstruction module to preserve important temporal variation information. Our approach uses the above mentioned embedding function f_θ as an encoder that maps the masked instance into latent representation and then reconstructs the full instance from the latent representation. Here, we use the random masking strategy. Our loss function computes the Mean Squared Error (MSE) between the reconstructed and the original value at each timestamp. Further, similar to BERT and MAE, we compute the MSE loss only on the masked timestamps in Eq.(6).

$$\mathcal{L}_{\text{recon}} = \frac{1}{2|\mathcal{B}|} \sum_i \|m_i \odot (\tilde{x}_i - x_i)\|_2^2 + \|m'_i \odot (\tilde{x}'_i - x'_i)\|_2^2 \tag{6}$$

Here, we denote $m_i \in \{0,1\}^{T \times F}$ as the observation mask for the i-th instance where $m_{i,t} = 0$ if $x_{i,t}$ is missing, and $m_{i,t} = 1$ if $x_{i,t}$ is observed, while $\tilde{x}_i$ is the generated reconstruction instance. Similar to the above notations, m'_i, $\tilde{x}'_i$ and x'_i have the same meaning.

The overall loss is defined as

$$\mathcal{L} = \mathcal{L}_{\text{dual}} + \alpha \mathcal{L}_{\text{recon}} \tag{7}$$

where α is a hyper-parameter to balance the two losses.

Experiments

In this section, to evaluate the generality and the downstream tasks performance of the representation learned by our TimesURL, we extensively experiment on 6 downstream tasks, including short- and long-term forecasting, imputation, classification, anomaly detection and transfer learning. The best results are highlighted in bold. More detailed experimental setups and other additional experiment results will be presented in the Appendix.

Implementation The summary of the benchmarks is in the Appendix. For TimesURL, we use Temporal Convolution Network (TCN) as the backbone encoder, which is similar to TS2Vec (Yue et al. 2022). More detailed information about the dataset and other experiment implementation information is in the Appendix.

Baselines Following the self-supervised learning setting, we extensively compare TimesURL with recent advanced models under a similar experimental setup. Since most existing self-supervised learning methods cannot learn universal representations for all kinds of tasks, we utilize each method only for tasks it is specifically designed for. Moreover, we also compare the SOTA models for each specific task as follows, where SSL, E2EL, USL are abbreviations for self-supervised, end-to-end, and unsupervised learning: ***Forecasting***: 1) **SSL**: CoST (Woo et al. 2022), TS2Vec (Yue et al. 2022), TNC (Tonekaboni, Eytan, and Goldenberg 2021), 2) **E2EL**: Informer (Zhou et al. 2021), LogTrans (Li et al. 2019), N-BEATS (Oreshkin et al. 2019); ***Classification***: 1) **SSL**: InfoTS (Luo et al. 2023), TS2Vec, TS-TCC (Eldele et al. 2021), TST (Zerveas et al. 2021),2) **USL**: DTW; ***Imputation***: **SSL**: TS2Vec, InfoTS; ***Anomaly detection***: 1) **SSL**: TS2Vec, 2) **USL**: SPOT(Siffer et al. 2017), DSPOT(Siffer et al. 2017), DONUT(Xu et al. 2018), SR(Ren et al. 2019).

Overall, about 15 baselines are included for a comprehensive comparison.

Classification

Setups Time series classification has practical significance in medical diagnosis, action recognition, etc. Our experiments are under the setting that class labels are on the instance. So instance-level classification is adopted to verify the model capacity in presentation learning. We select commonly used UEA (Bagnall et al. 2018) and UCR (Dau et al. 2019) Classification Archive. The representation dimensions of all classification methods except DTW are set to 320 and we then follow the same protocol as TS2Vec which uses an SVM classifier with RBF kernel to train on top of representations for classification.

