Pattern Recognition 169 (2026) 111928

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

Generalizing across non-stationary series via learning dynamic causal factors

Weifeng Zhang [a], Yan Liu [a], Xovee Xu [a], Fan Zhou [a,b,*], Ting Zhong [a], Kunpeng Zhang [c]

[a] University of Electronic Science and Technology of China, Chengdu, 610054, Sichuan, China
[b] Key Laboratory of Intelligent Digital Media Technology of Sichuan Province, Chengdu, 610054, Sichuan, China
[c] University of Maryland, College Park, 20742, MD, USA

ARTICLE INFO

Keywords:
Domain generalization
Non-stationary time series
Neural networks

ABSTRACT

Learning domain-invariant representations is a crucial task for achieving *out-of-distribution generalization*. Recent efforts have begun to incorporate causality into this process, aiming to identify and understand the *causal factors* relevant to various tasks. However, when confronted with non-stationary time series data, simply extending existing generalization methods may prove ineffective. This inadequacy stems from their failure to adequately model the underlying causal factors, exacerbated by *temporal domain shifts* in addition to source domain shifts. In this paper, we thoroughly examine the challenges posed by both source and temporal shifts through a causal lens in the context of generalizing non-stationary time series data. We introduce a novel model called the Dynamic Causal Sequential Variational Auto-Encoder (DCSVAE), designed specifically to learn dynamic causal factors. By effectively disentangling the representation of non-stationary time series data, our model distinguishes between dynamic causal, dynamic non-causal, and static non-causal factors, thereby facilitating temporal generalization. To enhance disentanglement, we introduce two constraints on latent variables based on mutual information. Theoretical guarantees rooted in information theory validate the feasibility of our approach. Our experiments, conducted on both synthetic and real datasets, demonstrate the superior performance of the proposed model in time series domain generalization tasks when compared to state-of-the-art benchmarks.

1. Introduction

1.1. Motivation

Many machine learning paradigms often fail to generalize well when training and test datasets do not comply with the conventional i.i.d. assumption [1], which is also known as *out-of-distribution (OOD) generalization*. This is often caused by overreliance on relations among features rather than causation [2]. To address this problem, recent studies have started paying attention to invariant causal representation [3,4]. They regard OOD generalization as a task aiming to extract invariant representation across domains, which has a great impact on various downstream applications in computer vision [5] and natural language processing [6]. However, the performance of above methods usually drops significantly when it comes to time series data.

The phenomenon of *temporal shifts* (cf. Section 3) is ubiquitously observed given the real world is non-stationary and constantly evolving. For example, in the clinical context [7], mortality might decrease with the improvement of critical care. This is also true in linguistic settings [8], where the content and styles of conversations change over time [9,10]. In addition, the *source shifts* may still be unavoidable due to the nature of data. For example, data might be collected from multiple heterogeneous sources. Different samples might be measured by different devices. These two latent distribution shifts make typical generalization methods fail to extract invariants from time series [11,12]. Thus, seeking new methods to tackle both shifts in a unified model is called for.

1.2. Research gaps

Recent works have attempted to understand generalization from the perspective of causality [13,14]. Inspired by this, we intend to model underlying causal factors in time series data upon which the invariant representation can be extracted and adapted to unseen domains. Note that non-causal factors always exist that can affect the data generation process. To explain this, we take the football classification task provided by [11] as an example. In this example, each instance is a video clip where sportsmen perform actions. The actions vary with time, e.g., running, receiving, or kicking (repeatedly) occurs at different

* Corresponding author at: University of Electronic Science and Technology of China, Chengdu, 610054, Sichuan, China.
E-mail addresses: weifzh@outlook.com (W. Zhang), yan.liu@std.uestc.edu.cn (Y. Liu), xovee@std.uestc.edu.cn (X. Xu), fan.zhou@uestc.edu.cn (F. Zhou), zhongting@uestc.edu.cn (T. Zhong), kpzhang@umd.edu (K. Zhang).

https://doi.org/10.1016/j.patcog.2025.111928
Received 22 October 2024; Received in revised form 16 April 2025; Accepted 28 May 2025
Available online 16 June 2025

time points. This indicates the existence of temporal shifts. The order of actions can also be completely different across videos. It is also possible that some videos are recorded by one device while others are from different devices. This suggests that source domain shifts might occur. In order to precisely classify the sport, the model needs to recognize factors that affect both the video itself and the outcome (i.e., the prediction label). These are what we call dynamic causal factors. In addition, there are two kinds of non-causal factors that are only related to the data (i.e., videos): the static one, e.g., the costumes of players, and the dynamic one, e.g., the changing viewpoints. To sum up, it is desirable to disentangle the representation into these three factors discussed above in time series tasks.

However, existing domain generalization methods for time series fail to explicitly address temporal and source shifts simultaneously. Additionally, they often rely on an explicit time index or a sequence of labels, which may not always be available. In this paper, we propose a novel model, namely *Dynamic Causal Sequential Variational Auto-Encoder* (DCSVAE), to tackle the challenge of generalizing non-stationary time series. Specifically, DCSVAE is a deep generative model designed to learn and disentangle three latent factors: dynamic causal, dynamic non-causal, and static non-causal components. To promote disentanglement, we align the model objective with mutual information principles. Furthermore, we provide theoretical guarantees based on information theory. Furthermore, we demonstrate that our model can generalize better than existing domain generalization methods when dealing with non-stationary time series. We validate DCSVAE on both synthetic and real data, and the results show that our model outperforms the state-of-the-art methods with effective disentanglement. Note that our model not only generalizes well on the test domain but also on the training domain [15].

1.3. Objectives

Suppose we are given training data $\mathcal{D}^{tr} = \{(\mathbf{x}^i_{1:T}, \mathbf{y}^i_{1:T})\}_{i=1}^N$, where $\mathbf{x}^i_{1:T} \in \mathcal{X} \subset \mathbb{R}^{T\times D}$ is the ith non-stationary time series, and $\mathbf{y}_i \in \mathcal{Y} \subset \mathbb{R}^{T\times C}$ is the corresponding labels, where T denotes the length, D is the input dimension for each time point and C is the number of classes. We denote $P^{tr}(\mathbf{x}, \mathbf{y})$ and $P^{te}(\mathbf{x}, \mathbf{y})$ as the distribution of training set and test set respectively, and there exists two shifts as discussed before (cf. Definitions 1 and 2). Our goal is to train a model $h : \mathcal{X} \to \mathcal{Y}$ to minimize the risk on an unseen but related target domain $\mathcal{D}^{te}$: $R_{\mathcal{D}^{te}}(h) = \mathbb{E}_{(\mathbf{x},\mathbf{y})\sim P^{te}}[\ell(h(\mathbf{x}), \mathbf{y})]$, where $\ell : \mathcal{Y}\times\mathcal{Y} \to \mathbb{R}_+$ is a loss function.

1.4. Key contributions

The contributions made by this study are summarized as follows:

- To our knowledge, we are among the first to analyze both source and temporal shifts in a causal view for non-stationary time series generalization tasks.
- We propose DCSVAE, a novel framework designed to effectively disentangle the representation of non-stationary time series data into dynamic causal, dynamic non-causal, and static non-causal factors, enabling improved temporal generalization. To the best of our knowledge, this is the first work in the field of time series domain generalization.
- To encourage the disentanglement, we construct a new objective combining evidence lower bound (ELBO) with constraints based on mutual information. More interestingly, our proposed model can be used as a feature extractor and provide off-the-shelf domain generalization methods with the generalization ability for non-stationary time series data.
- The performance on both synthetic and real datasets along with two model selection methods shows the superiority of our model and its ability to learn dynamic causal factors effectively.

2. A causal view of non-stationary series generalization

To demonstrate the challenges of non-stationary time series generalization and justify the necessity of specific settings in this work, we take a simple yet illustrative example from the temporal colored MNIST (TCMNIST) dataset [12], shown in Fig. 2. It is an extension of colored MNIST [16], which converts a static dataset into a time series one. The goal is to predict the parity (even or odd) of the sum of the current and the last frames under the following two distribution shifts One unique characteristic intentionally designed in TCMNIST is the spurious correlation between color and label, i.e., the pairs $(green, odd)$ and $(red, even)$.

To illustrate the generalization problem of nonstationary time series better, we define two kinds of distributional shifts separately as below:

Definition 1 (*Source Shift*)**.** Let $P^a(\mathbf{x}_{1:T}, \mathbf{y}_{1:T})$ and $P^b(\mathbf{x}_{1:T}, \mathbf{y}_{1:T})$ be two distributions from two sources $\mathcal{D}^a$ and $\mathcal{D}^b$ where $\mathbf{x}_{1:T}$ is a time series and $\mathbf{y}_{1:T}$ denotes the corresponding label, if there exists source domain shift between $\mathcal{D}^a$ and $\mathcal{D}^b$, then we have $P^a(\mathbf{x}_{1:T}, \mathbf{y}_{1:T}) \neq P^b(\mathbf{x}_{1:T}, \mathbf{y}_{1:T})$.

Definition 2 (*Temporal Shift*)**.** Let $\mathbf{x}_{1:T}$ and $\mathbf{y}_{1:T}$ be a time series and corresponding label respectively, if there exists temporal shift within the time series, then we have $P(\mathbf{x}_i, \mathbf{y}_i) \neq P(\mathbf{x}_j, \mathbf{y}_j)$, $\exists i, j \in [1, T]$.

