

# Generalizing to Evolving Domains with Latent Structure-Aware Sequential Autoencoder

Tixin Qin<sup>1</sup> Shiqi Wang<sup>1</sup> Haoliang Li<sup>1</sup>

## Abstract

Domain generalization aims to improve the generalization capability of machine learning systems to out-of-distribution (OOD) data. Existing domain generalization techniques embark upon stationary and discrete environments to tackle the generalization issue caused by OOD data. However, many real-world tasks in non-stationary environments (*e.g.*, self-driven car system, sensor measures) involve more complex and continuously evolving domain drift, which raises new challenges for the problem of domain generalization. In this paper, we formulate the aforementioned setting as the problem of *evolving domain generalization*. Specifically, we propose to introduce a probabilistic framework called Latent Structure-aware Sequential Autoencoder (LSSAE) to tackle the problem of evolving domain generalization via exploring the underlying continuous structure in the latent space of deep neural networks, where we aim to identify two major factors namely *covariate shift* and *concept shift* accounting for distribution shift in non-stationary environments. Experimental results on both synthetic and real-world datasets show that LSSAE can lead to superior performances based on the evolving domain generalization setting.

## 1. Introduction

The success of machine learning techniques typically lies on the assumption that training data and test data are sampled independently and identically from similar distributions. However, this assumption does not hold when deploying the trained model in many real-world environments where the distribution of test data varies from training data. This distribution discrepancy, so-called distribution shift, can lead

<sup>1</sup>City University of Hong Kong, Hong Kong. Correspondence to: Haoliang Li <haoliang.li1991@gmail.com>.

to the dramatic performance decrease of machine learning models (Torralba & Efros, 2011). To mitigate this issue, domain generalization (DG) has been proposed to learn a more robust model which can be better generalized to OOD data (Muandet et al., 2013; Balaji et al., 2018; Li et al., 2018b).

While some progress is being achieved so far, existing DG methods are limited to the setting of generalization among discrete and stationary environments. This setting can be problematic in some real-world applications where we require that model can be generalized among continuous domains (Hoffman et al., 2014). For example, a self-driving car system, when deployed in the real world, struggles to perform under an open environment where the accepted data changes naturally according to the geographic location, time intervals, and other factors in a gradual manner (Hoffman et al., 2014). For another example, the measures of sensors can also drift over time due to the outer environments and inter factors, such as aging (Vergara et al., 2012). In these scenarios, treating each domain in a separate manner is unlikely to yield the desired performance as it does not consider the property of continuous domain structure.

Another limitation of most of the existing DG methods is that they did not take “concept shift” into consideration. Such concept shift can also lead to performance drop (Federici et al., 2021). A typical example of concept shift would be that the incidence rate of a particular disease in certain groups may change over time due to the development of treatments and preventive measures. Therefore, existing DG techniques may fail to be applied to some other complex real-world applications in non-stationary environments (Sugiyama et al., 2013; Tahmasbi et al., 2021).

In this paper, we propose to focus on the problem of domain generalization based on the non-stationary setting, where data can evolve gradually with both covariate shift and concept shift. Particularly, we formulate this non-stationary scenario as *evolving domain generalization* where we only have access to adequate labeled examples from the sequential source domains. Our objective is to develop algorithms that can explore the underlying continuous structure of distribution shift and generalize well to evolving target domains where the samples are unavailable during the training stage.

Unlike existing DG methods that only focus on covariate shift based on the stationary environment, in this paper, we propose a novel framework called Latent Structure-aware Sequential Autoencoder (LSSAE), a dynamic probabilistic framework to model the underlying latent variables across domains. More specifically, we propose to use two latent variables to represent the sampling bias in data sample space (i.e., covariate shift) and data category space (i.e., concept shift), and propose a domain-related module and a category-related module to infer their dynamic transition functions based on different time stamps. We conduct extensive experiments to verify that our framework can successfully capture the underlying covariate shift and interpret the concept shift simultaneously. Last but not least, we show that our proposed LSSAE has a promising generation capability to predict unseen target domains, which can be helpful to the problem related to sequential data generation. The main contributions of this paper are summarized as follows.

- We propose to focus on the problem of non-stationary evolving domain generalization where both covariate shift and concept shift may exist in the setting.
- We propose a novel probabilistic framework LSSAE which incorporates variational inference to identify the continuous latent structures of these two shifts separately and simultaneously.
- We provide empirical results to show that the proposed approach yield better results than other DG methods across scenarios. Besides, it presents a powerful generation ability of predicting unseen evolving domains.

## 2. Related Work

**Domain Generalization (DG).** The goal of DG is to learn robust models which can generalize well towards the out-of-distribution samples from unseen domains. Existing DG methods commonly rely on multiple source domains to learn representative features that can be better generalized. According to various strategies used to learn these representations, we can roughly categorize them into three categories. The first type is feature-based methods, which aim to learn domain-invariant representation which can be better generalized to target domains. Specifically, it can be achieved by aligning the distribution of representations from all source domains (Blanchard et al., 2011; Li et al., 2018b; Albuquerque et al., 2021) and feature disentanglement (Ilse et al., 2020; Wang et al., 2021; Nguyen et al., 2021). The second category is meta-learning based methods, which utilize the model agnostic training procedure to stimulate the train/test shift for acquiring generalized models (Li et al., 2018a; Balaji et al., 2018; Dou et al., 2019). Last but not least, data augmentation-based techniques, which aim to manipulate the perturbation both in original images and features to stimulate the unseen target domains, can also benefit the problem

of domain generalization (Volpi et al., 2018; Shankar et al., 2018; Zhou et al., 2021).

**Continuous Domain Adaptation (CDA).** The problem of continuous domain adaptation (i.e., evolving domain adaptation) has attracted increasing attention recently, where the CDA methods can be categorized into the intermediate-domains based methods (Kumar et al., 2020; Gong et al., 2019; Chen & Chao, 2021), domain manifold based methods (Hoffman et al., 2014; Li et al., 2017), adversary-based approaches (Wang et al., 2020; Wulfmeier et al., 2018), and meta-learning based methods (Liu et al., 2020; Lao et al., 2020). More or less, they require some samples from target domains for adaptation. In Mancini et al. (2019), they propose to substitute this reliance with some metadata from target domains as additional supervision. Instead, our focus is continuous domain generalization where no information from target domains is accessible for model learning, which is a more challenging but realistic task for real-world applications.

**Sequential Data Generation.** Recent process in unsupervised sequence generation (Yingzhen & Mandt, 2018; Han et al., 2021; Park et al., 2021) suggests the importance of decoupling time-invariant and time-variant information during the representation learning procedure. However, these approaches only take sequence generation tasks into consideration and fail to consider the category-related information, which is important for the problem of domain generalization. Unlike these approaches, we propose jointly focusing on the dynamic modeling of time-variant information on both data sample space and category space across domains.

## 3. Methodology

In this section, we first formalize the problem of evolving domain generalization based on the non-stationary environment and then describe our framework LSSAE for addressing this problem. After that, we will provide theoretical analysis for the proposed framework in Sec. 3.3 and implementation in Sec. 3.4.