Method		TimesURL	InfoTS	TS2Vec	T-Loss	TNC	TS-TCC	TST	DTW
30 UEA datasets	Avg. ACC	**0.752 (+3.8%)**	0.714	0.704	0.658	0.670	0.668	0.617	0.629
	Avg. Rank	**1.367**	3.200	3.567	4.567	5.333	5.000	5.900	5.207
128 UCR datasets	Avg. ACC	**0.845 (+0.7%)**	0.838	0.836	0.806	0.761	0.757	0.639	0.729
	Avg. Rank	**1.844**	2.047	2.625	4.248	5.128	5.032	6.961	6.008

Table 1: Time series classification results.

Results The evaluation results are shown in Table 1. TimesURL achieves the best performance with an average accuracy of 75.2% for 30 univariate datasets in UEA and 84.5% for 128 multivariate datasets in UCR, surpassing the previous SOTA self-supervised method InfoTS (71.4%). Moreover, the best average rank also validates the significant outperformance of TimesURL. As mentioned before, it is easy to understand the failure of other methods, since TS2Vec lacks sufficient instance-level information, while some augmentations in InfoTS may introduce inappropriate inductive biases that damage the temporal properties, such as the trend for classification. Since TimesURL uses general FTAug, contains appropriate hard negatives and can capture both segment- and instance-level information, thus achieves better performance.

Imputation

Setups Under a realistic scenario, irregular and asynchronized sampling often happens, which may lead to missingness resulting in difficulty in downstream tasks. Imputation is a straightforward and widely used method to relieve this problem. We complete the task with ETT dataset (Zhou et al. 2021) under the electricity scenario, where the data-missing problem happens commonly. To compare the model capacity under different proportions of missing data, we randomly mask the time points in the ratio of $\{12.5\%, 25\%, 37.5\%, 50\%\}$. We then follow the same setting as TimesNet which uses a MLP network for the downstream tasks.

		TimesURL		InfoTS		TS2Vec	
Metrics		MSE	MAE	MSE	MAE	MSE	MAE
ETTh1	0.125	**0.659**	**0.640**	0.717	0.666	0.690	0.658
	0.250	**0.679**	**0.648**	0.726	0.674	0.710	0.668
	0.375	**0.702**	**0.656**	0.726	0.676	0.728	0.676
	0.500	**0.712**	0.693	0.783	0.695	0.751	**0.682**
ETTh2	0.125	**2.455**	1.215	2.491	**1.199**	2.866	1.288
	0.250	**2.560**	**1.239**	2.644	1.244	2.792	1.271
	0.375	**2.673**	1.269	2.757	**1.266**	2.793	1.271
	0.500	**2.701**	1.281	2.844	1.283	2.769	**1.267**
ETTm1	0.125	**0.644**	**0.650**	0.702	0.651	0.726	0.663
	0.250	**0.699**	0.668	0.732	0.664	0.719	**0.664**
	0.375	**0.718**	0.686	0.753	0.674	0.728	**0.670**
	0.500	**0.716**	0.680	0.759	0.677	0.739	**0.669**
Avg.		**1.326**	**0.860**	1.386	0.864	1.418	0.871

Table 2: Multivarite time series imputation results.

Results Since TimesURL contains a time construction module to capture underlying temporal patterns, we naturally extend it to a downstream task. As shown in Table 2, our proposed TimesNet still achieves SOTA performance on three datasets and proves to have the ability to capture temporal variation from complicated time series.

Short- and Long-Term Forecasting

Setups Time series forecasting is ubiquitous in our everyday life. For both short- and long-term forecasting we use ETT, Electricity and Weather datasets from various reality scenarios and the results of the two latter datasets are in the Appendix. For short-term forecasting, the horizon is 24 and 48, while for long-term forecasting the horizon ranges from 96 to 720. Here, we learn the representations once for each dataset and can be directly applied to various horizons with linear regressions. This helps demonstrate the universality of the learned representation.

Results We compare TimesURL with not only representation learning as well as end-to-end forecasting methods in 3 and indicate that TimesURL has established a new SOTA in most cases for both short- and long-term forecasting.