Source shift is a common phenomenon when it comes to non-temporal data such as images. Apparently, it can occur across time series. Source shift indicates that the distributions might be different across sources while staying the same within each source, although we may not know which domain/source the given time series belongs to. As shown in the left part of Fig. 2, instances in different rows might be collected from different sources. The correlation between color and label decreases from 90% at the top to 10% at the bottom. This spurious correlation is not a stable factor for label prediction even it may perform well when it is high. From a causal perspective, recent studies attribute the correlation between input and label to the common causes as shown in Fig. 1(a) [3,17], where $\mathbf{z}^c$ denotes the *causal factors* that are domain invariant, and $\mathbf{z}^n$ denotes the non-causal factors which vary with domains. For the setting of stationary series, the distributions of causal factors could change with time while non-causal factors do not for the purpose of satisfying the stationary property. Thus, we need to explicitly model the dynamic causal factors and static non-causal factors as shown in Fig. 1(b). Note that we do not exclude the situation where all causal factors are the same within a time series.

Temporal shift is ubiquitous in non-stationary time series, where the statistical property and distribution change continuously over time [18] as shown in the right part of Fig. 2. Recent works [19,20] notice this shift and propose to tackle it for evolving domain transfer. More recently, Lu et al. have brought temporal shifts into time series generalization formulation [21], which is similar to our paper. It is noted that methods of modeling temporal shifts are quite different, and they just model temporal shifts with discrete sub-domains. We argue that the continuous view is more natural, as well as separating sub-domains is not trivial.

Given the existence of two domain shifts and efforts made by researchers, two questions naturally arise.

Q1: How can we model non-stationary time series for generalization? If there only exist source shifts, one could treat it as a general domain generalization task, and extract factors that are invariant to domains as shown in Fig. 1(a). For adapting to time series, temporal information should be considered as shown in Fig. 1(b). When time series is not stationary, dynamic non-causal factors can be accounted for the temporal shifts. To this end, we propose a novel graphical model for modeling non-stationary time series (cf. Fig. 1(d)), aiming to learn the dynamic causal factors which account for both two shifts. Specifically, $\mathbf{z}^c_{1:T}$, $\mathbf{z}^n_{1:T}$ and $\mathbf{z}^s$ denote the dynamic causal, dynamic non-causal, and static non-causal factors, respectively, which correspond to the digits

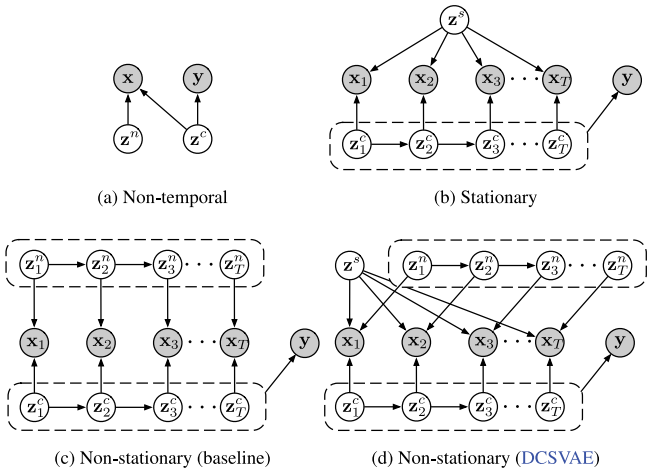


Fig. 1. Comparison of causal graphs for different settings: non-temporal, stationary, non-stationary time series (baseline and ours). Gray and white nodes denote observed and unobserved variables, respectively. (a) is the causal graph for non-temporal data, where $\mathbf{z}^c$, $\mathbf{z}^n$ and $\mathbf{y}$ are the causal factors, non-causal factors and label(s), respectively. (c) is the causal graph for the baseline setting on non-stationary time series. (b) and (d) are the causal graphs for stationary and non-stationary time series (ours), where $\mathbf{z}^c_{1:T}$, $\mathbf{z}^n_{1:T}$, $\mathbf{z}^s$ and $\mathbf{y}$ denote the dynamic causal, dynamic non-causal, static non-causal factors, and a series of labels, respectively.

shapes, colors, and background colors in Fig. 2. The $\mathbf{z}^s$ accounts for source shifts like general DG methods [3,17] without considering time-varying distribution shifts, and the $\mathbf{z}^n_{1:T}$ accounts for temporal shifts which is similar to recent works [19]. Note that if a given series is stationary, $\mathbf{z}^n_{1:T}$ would be the same for every timestamp, thus our proposed causal graph will degenerate to Fig. 1(b), showing its compatibility to stationary and non-stationary series. In Section 4, we present our proposed deep generative network to model the aforementioned three factors and optimize it with theoretical guarantees.

Q2: Can we directly use off-the-shelf domain generalization methods for non-stationary series? In this paper, we focus on domain generalization across non-stationary time series. More precisely, both source shifts and temporal shifts occur simultaneously, which is termed as *mixed shifts* in this paper. These two shifts can be roughly regarded as the distribution shifts originated from different sources and times. Though there are few studies about time series domain generalization, many efforts have been taken into domain generalization for source shifts in non-temporal data, such as IRM [16] and VREx [22]. These are model-agnostic and aim to find domain invariant representations, ignoring the differences between source and temporal shifts. Thus, they mix up the dynamic and static non-causal factors (cf. Fig. 1(c)). According to the above discussion, it is inappropriate to tackle non-stationary time series generalization with off-the-shelf methods. Therefore, we propose a novel generative model aiming to model non-stationary time series (cf. Fig. 1(d)).

3. Related work

3.1. Domain generalization and causality

Domain generalization aims to learn a model which can fit the data well in an unseen target domain [23,24]. Recent research efforts have brought causality into OOD generalization tasks [25]. Guided by the invariance principle of causality [2], a stream of literature focus on learning causal factors to represent the causality via the certain objective function or the generative process. IRM [16] and its extension [26] divided the observations into upstream causal and downstream non-causal factors with respect to labels, and proposed an objective minimizing the differences across environments via carefully

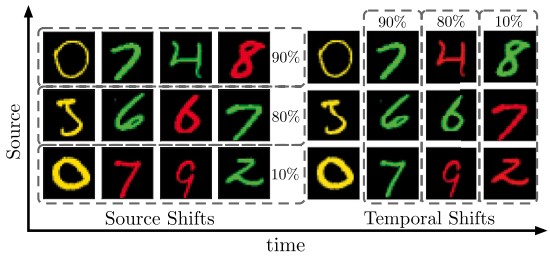


Fig. 2. An illustration example of source shifts (left) and temporal shifts (right) in non-stationary time series. The percentage denoting the spurious correlation between the color and the label varies with the source and the time, respectively. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

designed penalty terms. Recently, deep generative models have been taken into consideration for modeling causal and non-causal factors. DIVA [27] proposed a generative model by disentangling latent representations conditioned by domains and labels, which learned three disentangled latent factors given different source domains. Causal-HMM [28] combined the causal graph based on expertise and hidden Markov model for time series forecasting, showing the generalization ability when the distribution of data (e.g., the gender proportion) changes.

Though above methods try to model causal factors explicitly, they all ignore the non-stationary property which is ubiquitous for time series in real world. More recently, temporal generalization has arisen attention [29,30]. They both model the temporal shift with time-sensitive parameters. However, existing work does not consider the temporal shift in a causality view. To address the above issues, we propose a novel model based on a new causal graph which is reasonable and effective for non-stationary time series. Specifically, our proposed model aims to learn dynamic causal factors rather than static ones in previous work.

3.2. Disentangling time series

Disentangled representation is to learn several independent factors in latent space for data modeling, which is similar to human cognition, e.g., the Beuchet Chair illusion [31]. To do this, various generative models based on VAE and GAN have been proposed and achieved significant performance improvement for stationary data. For non-stationary time series, disentangling time-dependent and time-invariant features explicitly was found to be effective in controlled generation [32] based on sequential generative models. However, Locatello et al. [33] pointed out that it is impossible to ensure disentanglement without supervision. Therefore, self-supervised and contrastive learning are taken into consideration for introducing inductive bias [34,35]. More recently, LSSAE [19] leveraged the disentanglement for evolving domain generalization. Specifically, it aims to learn two latent factors related to inputs and labels which accounts for covariate and concept shifts.

Note that our work is similar to LSSAE [19], where we both model the dynamic and static factors for time series in the OOD setting. However, we explicitly model the dynamic causal factors rather than static factors in [19]. Our DCSVAE is also related to Causal-HMM in its efforts to learn dynamic causal factors for time series data. Essentially, DCSVAE does not need any expertise about the target task, while the causal graph in Causal-HMM is carefully designed based on expertise.