### 3.1. Problem Formulation

Suppose we are given  $T$  sequentially arriving *source* domains  $\mathcal{S} = \{\mathcal{D}_1, \mathcal{D}_2, \dots, \mathcal{D}_T\}$ , where each domain  $\mathcal{D}_t = \{(x_{t,i}, y_{t,i})\}_{i=1}^{n_t}$  is comprised of  $n_t$  labeled samples for  $t \in \{1, 2, \dots, T\}$ . The goal of our problem setting is to train a classification model on  $\mathcal{S}$  which can be generalized to  $M$  following arriving *target* domains  $\mathcal{T} = \{\mathcal{D}_{T+1}, \mathcal{D}_{T+2}, \dots, \mathcal{D}_{T+M}\}$ ,  $\mathcal{D}_t = \{(x_{t,i})\}_{i=1}^{n_t}$  ( $t \in \{T+1, T+2, \dots, T+M\}$ ), which are not available during training stage. For simplicity, we omit the index  $i$  whenever  $\mathbf{x}_i$  refers to a single data point. To further quantify the continuously evolving nature of domains, we denote

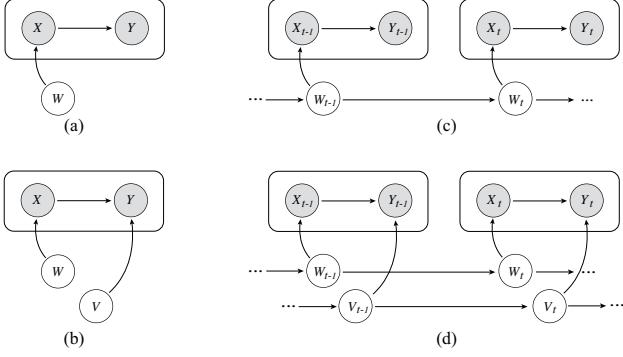


Figure 1. Comparison of causality diagram for stationary and non-stationary domain generalization scenarios. (a) represents the standard stationary DG settings which only contains covariate shift ( $P(X)$  varies for source and target domains). (b) is an extension version of (a) which contains both covariate shift and concept shift ( $P(Y|X)$  varies). (c) and (d) are corresponding non-stationary DG settings where there exist evolving patterns among adjacent domains.

$0 \leq \mathbb{D}(\mathcal{D}_t, \mathcal{D}_{t+1}) \leq \epsilon$  for two consecutive domains under some distribution distance function  $\mathbb{D}$  (e.g., Kullback-Leibler distance). In other words, the discrepancy between two consecutive domains is bounded.

Conventional DG setting only assumes that  $P(X)$  varies (i.e., covariate shift) for different domains (See Fig. 1 (a)), which may not be ideal since both  $P(X)$  and  $P(Y|X)$  can be non-stationary (i.e.,  $P(X)$  and  $P(Y|X)$  vary over time which lead to evolving covariate shift and concept shift, respectively). To tackle this problem, in this paper, we aim to explore the evolving patterns of covariate shift and concept shift across domains.

### 3.2. LSSAE: Latent Structure-aware Sequential Autoencoder

To model the dynamic in non-stationary systems, we consider two independent factors  $W$  and  $V$  (i.e.,  $W \perp\!\!\!\perp V$ ) which account for the distribution drift in data sample space (i.e., covariate shift) and data category space (i.e., concept shift) respectively according to different time stamps (See Fig. 1 (d)). For data  $(\mathbf{x}_t, \mathbf{y}_t)$  collected at time stamp  $t$ , we denote  $\mathbf{z}_t^w$  and  $\mathbf{z}_t^v$  as the latent variables of  $W$  and  $V$  at time stamp  $t$ . For completeness, we further consider time-invariant latent code  $\mathbf{z}^c$  to capture the static information of  $\mathbf{x}_t$ . We thus can define a probabilistic generative model for the joint distribution of all source domains as

$$\begin{aligned} p(\mathbf{x}_{1:T}, \mathbf{y}_{1:T}, \mathbf{z}^c, \mathbf{z}_{1:T}^w, \mathbf{z}_{1:T}^v) \\ = p(\mathbf{x}_{1:T}, \mathbf{z}_{1:T}^w, \mathbf{z}^c) p(\mathbf{y}_{1:T}, \mathbf{z}_{1:T}^v | \mathbf{z}^c). \end{aligned} \quad (1)$$

where the first term  $p(\mathbf{x}_{1:T}, \mathbf{z}_{1:T}^w, \mathbf{z}^c)$  and second term  $p(\mathbf{y}_{1:T}, \mathbf{z}_{1:T}^v | \mathbf{x}_{1:T})$  can be formulated by using Markov

chain model as

$$p(\mathbf{x}_{1:T}, \mathbf{z}_{1:T}^w, \mathbf{z}^c) = p(\mathbf{z}^c) \prod_{t=1}^T p(\mathbf{z}_t^w | \mathbf{z}_{<t}^w) \underbrace{p(\mathbf{x}_t | \mathbf{z}^c, \mathbf{z}_t^w)}_{\text{covariate shift}}, \quad (2)$$

$$p(\mathbf{y}_{1:T}, \mathbf{z}_{1:T}^v | \mathbf{z}^c) = \prod_{t=1}^T p(\mathbf{z}_t^v | \mathbf{z}_{<t}^v) \underbrace{p(\mathbf{y}_t | \mathbf{z}^c, \mathbf{z}_t^v)}_{\text{concept shift}}, \quad (3)$$

and  $p(\mathbf{x}_t | \mathbf{z}^c, \mathbf{z}_t^w)$  and  $p(\mathbf{y}_t | \mathbf{z}^c, \mathbf{z}_t^v)$  denote covariate shift and concept shift, respectively. Eq. 2 shows that the generation process of domains data  $\mathbf{x}_t$  at time stamp  $t$  depends on the corresponding dynamic latent code  $\mathbf{z}_t^w$  and static code  $\mathbf{z}^c$ , and Eq. 3 shows that the inference process (i.e., classifier to produce  $\mathbf{y}_t$ ) rely on the corresponding  $\mathbf{z}_t^v$  and  $\mathbf{z}^c$ . Our objective is to learn the classifier  $p(\mathbf{y}_t | \mathbf{z}^c, \mathbf{z}_t^v)$  which disposes of covariate shift through  $\mathbf{z}^c$  and concept shift with dynamic  $\mathbf{z}_t^v$  for the problem of evolving domain generalization.