Anomaly Detection

Setups Detecting anomalies from monitoring data is essential for industrial maintenance. We follow the setting of a streaming evaluation protocol (Ren et al. 2019) in time series anomaly detection that determines whether the last point x_t in time series slice $x_1, \ldots, x_t$ is an anomaly or not. During training, each time series sample is split into two halves according to the time order, where the first half is for training and the second is for evaluation. In this task, We compare models on two benchmark datasets, including KPI (Ren et al. 2019) a competition dataset that includes multiple minutely sampled real KPI curves and Yahoo (Nikolay Laptev 2015) including 367 hourly sampled time series.

Results Table 4 shows the performance of anomaly detection tasks with different methods on F1 score, precision and recall. In the normal setting, TimesURL has consistently good performance on both KPI and Yahoo datasets.

Transfer Learning

We complete the transfer learning task to demonstrate that the representation learned by TimesURL has good transferability that can achieve good performance when training on one condition (i.e., source domain) and testing it on other multiple conditions (i.e., target domains). Here, we present the transfer learning results achieved by training the model on two separate source domains, namely

Methods		Representation Learning								End-to-end Forecasting					
		TimesURL		CoST		TS2Vec		TNC		Informer		LogTrans		N-BEATS	
Metrics		MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
ETTh1	24	**0.036**	**0.142**	0.040	0.152	0.039	0.151	0.057	0.184	0.098	0.247	0.103	0.259	0.094	0.238
	48	**0.056**	**0.146**	0.060	0.186	0.062	0.189	0.094	0.239	0.158	0.319	0.167	0.328	0.210	0.367
	168	**0.096**	**0.233**	0.097	0.236	0.142	0.291	0.171	0.329	0.183	0.346	0.207	0.375	0.232	0.391
	336	0.121	0.267	**0.112**	**0.258**	0.160	0.316	0.192	0.357	0.222	0.387	0.230	0.398	0.232	0.388
	720	**0.145**	0.3068	0.148	**0.306**	0.179	0.345	0.235	0.408	0.269	0.435	0.273	0.463	0.322	0.490
ETTh2	24	0.083	0.219	**0.0790**	**0.207**	0.091	0.230	0.097	0.238	0.093	0.240	0.102	0.255	0.198	0.345
	48	**0.116**	**0.219**	0.1180	0.259	0.124	0.274	0.131	0.281	0.155	0.314	0.169	0.348	0.234	0.386
	168	**0.175**	**0.332**	0.1890	0.339	0.198	0.355	0.197	0.354	0.232	0.389	0.246	0.422	0.331	0.453
	336	**0.188**	**0.347**	0.2060	0.360	0.205	0.364	0.207	0.366	0.263	0.417	0.267	0.437	0.431	0.508
	720	**0.186**	**0.352**	0.2140	0.371	0.208	0.371	0.207	0.370	0.277	0.431	0.303	0.493	0.437	0.517
ETTm1	24	**0.013**	**0.084**	0.015	0.088	0.016	0.093	0.019	0.103	0.030	0.137	0.065	0.202	0.054	0.184
	48	**0.024**	0.1765	0.025	**0.117**	0.028	0.126	0.036	0.142	0.069	0.203	0.078	0.220	0.190	0.361
	96	**0.037**	**0.145**	0.038	0.147	0.045	0.162	0.054	0.178	0.194	0.372	0.199	0.386	0.183	0.353
	288	0.080	0.214	**0.077**	**0.209**	0.095	0.235	0.098	0.244	0.401	0.554	0.411	0.572	0.186	0.362
	672	0.114	**0.255**	**0.113**	0.257	0.142	0.290	0.136	0.290	0.512	0.644	0.598	0.702	0.197	0.368
Avg.		**0.098**	**0.229**	0.102	0.233	0.116	0.253	0.129	0.272	0.210	0.362	0.228	0.391	0.235	0.381

Table 3: Short- and Long-Term Forecasting Univariate forecasting results.