4. Dynamic causal sequential variational auto-encoder

4.1. Proposed model

Priors. In our model, there are three latent factors $\mathbf{z}^c_{1:T}$, $\mathbf{z}^n_{1:T}$, $\mathbf{z}^s$ denoting dynamic causal, dynamic non-causal, and static non-causal factors,

respectively. The joint prior distribution can be factorized as

$$p_\theta(\mathbf{z}^c_{1:T}, \mathbf{z}^n_{1:T}, \mathbf{z}^s) = p_\theta(\mathbf{z}^s) \prod_{t=1}^{T} p_\theta(\mathbf{z}^c_t|\mathbf{z}^c_{<t}) p_\theta(\mathbf{z}^n_t|\mathbf{z}^n_{<t}). \tag{1}$$

The priors of dynamic factors are defined as sequential priors $p_\theta(\mathbf{z}^c_t|\mathbf{z}^c_{<t}) = \mathcal{N}(\boldsymbol{\mu}(\mathbf{z}^c_{<t}), \boldsymbol{\sigma}(\mathbf{z}^c_{<t}))$ and $p_\theta(\mathbf{z}^n_t|\mathbf{z}^n_{<t}) = \mathcal{N}(\boldsymbol{\mu}(\mathbf{z}^n_{<t}), \boldsymbol{\sigma}(\mathbf{z}^n_{<t}))$, which can be parameterized by recurrent neural networks such as LSTM and GRU. And the prior of static factors is defined as a standard Gaussian distribution $p_\theta(\mathbf{z}^s) = \mathcal{N}(\mathbf{0}, \mathbf{I})$.

To model the dynamic causal factors in non-stationary time series, we define a probabilistic generative model for the joint distribution over observed and latent variables based on the causal graph as shown in Fig. 1(d). It can be factorized due to $\{\mathbf{x}_{1:T}, \mathbf{z}^s, \mathbf{z}^n_{1:T}\} \perp \mathbf{y}_{1:T}|\mathbf{z}^c_{1:T}$:

$$p_\theta(\mathbf{x}_{1:T}, \mathbf{y}_{1:T}, \mathbf{z}^c_{1:T}, \mathbf{z}^n_{1:T}, \mathbf{z}^s) = \underbrace{p_\theta(\mathbf{y}_{1:T}|\mathbf{z}^c_{1:T})}_{\text{prediction}} \underbrace{p_\theta(\mathbf{x}_{1:T}, \mathbf{z}^c_{1:T}, \mathbf{z}^n_{1:T}, \mathbf{z}^s)}_{\text{generation}}, \tag{2}$$

where the first term denotes the predictive process from dynamic causal factors $\mathbf{z}^c_{1:T}$. The causal factors learning was used by recent domain generalization works [3,17]. The second term denotes the generative process for time series which can be factorized by Markov chain as:

$$p_\theta(\mathbf{x}_{1:T}, \mathbf{z}^c_{1:T}, \mathbf{z}^n_{1:T}, \mathbf{z}^s) = p_\theta(\mathbf{z}^s) \prod_{t=1}^{T} p_\theta(\mathbf{z}^c_t|\mathbf{z}^c_{<t}) p_\theta(\mathbf{z}^n_t|\mathbf{z}^n_{<t}) p_\theta(\mathbf{x}_t|\mathbf{z}^c_t, \mathbf{z}^n_t, \mathbf{z}^s), \tag{3}$$

where the generation process $p_\theta(\mathbf{x}_t|\cdot)$ from latent variables can be implemented by a flexible function, e.g., a deconvolutional network for images or multi-layer perceptrons (MLP) for others.

Inference of dynamic causal factors. Our model exploits variational inference to learn an approximate posterior of three latent factors $q_{\boldsymbol{\phi}}$ given observed data. We train it with VAE. The objective function of latent factor learning can be optimized by maximizing the logarithm likelihood as:

$$\max_{\theta} \mathbb{E}_{P^{tr}} \left[p_\theta(\mathbf{x}_{1:T}, \mathbf{y}_{1:T}) \right]. \tag{4}$$

Following the past work towards time series generation [32,34], the posterior distribution $q_{\boldsymbol{\phi}}$ over latent variables can be factorized in two structures with respect to dynamic and static factors, i.e., full and factorized structures, whether inferring dynamic factors through static one or not specifically. Regarding our proposed method, we should consider not only dynamic and static relations but also causal and non-causal relations. The causal and non-causal information would be entangled if we adopt the full structure. Moreover, we also employ factorized structure when inferring dynamic and static factors for better disentanglement [34]. Therefore, we have the factorization of the inference model as follows:

$$q_{\boldsymbol{\phi}}(\mathbf{z}^c_{1:T}, \mathbf{z}^n_{1:T}, \mathbf{z}^s|\mathbf{x}_{1:T}) = q_{\boldsymbol{\phi}}(\mathbf{z}^s|\mathbf{x}_{1:T}) \prod_{t=1}^{T} q_{\boldsymbol{\phi}}(\mathbf{z}^c_t|\mathbf{x}_{<t}) q_{\boldsymbol{\phi}}(\mathbf{z}^n_t|\mathbf{x}_{<t}), \tag{5}$$

where the $q_{\boldsymbol{\phi}}(\mathbf{z}^s|\mathbf{x}_{1:T})$, $q_{\boldsymbol{\phi}}(\mathbf{z}^c_t|\mathbf{x}_{<t})$ and $q_{\boldsymbol{\phi}}(\mathbf{z}^n_t|\mathbf{x}_{<t})$ are all Gaussian distributions parameterized by sequential models. Specifically, we employ the bi-directional LSTM for $q_{\boldsymbol{\phi}}(\mathbf{z}^s|\mathbf{x}_{1:T})$ to model static information of the whole input sequence.

Theorem 1. *Combined with above causal graph of non-stationary time series as shown in Fig. 1(d), the evidence lower bound of likelihood (cf. Eq. (4)) in DCSVAE is:*

$$\begin{aligned}\mathcal{L}_{ELBO} = {} & \mathbb{E}_{P^{tr}}\left[\mathbb{E}\left[\log p_\theta(\mathbf{y}_{1:T}|\mathbf{z}^c_{1:T})\right]\right. \\ & + \mathbb{E}_{q_{\boldsymbol{\phi}}}[\log p_\theta(\mathbf{x}_{1:T}|\mathbf{z}^c_{1:T}, \mathbf{z}^n_{1:T}, \mathbf{z}^c)] \\ & - \alpha_s \mathbb{D}_{KL}(q_{\boldsymbol{\phi}}(\mathbf{z}^s|\mathbf{x}_{<t}) \parallel p_\theta(\mathbf{z}^s)) \\ & - \sum_{t=1}^{T} \alpha_n \mathbb{D}_{KL}(q_{\boldsymbol{\phi}}(\mathbf{z}^n_t|\mathbf{z}^n_{<t}, \mathbf{x}_t) \parallel p_\theta(\mathbf{z}^n_t|\mathbf{z}^n_{<t})) \\ & \left. - \sum_{t=1}^{T} \alpha_c \mathbb{D}_{KL}(q_{\boldsymbol{\phi}}(\mathbf{z}^c_t|\mathbf{z}^c_{<t}, \mathbf{x}_t) \parallel p_\theta(\mathbf{z}^c_t|\mathbf{z}^c_{<t}))\right], \end{aligned} \tag{6}$$

where α_s, α_n and α_c are hyperparameters for balancing the independent constraints and reconstructions.

Theorem 1 shows that with the help of variational inference, the latent space of non-stationary time series which contains three subspaces (i.e., $\mathbf{z}^c$, $\mathbf{z}^n$ and $\mathbf{z}^s$) can be jointly inferred and the learned causal factors can be improved while optimizing the predictor. However, we notice that it is not enough to ensure the causal information to be excluded from non-causal factors, which accordingly may result in $\mathbf{z}^n$ capturing both shape and color of digits in Fig. 2. Besides, recent works on time series generation focus on how to explicitly disentangle dynamic and static information, which is also worth considering in our model. To address these, we propose novel regularization terms based on the information theory to achieve latent factor disentanglement.

4.2. Mutual information constraints

Since there is no disentanglement without inductive bias as discussed by Locatello et al. [33], to ensure better disentanglement, we introduce two constraints for latent variables based on information theory. The first one is designed for disentangling the causal and non-causal information. To this end, we can minimize the terms $I(\mathbf{z}^c, \mathbf{z}^n)$ and $I(\mathbf{z}^c, \mathbf{z}^s)$, which denote the mutual information of causal factors $\mathbf{z}^c$ and non-causal factors $\mathbf{z}^n, \mathbf{z}^s$, respectively. The second one aims to disentangle dynamic and static information, which has been commonly seen in recent literature of disentangled sequential data generation [34,35]. This would benefit the disentanglement performance and improve the learned dynamic causal factors. Specifically, we minimize the upper bound of mutual information between dynamic and static factors, i.e., $I(\mathbf{z}^c, \mathbf{z}^s)$.

It is known that estimating mutual information is not easy, since it requires underlying marginal and joint distributions of continuous latent variables, i.e., $I(\mathbf{z}^a, \mathbf{z}^b) := \int_{\mathbf{z}^a, \mathbf{z}^b} p(\mathbf{z}^a, \mathbf{z}^b) \log \frac{p(\mathbf{z}^a, \mathbf{z}^b)}{p(\mathbf{z}^a)p(\mathbf{z}^b)}$. Although marginal priors can be normal distributions, the joint distribution still cannot be estimated due to unknown interactions between different variables. Previous studies have made a few attempts towards finding a tractable upper bound of mutual information [36,37]. However, they avoid the calculation of joint distribution by transforming it into a conditional form, i.e., $p(\mathbf{z}^a|\mathbf{z}^b)$ or $p(\mathbf{z}^b|\mathbf{z}^a)$, which is not interpretable and thus still hard to estimate in our problem.