**Domain-related module for covariate shift.** To model  $p(\mathbf{x}_{1:T}, \mathbf{z}_{1:T}^w, \mathbf{z}^c)$  where covariate shift involved, we set the prior distribution as  $p(\mathbf{z}^c) = \mathcal{N}(\mathbf{0}, \mathbf{I})$ ,  $p(\mathbf{z}_t^w | \mathbf{z}_{<t}^w) = \mathcal{N}(\boldsymbol{\mu}(\mathbf{z}_t^w), \boldsymbol{\sigma}^2(\mathbf{z}_t^w))$  which can be parameterized by some recurrent neural networks (e.g., LSTM (Hochreiter & Schmidhuber, 1997)) by setting  $\mathbf{z}_0^w = \mathbf{0}$  for initial state when  $t = 0$ , and  $p(\mathbf{x}_t | \mathbf{z}^c, \mathbf{z}_t^w)$  as a conditional decoder for the reconstruction of input data  $\mathbf{x}_t$ . To approximate the prior distributions  $p(\mathbf{z}_t^w | \mathbf{z}_{<t}^w)$ , we propose to use variational inference to learn an approximate posterior distribution  $q$  over latent variables given data which can be formulated as

$$q(\mathbf{z}_{1:T}^w, \mathbf{z}^c | \mathbf{x}_{1:T}) = q(\mathbf{z}^c | \mathbf{x}_{1:T}) \prod_{t=1}^T q(\mathbf{z}_t^w | \mathbf{z}_{<t}^w, \mathbf{x}_t), \quad (4)$$

where  $q(\mathbf{z}^c | \mathbf{x}_{1:T})$  and  $q(\mathbf{z}_t^w | \mathbf{z}_{<t}^w, \mathbf{x}_t)$  can be also parameterized by neural networks. The objective function for latent feature representation learning can be derived based on the evidence lower bound (ELBO) form (Kingma & Welling, 2014) given as

$$\begin{aligned} \mathcal{L}_d = & \sum_{t=1}^T \mathbb{E}_{q(\mathbf{z}^c | \mathbf{x}_t) q(\mathbf{z}_t^w | \mathbf{z}_{<t}^w, \mathbf{x}_t)} [\log p(\mathbf{x}_t | \mathbf{z}^c, \mathbf{z}_t^w)] \\ & - \lambda_1 \mathbb{D}_{KL}(q(\mathbf{z}^c | \mathbf{x}_{1:T}), p(\mathbf{z}^c)) \\ & - \lambda_2 \sum_{t=1}^T \mathbb{D}_{KL}(q(\mathbf{z}_t^w | \mathbf{z}_{<t}^w, \mathbf{x}_t), p(\mathbf{z}_t^w | \mathbf{z}_{<t}^w)), \end{aligned} \quad (5)$$

where the first term denotes the reconstruction term for input data  $\mathbf{x}_t$ , the second and third term denote KL divergence which are to align the posterior distributions  $\mathbf{z}^c$  and  $\mathbf{z}_t^w$  with the corresponding prior distributions.

**Category-related module for concept shift.** To model  $p(\mathbf{y}_{1:T}, \mathbf{z}_{1:T}^v | \mathbf{z}^c)$  for classification purpose where concept

shift is involved, we propose to introduce another module which can be easily integrated into our unified framework under domain generalization scenarios. Specifically, we propose to model  $p(\mathbf{y}_{1:T}, \mathbf{z}_{1:T}^v | \mathbf{z}^c)$  with a dynamic distribution  $p(\mathbf{y}_t, \mathbf{z}_t^v | \mathbf{z}^c)$  where the  $\mathbf{z}_t^v$  is varied and encode the shift information in the data category space (*e.g.*, the proportion of each category). The module can be optimized via maximizing the distribution  $p(\mathbf{y}_{1:T}, \mathbf{z}_{1:T}^v | \mathbf{z}^c)$  given a sequence of domains. In practise, we represent  $p(\mathbf{z}_t^v | \mathbf{z}_{<t}^v)$  as  $p(\mathbf{z}_t^v | \mathbf{z}_{<t}^v) = \text{Cat}(\boldsymbol{\pi}(\mathbf{z}_{<t}^v))$  which is a learnable categorical distribution. In a similar vein with  $\mathbf{z}_t^w$ , we utilize a distribution  $q$  to model the posterior distribution and approximate the prior distribution of  $\mathbf{z}_t^v$  by adopting variational inference given as

$$q(\mathbf{z}_{1:T}^v | \mathbf{y}_{1:T}) = \prod_{t=1}^T q(\mathbf{z}_t^v | \mathbf{z}_{<t}^v, \mathbf{y}_t), \quad (6)$$

where  $q(\mathbf{z}_t^v | \mathbf{z}_{<t}^v, \mathbf{y}_t)$  can be parameterized by a recurrent neural network with categorical distribution as output. We set  $\mathbf{z}_0^v = \mathbf{0}$  for initial state when  $t = 0$ . The proposed module can be jointly trained with the inference process in Eq. 6 as well as classification loss by maximizing

$$\begin{aligned} \mathcal{L}_c &= \sum_{t=1}^T \mathbb{E}_{q(\mathbf{z}_t^v | \mathbf{z}_{<t}^v, \mathbf{y}_t)} [\log p(\mathbf{y}_t | \mathbf{z}^c, \mathbf{z}_t^v)] \\ &\quad - \lambda_3 \sum_{t=1}^T \mathbb{D}_{KL}(q(\mathbf{z}_t^v | \mathbf{z}_{<t}^v, \mathbf{y}_t), p(\mathbf{z}_t^v | \mathbf{z}_{<t}^v)). \end{aligned} \quad (7)$$

Here, the first term denotes the classification loss (*i.e.*, maximizing log likelihood) and the second term denotes KL divergence which aims to align the posterior distribution of  $\mathbf{z}_t^v$  with its prior distribution.

**Temporal smooth constraint for better stability.** We empirically find that the estimation of conditional density (*i.e.*,  $p(\mathbf{z}_t^w | \mathbf{z}_{<t}^w)$  and  $p(\mathbf{z}_t^v | \mathbf{z}_{<t}^v)$ ) which is to model the complex dynamics over temporal transition may not be stable based on our formulation above. We conjecture the reason that some of the static information can be distorted by the dynamic inference module  $q(\mathbf{z}_t^w | \mathbf{z}_{<t}^w, \mathbf{x}_t)$  and  $q(\mathbf{z}_t^v | \mathbf{z}_{<t}^v, \mathbf{y}_t)$  for better reconstruction quality, which further yields sub-optimal results for recognition tasks. Intuitively, one can tackle this limitation by reducing the dimension of latent codes  $\mathbf{z}_t^w$  and  $\mathbf{z}_t^v$  or decreasing the learning rate of corresponding inference modules and prior modules manually to achieve better decoupling effect. However, we find that such strategy may not lead to desired performance. In our work, we propose to employ Lipschitz constrain over the temporal domain to stabilize the learning of the dynamic inference modules as follows

$$\begin{aligned} |q(\mathbf{z}_t^w | \mathbf{z}_{<t}^w, \mathbf{x}_t) - q(\mathbf{z}_{t-1}^w | \mathbf{z}_{<t-1}^w, \mathbf{x}_{t-1})| &\leq \alpha, \\ |q(\mathbf{z}_t^v | \mathbf{z}_{<t}^v, \mathbf{y}_t) - q(\mathbf{z}_{t-1}^v | \mathbf{z}_{<t-1}^v, \mathbf{y}_{t-1})| &\leq \alpha, \end{aligned} \quad (8)$$

where  $\alpha$  is referred to as a Lipschitz constant. The above regularization term is denoted as TS constraint for simplicity. We expect it can help with reducing the potential for volatile training.