	Yahoo			KPI		
	F_1	Prec.	Rec.	F_1	Prec.	Rec.
SPOT	0.338	0.269	0.454	0.217	0.786	0.126
DSPOT	0.316	0.241	0.458	0.521	0.623	0.447
DONUT	0.026	0.013	0.825	0.347	0.371	0.326
SR	0.563	0.451	0.747	0.622	0.647	0.598
TS2Vec	0.745	0.729	0.762	0.677	0.929	0.533
TimesURL	**0.749**	0.748	0.750	**0.688**	0.925	0.546

Table 4: Univariate time series anomaly detection results.

	Avg. Accuracy
TimesURL	**0.752**
w/o Frequency Mixing	0.709 (-4.3%)
w/o Instance Universum	0.720 (-3.2%)
w/o Temporal Universum	0.717 (-3.5%)
w/o Double Universum	0.716 (-3.6%)
w/o Time Reconstruction	0.735 (-1.8%)

Table 5: Ablation results on 30 UEA datasets.

CBF and CinCECGTorso in the UCR archive and evaluate performance on the downstream classification task across other 9 target domains in the first 10 datasets in the UCR archive. The average results are 0.864 for CBF and 0.895 for CinCECGTorso for the transfer scenario. The transformation results show competitive performance with no transfer scenario. More transfer learning results are in the Appendix.

Ablation Study

We emphasize the importance of FTAug, Double Universums and joint optimization strategies for learning universal representations, respectively. To verify the effectiveness of the above three modules in TimesURL, a comparison between full TimesURL and its five variants on 30 datasets in the UEA archive is shown in Table 5, where 1) w/o frequency mixing, 2) w/o instance Universum, 3) w/o temporal Universum, 4) w/o double Universum, 5) w/o time reconstruction. Results show that all the above components of TimesURL are indispensable. We have to mention that constructing either temporal- or instance-wise Universums cannot achieve optimal performance, while double Universums achieve better performance by providing sufficient and discriminative information for both temporal- and instance-wise contrastive learning.

Conclusion

In this paper, we propose a novel self-supervised framework termed TimesURL that can learn universal time series representations for various types of downstream tasks. We introduce a new augmentation method called FTAug to keep contextual consistency and temporal characteristics unchanged, which is suitable for various downstream tasks. Moreover, we inject double Universums into contrastive learning to enhance negative sample quantity and quality to boost the performance of contrastive learning. Last but not least, TimesURL jointly optimizes contrastive learning and time reconstruction for capturing both segment- and instance-levels of information for universal representation learning. Experimental results demonstrate the effectiveness of the above strategies and show that with suitable augmentation methods, enough hard negative samples and proper levels of information, TimesURL shows great performance on six downstream tasks.

Acknowledgments

The authors wish to thank all the donors of the original datasets, and everyone who provided feedback on this work. Specially, the authors wish to thank Meng Cao for assisting with the implementation of the code, Xiang Li and Jiaqiang Zhang for proofreading this manuscript. This work is supported by the Key Program of NSFC under Grant No.62076124, Postgraduate Research & Practice Innovation Program of Jiangsu Province under Grant No.KYCX21_0225 and Scientific and Technological Achievements Transferring Project of Jiangsu Province under Grant No. BA2021005.

References

Bagnall, A.; Dau, H. A.; Lines, J.; Flynn, M.; Large, J.; Bostrom, A.; Southam, P.; and Keogh, E. 2018. The UEA multivariate time series classification archive, 2018. *arXiv preprint arXiv:1811.00075*.

Bayer, J.; Soelch, M.; Mirchev, A.; Kayalibay, B.; and van der Smagt, P. 2020. Mind the Gap when Conditioning Amortised Inference in Sequential Latent-Variable Models. In *International Conference on Learning Representations*.

Cai, T.; Frankle, J.; Schwab, D. J.; and Morcos, A. S. 2020. Are all negatives created equal in contrastive instance discrimination? *ArXiv*, abs/2010.06682.

Chapelle, O.; Agarwal, A.; Sinz, F.; and Schölkopf, B. 2007. An analysis of inference with the universum. *Advances in neural information processing systems*, 20.

Chen, M.; Xu, Z.; Zeng, A.; and Xu, Q. 2023. FrAug: Frequency Domain Augmentation for Time Series Forecasting. *arXiv preprint arXiv:2302.09292*.