Therefore, we propose a novel tractable upper bound of $I(\mathbf{z}^c, \mathbf{z}^n)$, $I(\mathbf{z}^c, \mathbf{z}^s)$ and $I(\mathbf{z}^n, \mathbf{z}^s)$ without estimating conditional distributions.

Theorem 2. *Let $\mathbf{z}^a$ and $\mathbf{z}^b$ be any two of three latent factors inferred by DCSVAE, the upper bound of mutual information with respect to these factors can be formulated by:*

$$I(\mathbf{z}^a, \mathbf{z}^b) \le \mathbb{E}_{\hat{p}(\mathbf{z}^a, \mathbf{z}^b)}[\gamma(\mathbf{z}^a, \mathbf{z}^b)] - \mathbb{E}_{\hat{p}(\mathbf{z}^a)\hat{p}(\mathbf{z}^b)}[\gamma(\mathbf{z}^a, \mathbf{z}^b)] \triangleq \mathcal{L}_{MI}(\mathbf{z}^a, \mathbf{z}^b), \tag{7}$$

where $\hat{p}(\mathbf{z})$ denotes the marginal density function approximation of $p(\mathbf{z})$, and $\gamma(\mathbf{z}^a, \mathbf{z}^b)$ is a parameterized normalized critic function.

Discussions. To ensure disentanglement, we adopt mutual information constraints for latent factors. It is not feasible to calculate and optimize the mutual information directly. Therefore, we need a tractable bound of mutual information as a surrogate objective. There exist some upper bounds to mutual information [36–38]. However, their computation all require the known conditional probability (here is $p(\mathbf{z}^a|\mathbf{z}^b)$ or $p(\mathbf{z}^b|\mathbf{z}^a)$), which is intractable and uninterpretable in our latent factors disentanglement problem since there is no conditional relationship between them. For this reason, we transform the conditional distribution into the joint one and introduce the energy-based critic to make it tractable.

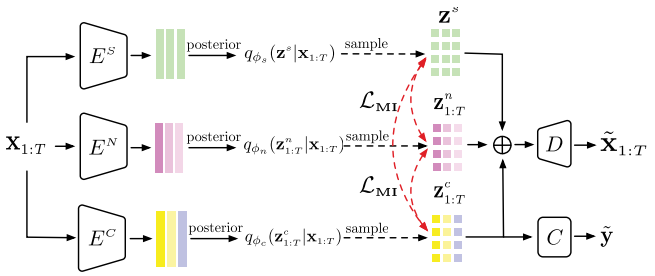


Fig. 3. Overview of our proposed DCSVAE. Given a sequential data $x_{1:T}$. Our model learns three disentangled latent factors: dynamic causal factors $\mathbf{z}^c_{1:T}$, dynamic and static non-causal factors $\mathbf{z}^n_{1:T}, \mathbf{z}^s$.

Algorithm 1 Training DCSVAE.

Input: training set $\{\mathbf{x}^i_{1:T}, \mathbf{y}^i_{1:T}\}^N_i$; fixed training epochs $\mathcal{E}$; mini-batch size $\mathcal{B}$; three latent encoders and predictor; model hyperparameters $\alpha_s, \alpha_n, \alpha_c, \beta$, and initialized parameters of DCSVAE.
Output: Optimized dynamic causal factors encoder $q_{\boldsymbol{\phi}^*}(\mathbf{z}^c_{1:T}|\mathbf{x}_{1:T})$ and predictor $p_{\theta^*}(\mathbf{y}_{1:T}|\mathbf{z}^c_{1:T})$.

```
Initialize e ← 1
while e ≤ E do
    Draw a mini-batch samples {x^i_{1:T}, y^i_{1:T}}_{i∈B} from training set sequentially;
    Infer the latent factors z^c_{1:T}, z^n_{1:T}, z^s as Eq. (5);
    Reconstruct x^i_{1:T} from latent factors as Eq. (3);
    Predict y^i_{1:T} from dynamic causal factors as Eq. (2);
    Minimize the objective via maximizing ELBO and minimizing upper bound of mutual information;
    Update the parameters of DCSVAE with Adam optimizer.
    e ← e + 1
end while
```

In practice, we evaluate $\mathbb{E}_{\hat{p}(\mathbf{z}^a,\mathbf{z}^b)}$ and $\mathbb{E}_{\hat{p}(\mathbf{z}^a)p(\mathbf{z}^b)}$ by drawing samples from the mini-batch. Specifically, we draw joint samples $(\mathbf{z}^a, \mathbf{z}^b)$ from $q_{\boldsymbol{\phi}}(\mathbf{z}^a|\mathbf{x}^{(i)}_{1:T})$ and $q_{\boldsymbol{\phi}}(\mathbf{z}^b|\mathbf{x}^{(i)}_{1:T})$, where (i) denotes a data point in mini-batch. Regarding the independent samples, we obtain them from distributions of different datapoints, i.e., $q_{\boldsymbol{\phi}}(\mathbf{z}^a|\mathbf{x}^{(i)}_{1:T})$ and $q_{\boldsymbol{\phi}}(\mathbf{z}^b|\mathbf{x}^{(j)}_{1:T})$.

Theorem 2 provides a tractable calculation to mutual information upper bound with the help of energy based critic function. Finally, our objective function can be written as:

$$\mathcal{L}_{DCSVAE} = -\mathcal{L}_{ELBO} + \beta(\mathcal{L}_{MI}(\mathbf{z}^c, \mathbf{z}^n) + \mathcal{L}_{MI}(\mathbf{z}^c, \mathbf{z}^s) + \mathcal{L}_{MI}(\mathbf{z}^n, \mathbf{z}^s)), \tag{8}$$

where β is the hyperparameter for balancing the capacity of variational inference and mutual information constraints. Fig. 3 summarizes the overview of our model.

4.3. Training & prediction

We now present the training and prediction phases of our proposed model. Algorithm 1 summarizes the training phase of our proposed model for dynamic causal factors learning.

After convergence, the dynamic causal factors encoder and predictor are fetched to predict unseen test domains (cf. Algorithm 2). Although the optimized predictor $p_{\theta}(\mathbf{y}_{1:T}|\mathbf{z}^c_{1:T})$ aims to predict a series of labels regarding the whole dynamic causal factors, our proposed DCSVAE still works if we have only one label of a time series, depending on the implementation of the predictor.

Algorithm 2 Prediction in DCSVAE.

Input: Non-stationary time series $\mathbf{x}^{te}_{1:T}$ from unseen test domain; optimized DCSVAE.
Output: Prediction of label $\hat{\mathbf{y}}_{1:T}$.

```
Fetch dynamic causal factors encoder q_{φ*}(z^c_{1:T}|x_{1:T}) and predictor p_{θ*}(y_{1:T}|z^c_{1:T}) from optimized DCSVAE;
Infer dynamic causal factors ẑ^c_{1:T} via q_{φ*}(z^c_{1:T}|x_{1:T}) from x^{te}_{1:T};
Predict ŷ_{1:T} from ẑ^c_{1:T} via predictor p_{θ*}(y_{1:T}|z^c_{1:T}).
```

4.4. Revisiting dcsvae

Two things that are worth noticing are: (1) The key data assumptions are that the time series exhibit non-stationarity and that both static and dynamic causal factors are present. Although DCSVAE aims to solve the domain generalization problem for non-stationary time series where both source and temporal shifts occur, our model can also perform well within a non-stationary series, a.k.a., evolving domain generalization [19]. While tackling this problem, separating the dynamic and static information would also benefit the generative model optimization [32]. (2) As discussed above, non-temporal generalization techniques (e.g., IRM, VREx and SD) are inappropriate if the time series is non-stationary, because the dynamic and static non-causal factors are mixed up. On the other hand, these methods would achieve better performance under stationary setting intuitively, where there only exists static information. Since the non-causal factors are modeled explicitly by DCSVAE, the learned causal factors would make the above domain generalization methods capture invariant representation easier. Therefore, our proposed model can be a feature extractor, and achieve better performance under non-stationary setting combining with downstream domain generalization methods.

5. Experiments

In this section, we compare our proposed DCSVAE with recent state-of-the-art baselines and conduct the ablation study to analyze the components. We also provide the visualization of latent factors with the trained inference model, showing the capability of our model disentangling the desired dynamic causal factors and others.

5.1. Experimental setup

5.1.1. Datasets

We conduct experiments on four synthetic datasets and two real-world datasets [12]. (1) *Fourier* consists of one-dimensional signals, generated by inverse Fourier transformations from invariant high frequency and spurious low frequency. We divide it into three source domains according to spurious frequency peak correlations. (2) *TCMNIST* has colored handwriting digit images which have a spurious relation between colors of digits and labels. The relations are manipulated and change with time, sources and both. The corresponding datasets are termed as TCMNIST-temporal, TCMNIST-source and TCMNIST-mixed, respectively. For TCMNIST-temporal, we divide the last three frames as three temporal domains which have different spurious correlations. For the other two datasets, we split them according to the sources with different spurious correlations. (3) *LSA64* records 64 signed words in Argentinian Sign Language from 10 signers, each series which consists of 20 frames representing one signed word. We divide every two signers into a source domain. (4) *Portraits* is constructed for gender classification given photos of American teenagers across 26 states over 108 years. We divide the dataset into 34 temporal domains by years. (5) *PCL* includes motor imagery EEG recordings for three datasets collected by different research groups: PhysionetMI, Cho2017, and Lee2019_MI. Each dataset represents a source domain, and the task involves generalizing motor imagery classification to unseen datasets using EEG measurements. We summarize the dataset used in our experiments and corresponding distribution shifts in Table 1.