**Objective function.** Given training data  $\mathcal{S}$ , our proposed framework can be optimized through the objective function  $\mathcal{L}_{LSSAE} = \mathcal{L}_d + \mathcal{L}_c$  with the temporal smooth constrains in Eq. 8, where the first term  $\mathcal{L}_d$  and second term  $\mathcal{L}_c$  aim to tackle the problem of covariate shift and concept shift, respectively.

**Discussion.** It is worth noting that there exists some works using probabilistic graph model for future video frame generation task (Yingzhen & Mandt, 2018; Han et al., 2021), which are similar to our proposed method at a high level. Nevertheless, our method is different since 1) our proposed framework can be applied not only to generation task (as we show in the ablation study of experimental section) but also to evolving domain generalization task (which is the main focus in our paper); 2) in order to fit our framework to the non-stationary recognition task, we introduce a novel category-related module to capture the concept shift, as such, better generalization performance can be achieved.

### 3.3. Theoretical Analysis

In this section, we aim to give a theoretical insight on our proposed method by extending variational inference from stationary environments to non-stationary environments.

**Probabilistic model for stationary environments.** We first elaborate our proposed framework based on the stationary condition, where  $\mathbf{z}^c$ ,  $\mathbf{z}^w$  and  $\mathbf{z}^v$  are introduced to capture the domain-invariant category information, domain-specific and category information, respectively. We thus have the following theorem.

**Theorem 3.1.** *For the data log likelihood  $\log p(\mathbf{x}, \mathbf{y})$  in a stationary environment, we have the evidence lower bound*

$$\begin{aligned} \max_{p,q} \mathbb{E}_{\mathbf{z}^c, \mathbf{z}^w, \mathbf{z}^v} &[\log p(\mathbf{x} | \mathbf{z}^c, \mathbf{z}^w) p(\mathbf{y} | \mathbf{z}^c, \mathbf{z}^v)] \\ &- \mathbb{D}_{KL}(q(\mathbf{z}^c | \mathbf{x}), p(\mathbf{z}^c)) - \mathbb{D}_{KL}(q(\mathbf{z}^w | \mathbf{x}), p(\mathbf{z}^w)) \\ &- \mathbb{D}_{KL}(q(\mathbf{z}^v | \mathbf{y}), p(\mathbf{z}^v)), \end{aligned} \quad (9)$$

where  $\mathbf{z}^c \sim q(\mathbf{z}^c | \mathbf{x})$ ,  $\mathbf{z}^w \sim q(\mathbf{z}^w | \mathbf{x})$ ,  $\mathbf{z}^v \sim q(\mathbf{z}^v | \mathbf{y})$ .

*Proof.* We consider the data generation procedure  $p(\mathbf{x}, \mathbf{y}, \mathbf{z}^c, \mathbf{z}^w, \mathbf{z}^v)$ . We have

$$\begin{aligned} \log p(\mathbf{x}, \mathbf{y}) &= \mathbb{D}_{KL}(q(\mathbf{z}^c, \mathbf{z}^w, \mathbf{z}^v | \mathbf{x}, \mathbf{y}), p(\mathbf{z}^c, \mathbf{z}^w, \mathbf{z}^v | \mathbf{x}, \mathbf{y})) \\ &\quad + \mathbb{E}_q \log \frac{p(\mathbf{x}, \mathbf{y}, \mathbf{z}^c, \mathbf{z}^w, \mathbf{z}^v)}{q(\mathbf{z}^c, \mathbf{z}^w, \mathbf{z}^v | \mathbf{x}, \mathbf{y})}, \end{aligned} \quad (10)$$

where the first term is the KL divergence of the approximate from the true posterior. Since this term is non-negative,

the second term is called the evidence lower bound on the marginal likelihood of data  $(\mathbf{x}, \mathbf{y})$ . This can be written as

$$\begin{aligned} \log p(\mathbf{x}, \mathbf{y}) &\geq \mathbb{E}_q \log \frac{p(\mathbf{x}, \mathbf{y}, \mathbf{z}^c, \mathbf{z}^w, \mathbf{z}^v)}{q(\mathbf{z}^c, \mathbf{z}^w, \mathbf{z}^v | \mathbf{x}, \mathbf{y})} \\ &= \mathbb{E}_{\mathbf{z}^c, \mathbf{z}^w, \mathbf{z}^v} [\log p(\mathbf{x} | \mathbf{z}^c, \mathbf{z}^w) p(\mathbf{y} | \mathbf{z}^c, \mathbf{z}^v)] \\ &\quad - \mathbb{D}_{KL}(q(\mathbf{z}^c | \mathbf{x}), p(\mathbf{z}^c)) - \mathbb{D}_{KL}(q(\mathbf{z}^w | \mathbf{x}), p(\mathbf{z}^w)) \\ &\quad - \mathbb{D}_{KL}(q(\mathbf{z}^v | \mathbf{y}), p(\mathbf{z}^v)). \end{aligned} \quad (11)$$

This completes the proof.

Theorem 3.1 shows that for a specified domain, the latent space can be decoupled into a domain invariant subspace, a domain related subspace and a category related subspace.

**Probabilistic model for non-stationary environments.** We now extend the aforementioned analysis to the non-stationary environment. Specifically, for the non-stationary environment, we can cast the dynamic variational inference framework via modeling the sequence of latent variables  $\mathbf{z}^w$  and  $\mathbf{z}^v$  as two parallel Markov chains (i.e.,  $p(\mathbf{z}^w) = p(\mathbf{z}_t^w | \mathbf{z}_{<t}^w)$  and  $p(\mathbf{z}^v) = p(\mathbf{z}_t^v | \mathbf{z}_{<t}^v)$ ). We thus have the following theorem.

**Theorem 3.2.** By denoting  $p(\mathbf{z}^w) = p(\mathbf{z}_t^w | \mathbf{z}_{<t}^w)$  and  $p(\mathbf{z}^v) = p(\mathbf{z}_t^v | \mathbf{z}_{<t}^v)$ ,  $\mathcal{L}_{LSSAE}$  is equivalent to the ELBO of the data log likelihood  $\log p(\mathbf{x}_{1:T}, \mathbf{y}_{1:T})$  based on the non-stationary environment setting.

*Proof.* We can reformulate the lower bound in Eq. 9 for a non-stationary environment as

$$\log p(\mathbf{x}_{1:T}, \mathbf{y}_{1:T}) \geq \mathbb{E}_q \log \frac{p(\mathbf{x}_{1:T}, \mathbf{y}_{1:T}, \mathbf{z}^c, \mathbf{z}_{1:T}^w, \mathbf{z}_{1:T}^v)}{q(\mathbf{z}^c, \mathbf{z}_{1:T}^w, \mathbf{z}_{1:T}^v | \mathbf{x}_{1:T}, \mathbf{y}_{1:T})}, \quad (12)$$