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

Chung, J.; Kastner, K.; Dinh, L.; Goel, K.; Courville, A. C.; and Bengio, Y. 2015. A recurrent latent variable model for sequential data. *Advances in neural information processing systems*, 28.

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.

Dempster, A.; Petitjean, F.; and Webb, G. I. 2020. ROCKET: exceptionally fast and accurate time series classification using random convolutional kernels. *Data Mining and Knowledge Discovery*, 34(5): 1454–1495.

Denton, E. L.; et al. 2017. Unsupervised learning of disentangled representations from video. *Advances in neural information processing systems*, 30.

Eldele, E.; Ragab, M.; Chen, Z.; Wu, M.; Kwoh, C.; Li, X.; and Guan, C. 2021. Time-Series Representation Learning via Temporal and Contextual Contrasting. In *International Joint Conference on Artificial Intelligence*.

Eldele, E.; Ragab, M.; Chen, Z.; Wu, M.; Kwoh, C. K.; Li, X.; and Guan, C. 2022. Self-supervised Contrastive Representation Learning for Semi-supervised Time-Series Classification. *arXiv preprint arXiv:2208.06616*.

Grill, J.-B.; Strub, F.; Altché, F.; Tallec, C.; Richemond, P.; Buchatskaya, E.; Doersch, C.; Avila Pires, B.; Guo, Z.; Gheshlaghi Azar, M.; et al. 2020. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33: 21271–21284.

Gutmann, M. U.; and Hyvärinen, A. 2012. Noise-Contrastive Estimation of Unnormalized Statistical Models, with Applications to Natural Image Statistics. *Journal of machine learning research*, 13(2).

Han, A.; and Chen, S. 2023. Universum-Inspired Supervised Contrastive Learning. In *Web and Big Data: 6th International Joint Conference, APWeb-WAIM 2022, Nanjing, China, November 25–27, 2022, Proceedings, Part II*, 459–473. Springer.

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.

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.

Kenton, J. D. M.-W. C.; and Toutanova, L. K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of NAACL-HLT*, 4171–4186.

Krishnan, R.; Shalit, U.; and Sontag, D. 2017. Structured inference networks for nonlinear state space models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31.

Lei, Q.; Yi, J.; Vaculin, R.; Wu, L.; and Dhillon, I. S. 2019. Similarity Preserving Representation Learning for Time Series Clustering. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, IJCAI'19, 2845–2851. AAAI Press. ISBN 9780999241141.

Li, S.; Jin, X.; Xuan, Y.; Zhou, X.; Chen, W.; Wang, Y.-X.; and Yan, X. 2019. Enhancing the locality and breaking the memory bottleneck of transformer on time series forecasting. *Advances in neural information processing systems*, 32.

Liu, J.; and Chen, S. 2019. Non-stationary multivariate time series prediction with selective recurrent neural networks. In *Pacific rim international conference on artificial intelligence*, 636–649. Springer.

Liu, M.; Zeng, A.; Chen, M.; Xu, Z.; Lai, Q.; Ma, L.; and Xu, Q. 2022a. Scinet: Time series modeling and forecasting with sample convolution and interaction. *Advances in Neural Information Processing Systems*, 35: 5816–5828.

Liu, Y.; and wei Liu, J. 2022. The Time-Sequence Prediction via Temporal and Contextual Contrastive Representation Learning. In *Pacific Rim International Conference on Artificial Intelligence*.

Liu, Y.; Wu, H.; Wang, J.; and Long, M. 2022b. Non-stationary Transformers: Exploring the Stationarity in Time Series Forecasting. In *Advances in Neural Information Processing Systems*.

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*.

Ma, Q.; Zheng, J.; Li, S.; and Cottrell, G. W. 2019. Learning representations for time series clustering. *Advances in neural information processing systems*, 32.

Malhotra, P.; TV, V.; Vig, L.; Agarwal, P.; and Shroff, G. 2017. TimeNet: Pre-trained deep recurrent neural network for time series classification. *arXiv preprint arXiv:1706.08838*.