Table 1
Datasets and types of distribution shifts.

Datasets	Source	Temporal	Sample size
Fourier	✓	✗	12 000
TCMNIST-source	✓	✗	52 500
TCMNIST-temporal	✗	✓	52 500
TCMNIST-mixed	✓	✓	52 500
LSA64	✓	✓	3200
Portraits	✓	✓	37 921
PCL	✓	✓	22 598

Table 2
Model architectures for each dataset.

Datasets	Extractor	Predictor	Decoder
Fourier	–	LSTM	FC
TCMNIST	ConvNet	LSTM	ConvTranNet
LSA64	ResNet-50	Attention LSTM	FC
Portraits	ResNet-18	FC	ConvTranNet
PCL	EEGNet	FC	FC

5.1.2. Baselines

We now provide detailed descriptions of baselines in our paper. We divide these baselines into two categories, model-agnostic methods and generative methods.

We first discuss the model-agnostic methods. Theoretically, these methods are suitable for both source and temporal generalization tasks because they only focus on distributions across domains instead of considering the source and temporal shifts separately. Generally, they all try to find invariance across domains from data except ERM. Here are these model-agnostic methods:

- ERM [39] minimizes the average empirical risks across different domains.
- IRM [16] is based on ERM, with a penalty term minimizing the local empirical risk across domains.
- IB-ERM & IB-IRM [26] are extensions of ERM and IRM, respectively. These approaches incorporate information bottleneck constraints into ERM and IRM for better generalization ability.
- VREx [22] performs ERM with a constraint minimizing the variance of empirical risks across domains.
- SD [40] extends ERM by regularizing the L2 norm of logits, which contributes to find domain invariant representation.

We also compare our DCSVAE with the recent generative methods for domain generalization:

- DIVA [27] aims to find domain invariant representation by a variational autoencoder. Specifically, it separates latent factors into label information, domain information, and remainder.
- LSSAE [19] is designed for temporal generalization, a.k.a., evolving domain generalization. It also models three independent factors, considering the covariate and concept shifts. It is noted that one latent factor is inferred by label series. Thus LSSAE will fail if we do not have label series.
- AIRL [41], which is short for adaptive invariant representation learning, aims at non-stationary setting in domain generalization as well. The core difference between this work and ours is that our proposed explicitly disentangle the causal factors and non-causal factors.

5.1.3. Model architectures

The model architectures for each dataset are summarized in Table 2. For fair comparison, we fix the feature extractor and predictor for each dataset.

- *Fourier*. Since Fourier consists of one-dimensional signals, we do not use any feature extractor. In addition, we take two LSTM layers with two fully connected layers as the predictor.
- *TCMNIST*. For the three TCMNIST datasets, all the inputs are the sequences of colored MNIST images, thus we use the same extractor and predictor for them. We leverage MNIST ConvNet provided by [15] to extract the representation from MNIST images. The predictor for TCMNIST is a network with an LSTM layer.
- *LSA64*. The size of each video frame from LSA64 is $3 \times 224 \times 224$. To extract representation, we use a frozen ResNet-50 model, which was pre-trained on ImageNet, as a feature extractor. The predictor is an LSTM models with a self-attention layer, and a fully connected network.
- *Portraits*. The feature extractor is a frozen ResNet-18 pre-trained on ImageNet which is similar to LSA64. The predictor is implemented by a fully connected layer.
- *PCL*. For this dataset, we employ a deep convolutional neural network (CNN) model, EEGNet, as proposed by Lawhern et al. [42]. This model was selected due to its strong recognition and widespread acceptance within the EEG research community.

Note that for baselines and our proposed model, the architectures of the above extractors and predictors are fixed for each dataset. The generative models (i.e., DIVA, LSSAE and ours) are implemented by encoder–decoder architectures, and the feature extractor can work as a encoder with reparameterization trick. Therefore we need the extra decoders for these methods. For Fourier, we implement it with a fully connected network. For TCMNIST and Portraits, we use transposed convolution and batch normalization to reconstruct the MNIST images. Additionally, we notice that reconstructing images in LSA64 is very time-consuming. Considering the efficiency, we aim to reconstruct the representation generated by ResNet-50 with a fully connected network, and this has been proven to be effective and efficient experimentally.

5.1.4. Evaluation settings

Here we evaluate our proposed model under two settings, i.e., *source generalization* and *temporal generalization*. The former is non-stationary time series domain generalization defined above where we intend to predict the labels of unseen source domains. We conduct experiments under this setting on Fourier, TCMNIST-source, TCMNIST-mixed and LSA64 datasets. The latter is evolving domain generalization, mainly considering temporal shift within one time series. The goal is to predict the label of future data given past data and labels. TCMNIST-temporal and Portraits are used in this setting.

5.1.5. Model selection

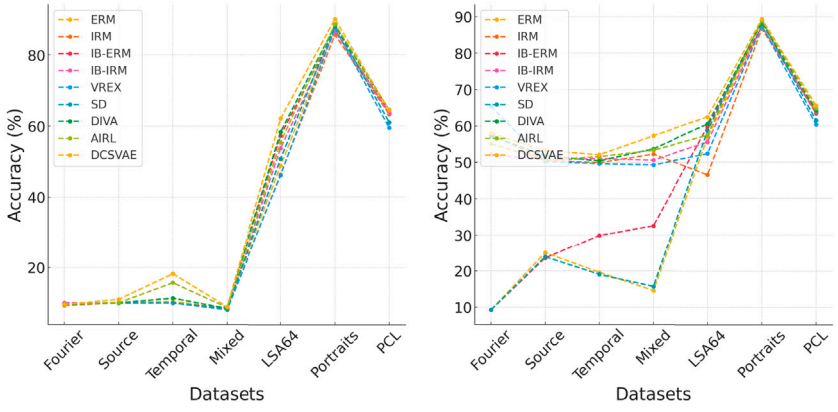
Since test domains are not accessible while OOD generalization training, it is vital for selecting the right model by validation as emphasized in [15]. Here we adopt two model selection methods to fully show generalization ability of proposed model: (1) *train-domain validation*. We use a validation set from train domain to select the model which complies with real generalization scenario; (2) *test-domain validation*. The validation set is from test domain. We train models in fixed epochs and select the model that performs the best in the final epoch. Though it is impossible to access to test domain in practice, this selection method could provide more insights about the generalization ability.

5.2. Performance comparison

The results of the above baselines and the proposed method are reported in Table 3. Note that LSSAE needs a series of labels, so it fails to work on datasets with only one label per sequence. For synthetic datasets, we have two high spurious correlation domains A, B and a low one C. We aim to predict C by a model trained on A and B to evaluate the generalization ability when distributions shift. For the Portraits dataset, our task is predicting the gender of future images trained on the past images and labels. It is noted that for the LSA64 dataset, we have five different source domains, and we evaluate these domains one by

Table 3
Comparison of test accuracy (%) for DCSVAE and baselines with the train-domain validation (top 10 rows) and the test-domain validation (bottom 10 rows). The best results are in **bold**, and the second best ones are underlined.

	Datasets	Fourier	TCMNIST			LSA64	Portraits	PCL	Average
	Methods		Source	Temporal	Mixed				
Train-domain Val	ERM	9.55	10.27	10.46	8.50	48.78	87.37	63.47	34.06
	IRM	9.35	10.04	10.04	8.10	46.31	85.68	63.22	33.25
	IB-ERM	**10.08**	9.99	10.04	8.58	57.28	86.87	63.76	35.23
	IB-IRM	9.97	10.05	10.04	8.10	53.71	86.63	63.53	34.58
	VREx	9.74	10.04	10.05	8.10	46.11	87.59	59.44	33.01
	SD	9.70	9.99	10.05	8.45	50.74	88.53	60.98	34.06
	DIVA	9.60	10.08	11.38	8.48	58.24	88.26	**64.46**	35.79
	LSSAE	–	10.04	16.41	8.89	–	89.06	62.06	–
	AIRL	9.69	10.08	15.74	8.75	55.35	88.65	64.52	36.11
	DCSVAE	9.59	**11.07**	**18.26**	**8.95**	**62.06**	**90.09**	64.37	**37.77**
Test-domain Val	ERM	9.28	25.03	19.58	14.60	56.82	87.59	63.51	39.49
	IRM	57.68	50.57	49.92	52.27	46.48	86.64	63.87	58.20
	IB-ERM	9.28	23.56	29.90	32.55	59.78	87.37	63.38	43.69
	IB-IRM	52.22	50.66	51.05	50.63	55.51	87.31	63.75	60.00
	VREx	**65.39**	50.20	49.67	49.30	52.32	87.44	60.31	59.23
	SD	9.28	23.89	19.03	15.73	58.62	88.92	61.46	39.56
	DIVA	57.31	51.81	50.42	53.72	60.48	88.41	64.11	60.89
	LSSAE	–	50.31	51.77	55.26	–	89.29	64.47	–
	AIRL	55.15	50.66	51.58	53.33	57.43	88.91	64.73	60.26
	DCSVAE	58.04	**53.25**	**52.11**	**57.32**	**62.44**	**89.35**	**65.55**	**62.58**



(a) Train-domain Validation. (b) Test-domain Validation.