which can be written as

$$\begin{aligned} &\log p(\mathbf{x}_{1:T}, \mathbf{y}_{1:T}) \\ &\geq \sum_{t=1}^T \mathbb{E}_{\mathbf{z}^c, \mathbf{z}_t^w, \mathbf{z}_t^v} [\log p(\mathbf{x}_t | \mathbf{z}^c, \mathbf{z}_t^w) p(\mathbf{y} | \mathbf{x}^c, \mathbf{z}_t^v)] \\ &\quad - \mathbb{D}_{KL}(q(\mathbf{z}^c | \mathbf{x}_{1:T}), p(\mathbf{z}^c)) \\ &\quad - \sum_{t=1}^T \mathbb{D}_{KL}(q(\mathbf{z}_t^w | \mathbf{z}_{<t}^w, \mathbf{x}_t), p(\mathbf{z}_t^w | \mathbf{z}_{<t}^w)) \\ &\quad - \sum_{t=1}^T \mathbb{D}_{KL}(q(\mathbf{z}_t^v | \mathbf{z}_{<t}^v, \mathbf{y}_t), p(\mathbf{z}_t^v | \mathbf{z}_{<t}^v)), \end{aligned} \quad (13)$$

where  $\mathbf{z}^c \sim q(\mathbf{z}^c | \mathbf{x}_{1:T})$ ,  $\mathbf{z}_t^w \sim q(\mathbf{z}_t^w | \mathbf{z}_{<t}^w, \mathbf{x}_t)$  and  $\mathbf{z}_t^v \sim q(\mathbf{z}_t^v | \mathbf{z}_{<t}^v, \mathbf{y}_t)$ . The detailed derivation procedure is provided in App. A. We can see that the reconstruction part for  $\mathbf{x}_t$  in the first term together with the second and third terms can form our objective  $\mathcal{L}_d$  for domain-related module (i.e., covariate shift), and the combination of the reconstruction part for  $\mathbf{y}_t$  in the first term and the last term can form  $\mathcal{L}_c$  for category-related module (i.e., concept shift). As a result, the formulation above is equivalent to our objective  $\mathcal{L}_{LSSAE}$ .

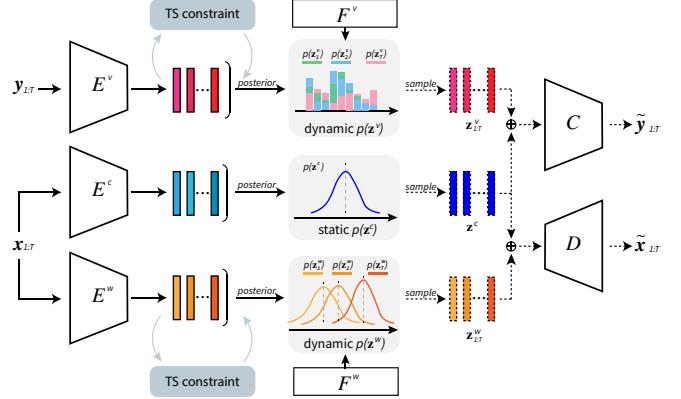


Figure 2. An overview of network architecture for LSSAE. Our framework consists of the static variational encoding network  $E^c$ , dynamic variational encoding networks  $E^w$  and  $E^v$ , dynamic prior networks  $F^w$  and  $F^v$ , a decoder  $D$  and a classifier  $C$ . It is worth noting that we do not require  $E^v$  (i.e., only data from target domain  $\mathcal{T}$  available) during inference stage.

### 3.4. Implementation

The implementation of network architecture for LSSAE is depicted in Fig. 2. It is composed of two parts: (1) the domain-related module (middle and bottom region); (2) the category-related module (top region).

The domain-related module consists of a static variational encoding network  $E^c$  for  $q(\mathbf{z}^c | \mathbf{x}_{1:T})$ , a dynamic variational encoding network  $E^w$  associated with  $q(\mathbf{z}_t^w | \mathbf{z}_{<t}^w, \mathbf{x}_t)$ , a dynamic prior network  $F^w$  working for  $p(\mathbf{z}_t^w | \mathbf{z}_{<t}^w)$  and a decoder  $D$  corresponding to  $p(\mathbf{x}_t | \mathbf{z}^c, \mathbf{z}_t^w)$ . Similar to VAE (Kingma & Welling, 2014), we can apply the reparameterization trick (Kingma & Welling, 2014) to optimize parameters of  $E^c$ ,  $E^w$  and  $F^w$ . Specifically, we implement  $E^c$  by a feature extractor which will be also utilized to extract features during test time.  $E^w$  can be implemented by a feature extractor with same architecture of  $E^c$  but not sharing network parameters, and followed by a LSTM network.  $F^w$  is implemented as a one-layer LSTM network. For the category-related module, we design a dynamic inference network  $E^v$  which takes the one-hot code of label  $y_t$  as the input and with the output of the categorical distribution  $q(\mathbf{z}_t^v | \mathbf{z}_{<t}^v, \mathbf{y}_t)$ , and a classifier  $C$  which takes  $\mathbf{z}^c$  and  $\mathbf{z}_t^v$  as the input for  $p(\mathbf{y}_t | \mathbf{z}^c, \mathbf{z}_t^v)$ . Similarly, the prior network  $F^v$  for  $p(\mathbf{z}_t^v | \mathbf{z}_{<t}^v)$  is a LSTM network with a categorical distribution as the output. After that, we use Gumbel-Softmax reparameterization trick (Jang et al., 2017; Maddison et al., 2017) to sample  $\mathbf{z}_t^v$  for optimization. Regarding the classifier  $C$ , we utilize a linear layer by following Gulrajani & Lopez-Paz (2021). The details of our architecture and hyperparameters for objective function can be found in the supplementary material.

**Optimization.** We first initialize  $\mathbf{z}^w$  and  $\mathbf{z}^v$  as  $\mathbf{z}_0^w = \mathbf{0}$  and

$\mathbf{z}_0^v = \mathbf{0}$ , respectively. To train our framework, we generate the dynamic prior distributions  $p(\mathbf{z}_t^w | \mathbf{z}_{<t}^w)$  and  $p(\mathbf{z}_t^v | \mathbf{z}_{<t}^v)$  through  $F^w$  and  $F^v$  respectively based on  $T$  source domains  $\mathcal{S}$ . For time stamp  $t$ , we sample a batch of data  $\mathbf{x}_t$  and take it as input for  $E^c$  and  $E^w$  to obtain the parameters of their posterior distributions. After that, the latent features  $\mathbf{z}^c$  and  $\mathbf{z}_t^w$  are resampled separately through reparameterization trick and then concatenated together. Finally the decoder  $D$  outputs the reconstruction data of  $\mathbf{x}_t$ . Meanwhile,  $E^v$  takes the corresponding labels  $\mathbf{y}_t$  of  $\mathbf{x}_t$  in one-hot format as input and outputs the latent features, and then the latent features are also resampled through reparameterization trick to obtain  $\mathbf{z}_t^v$ . The classifier  $C$  takes  $\mathbf{z}^c$  and  $\mathbf{z}_t^v$  as the input to predict the labels of  $\mathbf{x}_t$ . During training, we sample a mini-batch from each single domain with the same number of data sample to form a large batch in order to suit our framework to the temporal smooth constraint for stable training. This optimization procedure of LSSAE is depicted in Algorithm 1.