Nikolay Laptev, Y. B., Saeed Amizadeh. 2015. A Benchmark Dataset for Time Series Anomaly Detection. https://yahooresearch.tumblr.com/post/114590420346/a-benchmark-dataset-for-time-series-anomaly.

Oreshkin, B. N.; Carpov, D.; Chapados, N.; and Bengio, Y. 2019. N-BEATS: Neural basis expansion analysis for interpretable time series forecasting. *arXiv preprint arXiv:1905.10437*.

Pagliardini, M.; Gupta, P.; and Jaggi, M. 2017. Unsupervised learning of sentence embeddings using compositional n-gram features. *arXiv preprint arXiv:1703.02507*.

Ren, H.; Xu, B.; Wang, Y.; Yi, C.; Huang, C.; Kou, X.; Xing, T.; Yang, M.; Tong, J.; and Zhang, Q. 2019. Time-series anomaly detection service at microsoft. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, 3009–3017.

Robinson, J.; Chuang, C.-Y.; Sra, S.; and Jegelka, S. 2020. Contrastive learning with hard negative samples. *arXiv preprint arXiv:2010.04592*.

Shi, X.; Chen, Z.; Wang, H.; Yeung, D.-Y.; Wong, W.-K.; and Woo, W.-c. 2015. Convolutional LSTM network: A machine learning approach for precipitation nowcasting. *Advances in neural information processing systems*, 28.

Siffer, A.; Fouque, P.-A.; Termier, A.; and Largouet, C. 2017. Anomaly detection in streams with extreme value theory. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1067–1075.

Tonekaboni, S.; Eytan, D.; and Goldenberg, A. 2021. Unsupervised Representation Learning for Time Series with Temporal Neighborhood Coding. In *International Conference on Learning Representations*.

Um, T. T.; Pfister, F. M.; Pichler, D.; Endo, S.; Lang, M.; Hirche, S.; Fietzek, U.; and Kulić, D. 2017. Data augmentation of wearable sensor data for parkinson's disease monitoring using convolutional neural networks. In *Proceedings of the 19th ACM international conference on multimodal interaction*, 216–220.

Vapnik, V. 2006. Transductive Inference and Semi-Supervised Learning. *Semi-Supervised Learning*, 453–472.

Wang, X.; and Gupta, A. 2015. Unsupervised learning of visual representations using videos. In *Proceedings of the IEEE international conference on computer vision*, 2794–2802.

Woo, G.; Liu, C.; Sahoo, D.; Kumar, A.; and Hoi, S. 2022. CoST: Contrastive Learning of Disentangled Seasonal-Trend Representations for Time Series Forecasting. In *International Conference on Learning Representations*.

Wu, H.; Hu, T.; Liu, Y.; Zhou, H.; Wang, J.; and Long, M. 2023. TimesNet: Temporal 2D-Variation Modeling for General Time Series Analysis. In *The Eleventh International Conference on Learning Representations*.

Wu, H.; Xu, J.; Wang, J.; and Long, M. 2021. Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting. *Advances in Neural Information Processing Systems*, 34: 22419–22430.

Xu, H.; Chen, W.; Zhao, N.; Li, Z.; Bu, J.; Li, Z.; Liu, Y.; Zhao, Y.; Pei, D.; Feng, Y.; et al. 2018. Unsupervised anomaly detection via variational auto-encoder for seasonal kpis in web applications. In *Proceedings of the 2018 world wide web conference*, 187–196.

Xu, L.; Lian, J.; Zhao, W. X.; Gong, M.; Shou, L.; Jiang, D.; Xie, X.; and Wen, J.-R. 2022. Negative sampling for contrastive representation learning: A review. *arXiv preprint arXiv:2206.00212*.

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.

Zhou, H.; Zhang, S.; Peng, J.; Zhang, S.; Li, J.; Xiong, H.; and Zhang, W. 2021. Informer: Beyond efficient transformer for long sequence time-series forecasting. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, 11106–11115.