Fig. 4. Comparison of test accuracy of our proposed DCSVAE and baselines (Source, Temporal, and Mixed are short for TCMNIST-source, TCMNIST-temporal, and TCMNIST-mixed, respectively).

one. Note that we compare our proposed with baselines in train-domain and test-domain validation settings, and the overall result is shown in Fig. 4.

We first discuss the performances on synthetic datasets. Our proposed model does not perform the best on the Fourier dataset. The plausible explanation is that the information embedded in a one-dimensional series is not enough to capture desired causal factors. In contrast, DCSVAE outperforms all baselines across three TCMNIST image datasets. Furthermore, note that the model-agnostic baselines drop in performance with the train-domain validation, but our model achieves state-of-the-art performance regardless of which validation setting is adopted. We attribute this amazing improvement to well-learned dynamic causal factors. For real-world datasets, the non-stationary property is not as significant as carefully constructed synthetic datasets. Therefore, the performance of generalization methods is not remarkable, and yet our model achieves superior performance against baselines. The real-world dataset we use including LSA64, Portrait, and PCL, the last one can be considered representative of healthcare data. The results on these datasets suggest that our proposed model has notable advantages, as patient records inherently involve both source and temporal shifts. These characteristics are effectively addressed by DCSVAE, demonstrating its potential to tackle complex, real-world datasets in domains like healthcare. The ability to handle such challenges underscores the model's robustness and practical applicability.

5.3. Ablation study

Overall, our proposed DCSVAE has two major components: encoder–decoder (P1) and mutual information constraints (P2). To analyze the effect of each component, we intentionally construct two groups of variants w.r.t. the two components. There are three variants in group P1, including w/o $\mathbf{z}^s$&$\mathbf{z}^n$, w/o $\mathbf{z}^s$ and w/o $\mathbf{z}^n$. Specifically, w/o $\mathbf{z}^s$&$\mathbf{z}^n$ only learns a series of latent factors like VRNN, w/o $\mathbf{z}^s$ means source shifts are not taken into consideration, and w/o $\mathbf{z}^n$ denotes ignoring the temporal shifts within time series. As for group P2, we

Table 4
Ablation results of DCSVAE and variants, P1 denotes changes upon encoder–decoder architectures, and P2 indicates different kinds of constraints.

	Datasets	TCMNIST		
	Methods	Source	Temporal	Mixed
P1	w/o $\mathbf{z}^s$ & $\mathbf{z}^n$	9.77 ± 0.39	10.03 ± 0.25	8.45 ± 0.34
	w/o $\mathbf{z}^s$	10.17 ± 0.62	15.98 ± 1.33	8.33 ± 0.12
	w/o $\mathbf{z}^n$	10.73 ± 0.89	13.69 ± 1.05	8.10 ± 0.35
P2	w/o causal	10.79 ± 0.37	18.13 ± 1.72	8.73 ± 1.95
	w/o dynamic	10.94 ± 0.66	17.94 ± 3.86	8.49 ± 0.84
	w/o both	10.75 ± 0.50	16.51 ± 0.44	8.51 ± 0.69
	DCSVAE	**11.02 ± 0.83**	**18.26 ± 5.41**	**8.95 ± 0.40**

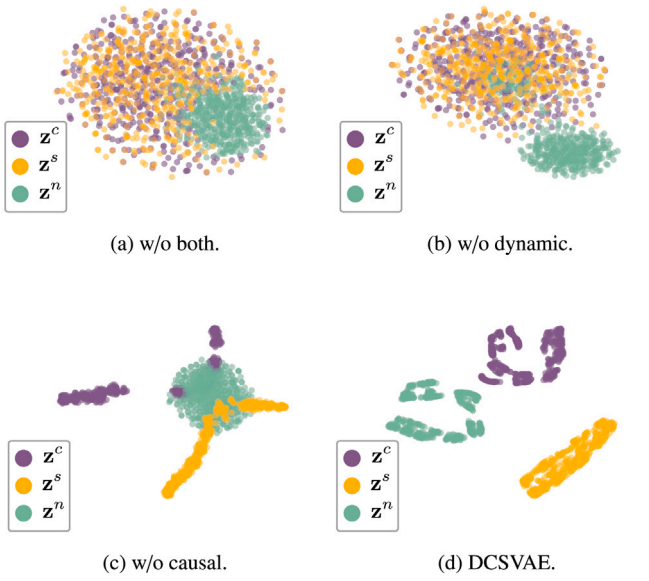


Fig. 5. Visualization of three latent factors learned by our DCSVAE and its variants in group P2 with 2-D UMAP.

have three variants which drop the constraints of causal/non-causal information (w/o causal), dynamic/static information (w/o dynamic) and both (w/o both). We conduct these ablation experiments on TCMNIST and the results are represented in Table 4.

DCSVAE outperforms all other variants, indicating the effectiveness of each component we designed. Furthermore, the mutual information constraints we employ aim to disentangle latent factors. To understand the latent factors learned by our proposed model, we extract the latent factors learned by DCSVAE and variants in P2 on TCMNIST-mixed and visualize these three embedding vectors in a 2D space using Umap. The visualization results are shown in Fig. 5. We find that DCSVAE without mutual information constraints could mix up all latent factors, resulting the overlapping in the latent space. In contrast, minimizing the mutual information can force the latent factors to be disentangled (cf. Fig. 5(d)). More interestingly, comparing Figs. 5(b) and 5(c), the constraint of causal/non-causal information is less important than another one, we think prediction from causal factors (cf. Eq. (2)) is beneficial to separating the causal/non-causal factors to some extent.

5.4. On the off-the-shelf generalization

For the proposed DCSVAE and its variants, we incorporate the above model-agnostic generalization methods. They are evaluated on the TCMNIST dataset, and the results are reported in Table 5, where "w/ baseline" denotes DCSVAE with the corresponding baseline. Note that we evaluate these methods with the test-domain validation for comparison.

Table 5
Results of DCSVAE and its variants on improving off-the-shelf generalization methods.

Datasets	TCMNIST		
Methods	Source	Temporal	Mixed
w/ ERM	11.02 ± 0.83	18.26 ± 5.41	**8.95 ± 0.40**
w/ IRM	10.21 ± 0.09	18.17 ± 1.46	8.79 ± 0.34
w/ IB-ERM	10.96 ± 0.87	18.30 ± 5.38	8.93 ± 0.40
w/ IB-IRM	10.22 ± 0.11	18.16 ± 1.43	8.79 ± 0.36
w/ VREx	10.22 ± 0.08	18.13 ± 0.76	8.81 ± 0.34
w/ SD	**11.04 ± 1.09**	**20.03 ± 7.60**	8.86 ± 0.39

By comparing the variants with raw baselines, we notice that our proposed model improves the performance of baselines in general. The results are in line with our expectation, since the learned dynamic causal factors could make prediction easier. Another interesting observation is that SD works worse than other baselines but the performance of w/ SD is very close to others (cf. Tables 3 and 5). We think the reversal of relations happens because SD does not exclude spurious non-causal factors explicitly [40], and DCSVAE helps SD separate causal and non-causal information, thus performing better than w/ ERM.

6. Conclusion

In this paper, we proposed DCSVAE, a generative model that can disentangle three latent factors through constraining mutual information for better generalizing across non-stationary time series. Extensive empirical results demonstrated its superior performance over conventional domain generalization methods in both source and temporal generalization tasks. Additionally, the learned dynamic causal factors can improve the performance of conventional domain generalization methods under the non-stationary setting, because the dynamic and static factors can be mixed up easily and can be excluded from causal factors in our proposed method.

However, there are some limitations to the proposed method. Firstly, DCSVAE incurs additional computational costs due to mutual information constraints, which raises concerns about scalability. Specifically, when applied to very large datasets, the length of dynamic factors can become a bottleneck for computational efficiency. Secondly, the process of learning distinct latent factors may require a substantial amount of labeled data. Besides, we notice that dynamic causal factors do not necessarily change all the time. This may hinder modeling causal factors which are assumed to vary along time. Thus, we expect to capture segmented dynamic causal factors rather at the time step level, which might benefit non-stationary time series modeling. Due to the distributional assumptions of the data, careful consideration is required when applying the model to new domains, such as climate modeling, where the data may not adhere to these assumptions. This could be a potential improvement in this direction.

CRediT authorship contribution statement

Weifeng Zhang: Writing – original draft, Visualization, Validation, Software, Methodology. **Yan Liu:** Writing – review & editing, Conceptualization. **Xovee Xu:** Writing – review & editing, Conceptualization. **Fan Zhou:** Writing – review & editing, Supervision, Investigation, Funding acquisition, Conceptualization. **Ting Zhong:** Writing – review & editing, Supervision. **Kunpeng Zhang:** Writing – review & editing, Conceptualization.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Fan Zhou reports financial support was provided by National Natural Science Foundation of China. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant Nos. 62072077, 62176043, and U22A2097.