**Inference.** To predict the label of  $\mathbf{x}_t$  sampled from one of the target domains in  $\mathcal{D}_t$  in  $\mathcal{T}$ , we adopt  $F^v$  to infer the latent code  $\mathbf{z}_t^v$  and  $E^c$  to extract the latent features  $\mathbf{z}^c$ , and then apply the classifier  $C$  for the prediction purpose. It is worth noting that we do not require  $E^v$  (i.e., only data from target domain  $\mathcal{T}$  available) during inference stage. We show this procedure in Algorithm 2.

## 4. Experiments

In this section, we present experimental results to validate the effectiveness of our proposed LSSAE based on the setting of evolving domain generalization.

### 4.1. Experimental Setup

We compare the proposed LSSAE with other DG methods on two synthetic datasets (Circle and Sine) and four real-world datasets (Rotated MNIST, Portraits, Caltran, Power-Supply). We also evaluate the results on two variants named Circle-C and Sine-C derived from Circle and Sine via synthesizing concept shift manually. We split the domains into source domains, intermediate domains and target domains with the ratio of  $\{1/2 : 1/6 : 1/3\}$ . The intermediate domains are utilized as validation set. More details of dataset construction can be found in the supplementary material.

(1) **Circle/-C** (Pesaranahader & Viktor, 2016). This dataset contains evolving 30 domains where the instances are sampled from 30 2D Gaussian distributions. For Circle-C, concept shift is introduced via changing the center and radius of decision boundary in a gradual manner over time.

(2) **Sine/-C** (Pesaranahader & Viktor, 2016). We rearrange this dataset by extending it to 24 evolving domains. To simulate concept drift for Sine-C, labels are reversed (i.e.,

---

### Algorithm 1 Optimization procedure for LSSAE

---

**Input:** sequential source labeled datasets  $\mathcal{S}$ ; static feature extractor  $E^c$ ; dynamic inference networks  $E^w, E^v$  and their corresponding prior networks  $F^w, F^v$ ; decoder  $D$  and classifier  $C$ .  
 Randomly initialize  $E^c, E^w, E^v, F^w, F^v, D, C$   
 Assign  $\mathbf{z}_0^w, \mathbf{z}_0^v \leftarrow \mathbf{0}$   
**for**  $t = 1, 2, \dots, K$  **do**  
     Generate prior distribution  $p(\mathbf{z}_t^w | \mathbf{z}_{<t}^w)$  via  $F^w$   
     Generate prior distribution  $p(\mathbf{z}_t^v | \mathbf{z}_{<t}^v)$  via  $F^v$   
     **for**  $i = 1, 2, \dots$  **do**  
         Sample a batch of data  $(\mathbf{x}_t, \mathbf{y}_t)$  from  $\mathcal{S}_t$   
         ▷ Calculate  $\mathcal{L}_d$  by Eq. 5 for  $E^c, E^w, D$  and  $F^w$   
         ▷ Calculate  $\mathcal{L}_c$  by Eq. 7 for  $E^c, E^v, C$  and  $F^v$   
         ▷ Calculate temporal smooth constriction by Eq. 8  
         Update all modules by the summary of these loss  
     **end for**  
**end for**

---

### Algorithm 2 Inference procedure for LSSAE

---

**Input:** sequential target datasets  $\mathcal{T}$ ; static feature extractor  $E^c$ ; dynamic prior network  $F^v$  and classifier  $C$ .  
 Assign  $\mathbf{z}_0^v \leftarrow \mathbf{0}$   
**for**  $t = 1, 2, \dots$  **do**  
     Sample  $\mathbf{z}_t^v \sim p(\mathbf{z}_t^v | \mathbf{z}_{<t}^v)$  via  $F^v$   
     **for**  $i = 1, 2, \dots$  **do**  
         ▷ Extract the feature for data  $\mathbf{x}_t$  via  $E^c$   
         ▷ Generate the prediction  $\tilde{\mathbf{y}}_t$  via  $C$   
     **end for**  
**end for**

---

from 0 to 1 or from 1 to 0) from the 6-th domain to the last one.

(3) **Rotated MNIST (RMNIST)** (Ghifary et al., 2015). Rotated MNIST (RMNIST) is composed of MNIST digits of various rotations. We extend it to 19 evolving domains via applying the rotations with degree of  $\mathcal{R} = \{0^\circ, 15^\circ, 30^\circ, \dots, 180^\circ\}$  in order.

(4) **Portraits** (Ginosar et al., 2015). This dataset comprises photos of high-school seniors from the 1905s to the 2005s for gender classification. We split the dataset into 34 domains by a fixed internal over time.

(5) **Caltran** (Hoffman et al., 2014). Caltran is a real-world surveillance dataset comprising images captured from a fixed traffic camera deployed in an intersection. The task is to predict the type of scene based on continuously evolving data. We divide it into 34 domains based on different times.

(6) **PowerSupply** (Dau et al., 2019). PowerSupply is constructed for the time-section prediction of current power supply based on the hourly records of an Italy electricity company. The concept shift may raise from the change in

Table 1. The comparison of accuracy (%) between LSSAE and other DG baselines on various datasets. As DG baselines do not take concept shift into consideration, we report the results on datasets without concept shift and with concept shift separately.

Algorithm	Circle	Sine	RMNIST	Portraits	Caltran	Avg	Circle-C	Sine-C	PowerSupply	Avg
ERM	49.9	63.0	43.6	87.8	66.3	62.1	34.0	61.5	71.0	55.5
Mixup	48.4	62.9	44.9	87.8	66.0	62.0	33.9	60.9	70.8	55.2
MMD	50.7	55.8	44.8	87.3	57.1	59.1	33.7	52.7	70.9	52.4
MLDG	50.8	63.2	43.1	88.5	66.2	62.3	34.6	62.0	70.8	55.8
IRM	51.3	63.2	39.0	85.4	64.1	60.6	38.5	61.2	70.8	56.8
RSC	48.0	61.5	41.7	87.3	67.0	61.1	33.7	61.5	70.9	55.4
MTL	51.2	62.9	41.7	89.0	68.2	62.6	33.9	61.4	70.7	55.3
Fish	48.8	62.3	44.2	88.8	68.6	62.5	34.3	<b>62.7</b>	70.8	55.9
CORAL	53.9	51.6	44.5	87.4	65.7	60.6	34.1	59.0	71.0	54.7
AndMask	47.9	69.3	42.8	70.3	56.9	57.4	37.7	52.7	70.7	53.7
DIVA	67.9	52.9	42.7	88.2	69.2	64.2	33.9	52.9	70.8	55.1
LSSAE (Ours)	<b>73.8</b>	<b>71.4</b>	<b>46.4</b>	<b>89.1</b>	<b>70.6</b>	<b>70.3</b>	<b>44.8</b>	60.8	<b>71.1</b>	<b>58.9</b>

season, weather or price. We form 30 domains according to days.

The methods for comparison include: (1) ERM (Vapnik, 1998); (2) Mixup (Yan et al., 2020); (3) MMD (Li et al., 2018b); (4) MLDG (Li et al., 2018a); (5) IRM (Rosenfeld et al., 2021); (6) RSC (Huang et al., 2020); (7) MTL (Blanchard et al., 2021); (8) Fish (Shi et al., 2021); (9) CORAL (Sun & Saenko, 2016); (10) AndMask (Parascandolo et al., 2021); (11) DIVA (Ilse et al., 2020). All of our experiments are implemented in the PyTorch platform based on DomainBed package (Gulrajani & Lopez-Paz, 2021). For a fair comparison, we keep the neural network architecture of encoding part and classification part to be same for all baselines for different benchmarks.

## 4.2. Quantitative Results

The results of our proposed LSSAE and baselines are presented in Table 1. As conventional DG methods focus upon covariate shift only, we separate the datasets according to with or without concept shift into two parts for fairness. We can see that LSSAE consistently outperforms other baselines over all datasets, it achieves 70.3% accuracy when there exists covariate shift only (Circle, Sine, RMNIST, Portraits and Caltran), and achieves 58.9% accuracy when there exist concept shift (Circle-C, Sine-C, PowerSupply). The results are significantly better than the compared DG approaches, which are reasonable since existing DG methods cannot deal with distribution shift well in non-stationary environments but our proposed LSSAE can properly capture the evolving patterns to gain better performance. More results can be found in the supplementary materials.

To better understand the rationality of our method, we visualize the decision boundaries of our method and ERM baseline on two synthetic datasets: Circle and Sine by comparing our proposed method with ERM baseline. The visualization results are depicted in Fig. 3 and Fig. 4. As

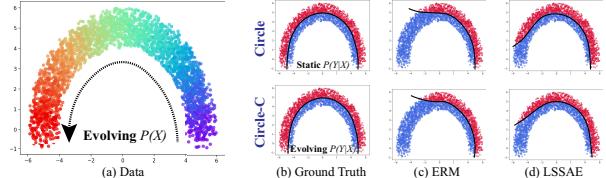


Figure 3. Decision boundary visualization of Circle and Circle-C datasets each with 30 domains. (a) presents the original data in different domains by color, where the right half part are source domains. (b) shows the positive and negative labels in red and blue dots. (c) and (d) are decision boundaries learned by ERM and LSSAE, respectively.

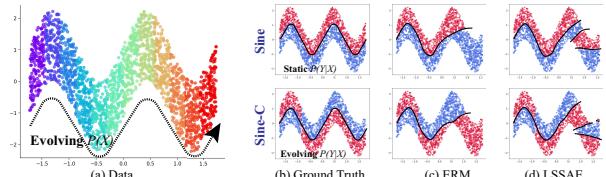


Figure 4. Decision boundary visualization of Sine and Sine-C datasets. (a) presents 24 domains indexed by different colors, where the left half part are source domains. (b) shows the positive and negative labels in red and blue dots. (c) and (d) are decision boundaries learned by ERM and LSSAE, respectively.

shown in Fig. 3, both of ERM and LSSAE can fit the source domains well. However, different from ERM which only fit the source domains, LSSAE shows a desired generalization ability to unseen target domains. This validates that our LSSAE can capture the underlying evolving patterns across domains to achieve better results. As for Circle-C where we introduce evolving concept shift via modifying the center and radius of decision boundary over a period of time, we observe that our proposed LSSAE can still produce a more accurate decision boundary compared with ERM. Similar observation can be found in Fig. 4, where LSSAE can be better generalized to evolving domains compared with ERM. However, we find that our proposed LSSAE may not be able

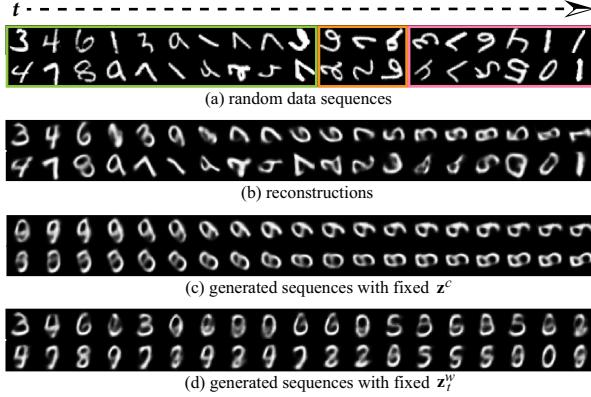


Figure 5. Visualisation of generated and reconstructed data sequences on RMNIST dataset.

to obtain a desired boundary based on the setting of Sine-C, where we introduce concept shift by reversing the labels. We conjecture the reason that our proposed LSSAE may not be able to well handle the abrupt concept shift where no continuous evolving pattern exists.

#### 4.3. Ablation Study

**Analysis on domain-related module.** To better evaluate whether our proposed LSSAE can capture domain specific information, we conduct experiments by evaluating the reconstruction and generation capability of LSSAE on RMNIST dataset. The reconstruction and generation results are shown in Fig. 5. Each subfigure shows two sequences which follow the domain order (i.e., rotation degree gradually evolves) from the left to the right. Subfigure (a) shows the original data sequence sampled from source domains (green bounding box), intermediate domains (orange bounding box) and target domains (pink bounding box) in order. Noted that we are only interested in degree of rotation thus the category of images from different domains may not be consistent. Subfigure (b) shows the reconstructions for each sample in Subfigure (a). Based on the results, we find that we can achieve a desired reconstruction quality, which suggests that the domain-related information can be captured even we random the category related information. To further show that our proposed method is capable for generation task, in subfigure (c), we visualize the randomly generated samples using  $\mathbf{z}_t^w \sim p(\mathbf{z}_t^w | \mathbf{z}_{<t}^w)$  while keeping  $\mathbf{z}^c$  the same for all domains. We can see that the category information (i.e., digit information) remains unchanged when we fix  $\mathbf{z}^c$ , and can generated samples for unseen future domains where the rotation degree of digit is gradually changed with evolving prior  $p(\mathbf{z}_t^w | \mathbf{z}_{<t}^w)$ . This observation suggests that our proposed method can be used for data augmentation when adapting to the unseen target domains. However, we also observe that there exists image quality distortion when generating new domains. We conjecture the reason that the

Table 2. Comparison of different prior distributions for category-related module on PowerSupply dataset.

Prior Type	Accuracy (%)
Without $\mathbf{z}^v$	70.7
Gaussian	70.1
Uniform	71.0
Categorical (Ours)	<b>71.1</b>

range of degrees for training may not be sufficiently diverse, which limits the performance of generation to unseen digit rotation degrees. Subfigure (d) shows randomly generated data via sampling  $\mathbf{z}^c \sim p(\mathbf{z}^c)$  while keeping  $\mathbf{z}_t^w$  the same for all domains at different time stamp  $t$  (i.e.,  $\mathbf{z}_t^w = \mathbf{z}_1^w$ ), where we find that the generated digit images belong to different categories but with the same rotation degree, which further indicates that our proposed method can successfully extract domain related information.

**Analysis on category-related module.** We then evaluate the effectiveness of our proposed category-related module on PowerSupply dataset. To this end, we consider three different ablation studies by 1) removing category-related module, 2) replacing the learnable categorical prior with a learnable Gaussian prior, 3) replacing the learnable categorical prior with a fix Uniform prior. The results are shown in Table 2. As we can see, our proposed LSSAE based on categorical prior can achieve the best performance among other baselines, which shows the effectiveness of our proposed method. However, we observe that the performance drops to some extent by replacing the categorical prior with Gaussian prior, which we conjecture the reason that categorical distribution can better capture the category related information which is discrete. We also observe that better performance can be achieved by comparing with the results using Uniform distribution as prior, which is reasonable since Uniform distribution does not change over time.