Appendix. Details of theoretical proof

Theorem 1. Combined with above causal graph of non-stationary time series as shown in Fig. 1(d), the evidence lower bound of likelihood (cf. Eq. (4)) in DCSVAE is:

$$\begin{aligned}\mathcal{L}_{ELBO} &= \mathbb{E}_{p^{tr}}\Big[\mathbb{E}\left[\log p_\theta(\mathbf{y}_{1:T}|\mathbf{z}^c_{1:T})\right] \\ &+ \mathbb{E}_{q_\phi}[\log p_\theta(\mathbf{x}_{1:T}|\mathbf{z}^c_{1:T},\mathbf{z}^n_{1:T},\mathbf{z}^c)] \\ &- \alpha_s \mathbb{D}_{KL}(q_\phi(\mathbf{z}^s|\mathbf{x}_{<t}) \parallel p_\theta(\mathbf{z}^s)) \\ &- \sum_{t=1}^{T} \alpha_n \mathbb{D}_{KL}(q_\phi(\mathbf{z}^n_t|\mathbf{z}^n_{<t},\mathbf{x}_t) \parallel p_\theta(\mathbf{z}^n_t|\mathbf{z}^n_{<t})) \\ &- \sum_{t=1}^{T} \alpha_c \mathbb{D}_{KL}(q_\phi(\mathbf{z}^c_t|\mathbf{z}^c_{<t},\mathbf{x}_t) \parallel p_\theta(\mathbf{z}^c_t|\mathbf{z}^c_{<t}))\Big], \end{aligned} \tag{A.1}$$

where α_s, α_n and α_c are hyperparameters for balancing the independent constraints and reconstructions.

Proof. It is intractable to estimate the likelihood straightly given training data, since $p_\theta(\mathbf{x}_{1:T}, y) = \int p_\theta(\mathbf{x}_{1:T}, y, \mathbf{z}^c, \mathbf{z}^n, \mathbf{z}^s)\mathrm{d}\mathbf{z}^c\ \mathrm{d}\mathbf{z}^n\mathrm{d}\mathbf{z}^s$ is difficult to estimate. Following past work on variational inference, we introduce inference models where latent factors can be sampled easily, making likelihood estimation tractable. According to the above discussion about generative and inference models, the latent space would be split into three spaces, denoted as $\mathbf{z}^c$, $\mathbf{z}^n$ and $\mathbf{z}^s$. We substitute these prior and posterior distributions with Eqs. (1)–(3) and Eq. (5), then the ELBO objective can be derived as (we omit the expectation note $\mathbb{E}_{p^{tr}}(\cdot)$ for brevity):

$$\log \mathbb{E}\left[p_\theta(\mathbf{x}_{1:T}, y, \mathbf{z}^c_{1:T}, \mathbf{z}^n_{1:T}, \mathbf{z}^s)\right] \tag{A.2}$$

$$\geq \mathbb{E}\left[\log p_\theta(\mathbf{x}_{1:T}, y, \mathbf{z}^c_{1:T}, \mathbf{z}^n_{1:T}, \mathbf{z}^s)\right] \tag{A.3}$$

$$= \mathbb{E}\left[\log\left[p_\theta(y|\mathbf{z}^c_{1:T})p_\theta(\mathbf{x}_{1:T}, \mathbf{z}^c_{1:T}, \mathbf{z}^n_{1:T}, \mathbf{z}^s)\right]\right] \tag{A.4}$$

$$\begin{aligned}&= \mathbb{E}_{q_\phi}\left[\log p_\theta(y|\mathbf{z}^c_{1:T})\right] \\ &+ \mathbb{E}_{q_\phi}\left[\log \frac{p_\theta(\mathbf{x}_{1:T}, \mathbf{z}^c_{1:T}, \mathbf{z}^n_{1:T}, \mathbf{z}^s)}{q_\phi(\mathbf{z}^c_{1:T}, \mathbf{z}^n_{1:T}, \mathbf{z}^s|\mathbf{x}_{1:T})}\right]\end{aligned} \tag{A.5}$$

$$\begin{aligned}&= \mathbb{E}_{q_\phi}\left[\log p_\theta(y|\mathbf{z}^c_{1:T})\right] + \mathbb{E}_{q_\phi}[\log p_\theta(\mathbf{x}_{1:T}|\mathbf{z}^c_{1:T},\mathbf{z}^n_{1:T},\mathbf{z}^c)] \\ &- \mathbb{D}_{KL}(q_\phi(\mathbf{z}^s|\mathbf{x}_{<t}) \parallel p_\theta(\mathbf{z}^s)) \\ &- \sum_{t=1}^{T} \mathbb{D}_{KL}(q_\phi(\mathbf{z}^n_t|\mathbf{z}^n_{<t},\mathbf{x}_t) \parallel p_\theta(\mathbf{z}^n_t|\mathbf{z}^n_{<t})) \\ &- \sum_{t=1}^{T} \mathbb{D}_{KL}(q_\phi(\mathbf{z}^c_t|\mathbf{z}^c_{<t},\mathbf{x}_t) \parallel p_\theta(\mathbf{z}^c_t|\mathbf{z}^c_{<t})),\end{aligned} \tag{A.6}$$

where the inequality holds due to the concavity of logarithm function, the first term is the prediction for labels, the second term denotes the reconstruction of input time series and the last three terms are KL-divergence values which are regularizations aligning posterior distributions with corresponding prior distributions. Inspired by beta-vae [43], we introduce three coefficients α_s, α_n and α_c to balance the independent constraints and reconstructions. ■

Theorem 2. Let $\mathbf{z}^a$ and $\mathbf{z}^b$ be any two of three latent factors inferred by DCSVAE, the upper bound of mutual information with respect to these factors can be formulated by:

$$I(\mathbf{z}^a, \mathbf{z}^b) \leq \mathbb{E}_{\hat{p}(\mathbf{z}^a,\mathbf{z}^b)}[\gamma(\mathbf{z}^a, \mathbf{z}^b)] - \mathbb{E}_{\hat{p}(\mathbf{z}^a)\hat{p}(\mathbf{z}^b)}[\gamma(\mathbf{z}^a, \mathbf{z}^b)] \triangleq \mathcal{L}_{MI}(\mathbf{z}^a, \mathbf{z}^b), \tag{A.7}$$

where $\hat{p}(\mathbf{z})$ denotes the marginal density function approximation of $p(\mathbf{z})$, and $\gamma(\mathbf{z}^a, \mathbf{z}^b)$ is a parameterized normalized critic function.

Proof. Here, the normalized critic function $\gamma(\mathbf{z}^a, \mathbf{z}^b)$ denotes an energy-based variational family of the joint distribution [44], i.e.:

$$p(\mathbf{z}^a, \mathbf{z}^b) = \frac{p(\mathbf{z}^a)p(\mathbf{z}^b)}{\mathcal{Z}} e^{\gamma(\mathbf{z}^a,\mathbf{z}^b)}, \tag{A.8}$$

where $\mathcal{Z} = \mathbb{E}_{\hat{p}(\mathbf{z}^a)\hat{p}(\mathbf{z}^b)}[e^{\gamma(\mathbf{z}^a,\mathbf{z}^b)}]$ is a expected value that is irrelevant to $\mathbf{z}^a$ and $\mathbf{z}^b$. With this formula, we can derive Eq. (A.7) by applying it into an existing upper bound CLUB [37]:

$$I(\mathbf{z}^a, \mathbf{z}^b) \leq I_{\mathrm{CLUB}} \tag{A.9}$$

$$= \mathbb{E}_{\hat{p}(\mathbf{z}^a,\mathbf{z}^b)}[\log \hat{p}(\mathbf{z}^a|\mathbf{z}^b)] - \mathbb{E}_{\hat{p}(\mathbf{z}^a)\hat{p}(\mathbf{z}^b)}[\log p(\mathbf{z}^a|\mathbf{z}^b)] \tag{A.10}$$

$$= \mathbb{E}_{\hat{p}(\mathbf{z}^a,\mathbf{z}^b)}[\log p(\mathbf{z}^a, \mathbf{z}^b)] - \mathbb{E}_{\hat{p}(\mathbf{z}^a)\hat{p}(\mathbf{z}^b)}[\log p(\mathbf{z}^a, \mathbf{z}^b)] \tag{A.11}$$

$$= \mathbb{E}_{\hat{p}(\mathbf{z}^a,\mathbf{z}^b)}[\log p(\mathbf{z}^a) + \log p(\mathbf{z}^b) + \gamma(\mathbf{z}^a, \mathbf{z}^b) - \log \mathcal{Z}] \tag{A.12}$$

$$- \mathbb{E}_{\hat{p}(\mathbf{z}^a)\hat{p}(\mathbf{z}^b)}[\log p(\mathbf{z}^a) + \log p(\mathbf{z}^b) + \gamma(\mathbf{z}^a, \mathbf{z}^b) - \log \mathcal{Z}] \tag{A.13}$$

$$= \mathbb{E}_{\hat{p}(\mathbf{z}^a,\mathbf{z}^b)}[\gamma(\mathbf{z}^a, \mathbf{z}^b)] - \mathbb{E}_{\hat{p}(\mathbf{z}^a)\hat{p}(\mathbf{z}^b)}[\gamma(\mathbf{z}^a, \mathbf{z}^b)] \tag{A.14}$$

$$\triangleq \mathcal{L}_{MI}(\mathbf{z}^a, \mathbf{z}^b). \tag{A.15}$$

Thus, our proposed $\mathcal{L}_{MI}$ is a valid upper bound to mutual information. ■

Data availability

The datasets are publicly available and well-known in the literature.

References

[1] A. Torralba, A.A. Efros, Unbiased look at dataset bias, in: CVPR, 2011, pp. 1521–1528.
[2] B. Schölkopf, F. Locatello, S. Bauer, N.R. Ke, N. Kalchbrenner, A. Goyal, Y. Bengio, Toward causal representation learning, Proc. IEEE 109 (5) (2021) 612–634.
[3] F. Lv, J. Liang, S. Li, B. Zang, C.H. Liu, Z. Wang, D. Liu, Causality inspired representation learning for domain generalization, in: CVPR, 2022, pp. 8046–8056.
[4] Y. Wang, K. Yu, G. Xiang, F. Cao, J. Liang, Discovering causally invariant features for out-of-distribution generalization, Pattern Recognit. 150 (2024) 110338.
[5] A. Miyai, J. Yang, J. Zhang, Y. Ming, Y. Lin, Q. Yu, G. Irie, S. Joty, Y. Li, H. Li, et al., Generalized out-of-distribution detection and beyond in vision language model era: A survey, 2024, arXiv preprint arXiv:2407.21794.
[6] J. Wang, C. Lan, C. Liu, Y. Ouyang, T. Qin, Generalizing to unseen domains: A survey on domain generalization, in: IJCAI, 2021, pp. 4627–4635.
[7] B. Nestor, M.B. McDermott, W. Boag, G. Berner, T. Naumann, M.C. Hughes, A. Goldenberg, M. Ghassemi, Feature robustness in non-stationary health records: Caveats to deployable model performance in common clinical machine learning tasks, in: MLHC, 2019, pp. 381–405.
[8] A. Lazaridou, A. Kuncoro, E. Gribovskaya, D. Agrawal, A. Liska, T. Terzi, M. Gimenez, C. de Masson d'Autume, T. Kocisky, S. Ruder, et al., Mind the gap: Assessing temporal generalization in neural language models, in: NeurIPS, 2021.
[9] D. Zhang, Z. Zhang, N. Chen, Y. Wang, Dynamic convolutional time series forecasting based on adaptive temporal bilateral filtering, Pattern Recognit. 158 (2025) 110985.
[10] W. Wang, E. Zuo, C. Chen, C. Chen, J. Zhong, Z. Yan, X. Lv, Efficient time series adaptive representation learning via dynamic routing sparse attention, Pattern Recognit. (2024) 111058.
[11] Z. Yao, Y. Wang, J. Wang, P. Yu, M. Long, Videodg: Generalizing temporal relations in videos to novel domains, IEEE TPAMI (2021).

[12] J.-C. Gagnon-Audet, K. Ahuja, M.-J. Darvishi-Bayazi, G. Dumas, I. Rish, WOODS: Benchmarks for out-of-distribution generalization in time series tasks, TMLR (2023).
[13] Y. Sui, W. Mao, S. Wang, X. Wang, J. Wu, X. He, T.-S. Chua, Enhancing out-of-distribution generalization on graphs via causal attention learning, ACM Trans. Knowl. Discov. from Data 18 (5) (2024) 1–24.
[14] P. Sheth, R. Moraffah, K.S. Candan, A. Raglin, H. Liu, Domain generalization–a causal perspective, 2022, arXiv preprint arXiv:2209.15177.
[15] I. Gulrajani, D. Lopez-Paz, In search of lost domain generalization, in: ICLR, 2021.
[16] M. Arjovsky, L. Bottou, I. Gulrajani, D. Lopez-Paz, Invariant risk minimization, 2019, ArXiv:1907.02893.
[17] C. Liu, X. Sun, J. Wang, H. Tang, T. Li, T. Qin, W. Chen, T.-Y. Liu, Learning causal semantic representation for out-of-distribution prediction, in: NeurIPS, vol. 34, 2021.
[18] R.J. Hyndman, G. Athanasopoulos, Forecasting: Principles and Practice, OTexts, 2018.
[19] T. Qin, S. Wang, H. Li, Generalizing to evolving domains with latent structure-aware sequential autoencoder, in: ICML, vol. 162, 2022, pp. 18062–18082.
[20] W. Wang, G. Xu, R. Pu, J. Li, F. Zhou, C. Shui, C. Ling, C. Gagné, B. Wang, Evolving domain generalization, 2022, ArXiv:2206.00047.
[21] W. Lu, J. Wang, X. Sun, Y. Chen, X. Xie, Out-of-distribution representation learning for time series classification, in: ICLR, 2022.
[22] D. Krueger, E. Caballero, J.-H. Jacobsen, A. Zhang, J. Binas, D. Zhang, R. Le Priol, A. Courville, Out-of-distribution generalization via risk extrapolation (rex), in: ICML, 2021, pp. 5815–5826.
[23] J. Hu, L. Qi, J. Zhang, Y. Shi, Domain generalization via inter-domain alignment and intra-domain expansion, Pattern Recognit. 146 (2024) 110029.
[24] J.-Z. Chu, B. Pan, X. Xu, T.-Y. Shi, Z.-W. Shi, T. Li, Joint variational inference network for domain generalization, Pattern Recognit. 154 (2024) 110587.
[25] R. Christiansen, N. Pfister, M.E. Jakobsen, N. Gnecco, J. Peters, A causal framework for distribution generalization, IEEE TPAMI (2021) 6614–6630.
[26] K. Ahuja, E. Caballero, D. Zhang, J.-C. Gagnon-Audet, Y. Bengio, I. Mitliagkas, I. Rish, Invariance principle meets information bottleneck for out-of-distribution generalization, in: NeurIPS, 2021.
[27] M. Ilse, J.M. Tomczak, C. Louizos, M. Welling, Diva: Domain invariant variational autoencoders, in: Medical Imaging with Deep Learning, 2020, pp. 322–348.
[28] J. Li, B. Wu, X. Sun, Y. Wang, Causal hidden Markov model for time series disease forecasting, in: CVPR, 2021, pp. 12105–12114.
[29] A. Nasery, S. Thakur, V. Piratla, A. De, S. Sarawagi, Training for the future: A simple gradient interpolation loss to generalize along time, in: NeurIPS, 2021, pp. 19198–19209.
[30] G. Bai, C. Ling, L. Zhao, Temporal domain generalization with drift-aware dynamic neural networks, in: ICLR.
[31] B. Schölkopf, Causality for machine learning, in: Probabilistic and Causal Inference: The Works of Judea Pearl, 2022, pp. 765–804.
[32] L. Yingzhen, S. Mandt, Disentangled sequential autoencoder, in: ICML, 2018, pp. 5670–5679.
[33] F. Locatello, S. Bauer, M. Lucic, G. Raetsch, S. Gelly, B. Schölkopf, O. Bachem, Challenging common assumptions in the unsupervised learning of disentangled representations, in: ICML, 2019, pp. 4114–4124.
[34] Y. Zhu, M.R. Min, A. Kadav, H.P. Graf, S3vae: Self-supervised sequential vae for representation disentanglement and data generation, in: CVPR, 2020, pp. 6538–6547.
[35] J. Bai, W. Wang, C.P. Gomes, Contrastively disentangled sequential variational autoencoder, in: NeurIPS, 2021, pp. 10105–10118.
[36] B. Poole, S. Ozair, A. Van Den Oord, A. Alemi, G. Tucker, On variational bounds of mutual information, in: ICML, 2019, pp. 5171–5180.
[37] P. Cheng, W. Hao, S. Dai, J. Liu, Z. Gan, L. Carin, Club: A contrastive log-ratio upper bound of mutual information, in: ICML, 2020, pp. 1779–1788.
[38] A.A. Alemi, I. Fischer, J.V. Dillon, K. Murphy, Deep variational information bottleneck, in: ICLR, 2017.
[39] V. Vapnik, Statistical learning theory new york, NY: Wiley 1 (2) (1998) 3.
[40] M. Pezeshki, O. Kaba, Y. Bengio, A.C. Courville, D. Precup, G. Lajoie, Gradient starvation: A learning proclivity in neural networks, in: NeurIPS, 2021, pp. 1256–1272.
[41] T.-H. Pham, X. Zhang, P. Zhang, Non-stationary domain generalization: Theory and algorithm, 2024, arXiv preprint arXiv:2405.06816.
[42] V.J. Lawhern, A.J. Solon, N.R. Waytowich, S.M. Gordon, C.P. Hung, B.J. Lance, EEGNet: A compact convolutional neural network for EEG-based brain–computer interfaces, J. Neural Eng. 15 (5) (2018) 056013.
[43] I. Higgins, L. Matthey, A. Pal, C.P. Burgess, X. Glorot, M.M. Botvinick, S. Mohamed, A. Lerchner, beta-VAE: Learning basic visual concepts with a constrained variational framework, in: ICLR, 2017.
[44] X. Nguyen, M.J. Wainwright, M.I. Jordan, Estimating divergence functionals and the likelihood ratio by convex risk minimization, IEEE Trans. Inform. Theory 56 (11) (2010) 5847–5861.