**Analysis on temporal smooth constraint.** The smooth constraint is mainly designed for stabilizing the training procedure. To verify this, we conduct experiments by considering “with & without this constraint” on RMNIST and CalTran datasets while keeping other hyperparameters the same. The results are reported as the variance of the test accuracy in the last five epochs (See Table 3). For RMNIST, when training with constraint, the variance is 0.4; without constraint, the variance is 3.6. For CalTran, when training with constraint, the variance is 4.7, and without constraint, the variance is 10.0. We can see that there exists an obvious decrease in terms of variance when incorporating with our proposed TS constraint. Besides, as our smooth constraint helps stabilize the training process, better performance can be expected. For RMNIST, training with our smooth constraint can achieve 0.9% performance gain; For CalTran,

Table 3. Ablation study on temporal smooth constraint.

TS Constraint	RMNIST		CalTran	
	Var ↓	Acc (%) ↑	Var ↓	Acc (%) ↑
✗	3.6	45.5	10.0	69.9
✓	0.4	46.4	4.7	70.6

the performance gain is 0.7%.

## 5. Conclusion

In this paper, we propose to focus on the problem of evolving domain generalization, where the covariate shift and concept shift vary over time. To tackle this problem, we propose a novel framework LSSAE to model the dynamics of distribution shift (i.e., covariate shift and concept shift). We also provide theoretical analysis, which shows that our proposed method is equivalent to maximizing the ELBO based on the non-stationary environment setting, and justifies the rationality of our proposed method for the problem of evolving domain generalization. Experimental results on both toy data and real-world datasets across multiple domains further indicate the significance of our proposed method based on this setting.

## Acknowledgements

This work was supported in part by CityU New Research Initiatives/Infrastructure Support from Central (APRC 9610528), CityU Applied Research Grant (ARG 9667244) and Hong Kong Innovation and Technology Commission (InnoHK Project CIMDA). Besides, this work was also supported in part by the National Natural Science Foundation of China under 62022002, in part by the Hong Kong Research Grants Council General Research Fund (GRF) under Grant 11203220.

## References

- Albuquerque, I., Monteiro, J., Darvishi, M., Falk, T. H., and Mitliagkas, I. Generalizing to unseen domains via distribution matching. *arXiv preprint arXiv:1911.00804*, 2021.
- Balaji, Y., Sankaranarayanan, S., and Chellappa, R. Metareg: Towards domain generalization using meta-regularization. *Advances in Neural Information Processing Systems*, 31: 998–1008, 2018.
- Blanchard, G., Lee, G., and Scott, C. Generalizing from several related classification tasks to a new unlabeled sample. *Advances in Neural Information Processing Systems*, 24: 2178–2186, 2011.
- Blanchard, G., Deshmukh, A. A., Dogan, Ü., Lee, G., and Scott, C. Domain generalization by marginal transfer learning. *J. Mach. Learn. Res.*, 22:2–1, 2021.
- Chen, H.-Y. and Chao, W.-L. Gradual domain adaptation without indexed intermediate domains. In *Thirty-Fifth Conference on Neural Information Processing Systems*, 2021.
- Dau, H. A., Bagnall, A., Kamgar, K., Yeh, C.-C. M., Zhu, Y., Gharghabi, S., Ratanamahatana, C. A., and Keogh, E. The ucr time series archive. *IEEE/CAA Journal of Automatica Sinica*, 6(6):1293–1305, 2019.
- Dou, Q., Coelho de Castro, D., Kamnitsas, K., and Glocker, B. Domain generalization via model-agnostic learning of semantic features. *Advances in Neural Information Processing Systems*, 32:6450–6461, 2019.
- Federici, M., Tomioka, R., and Forré, P. An information-theoretic approach to distribution shifts. In *Thirty-Fifth Conference on Neural Information Processing Systems*, 2021.
- Ghifary, M., Kleijn, W. B., Zhang, M., and Balduzzi, D. Domain generalization for object recognition with multi-task autoencoders. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2551–2559, 2015.
- Ginosar, S., Rakelly, K., Sachs, S., Yin, B., and Efros, A. A. A century of portraits: A visual historical record of american high school yearbooks. In *2015 IEEE International Conference on Computer Vision Workshop (ICCVW)*, pp. 652–658, 2015.
- Gong, R., Li, W., Chen, Y., and Gool, L. V. Dlow: Domain flow for adaptation and generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2477–2486, 2019.
- Gulrajani, I. and Lopez-Paz, D. In search of lost domain generalization. In *International Conference on Learning Representations*, 2021.
- Han, J., Min, M. R., Han, L., Li, L. E., and Zhang, X. Disentangled recurrent wasserstein autoencoder. In *International Conference on Learning Representations*, 2021.
- Hochreiter, S. and Schmidhuber, J. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.
- Hoffman, J., Darrell, T., and Saenko, K. Continuous manifold based adaptation for evolving visual domains. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 867–874, 2014.
- Huang, Z., Wang, H., Xing, E. P., and Huang, D. Self-challenging improves cross-domain generalization. In

- Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pp. 124–140, 2020.
- Ilse, M., Tomczak, J. M., Louizos, C., and Welling, M. Diva: Domain invariant variational autoencoders. In *Medical Imaging with Deep Learning*, pp. 322–348. PMLR, 2020.
- Jang, E., Gu, S., and Poole, B. Categorical reparameterization with gumbel-softmax. *International Conference on Learning Representations*, 2017.
- Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.
- Kingma, D. P. and Welling, M. Auto-encoding variational bayes. *International Conference on Learning Representations*, 2014.
- Kumar, A., Ma, T., and Liang, P. Understanding self-training for gradual domain adaptation. In *International Conference on Machine Learning*, pp. 5468–5479. PMLR, 2020.
- Lao, Q., Jiang, X., Havaei, M., and Bengio, Y. Continuous domain adaptation with variational domain-agnostic feature replay. *arXiv preprint arXiv:2003.04382*, 2020.
- Li, D., Yang, Y., Song, Y.-Z., and Hospedales, T. M. Learning to generalize: Meta-learning for domain generalization. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018a.
- Li, H., Pan, S. J., Wang, S., and Kot, A. C. Domain generalization with adversarial feature learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5400–5409, 2018b.
- Li, W., Xu, Z., Xu, D., Dai, D., and Van Gool, L. Domain generalization and adaptation using low rank exemplar svms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(5):1114–1127, 2017.
- Liu, H., Long, M., Wang, J., and Wang, Y. Learning to adapt to evolving domains. In *Advances in Neural Information Processing Systems*, volume 33, pp. 22338–22348, 2020.
- Maddison, C. J., Mnih, A., and Teh, Y. W. The concrete distribution: A continuous relaxation of discrete random variables. *International Conference on Learning Representations*, 2017.
- Mancini, M., Bulo, S., Caputo, B., and Ricci, E. Adagraph: Unifying predictive and continuous domain adaptation through graphs. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6561–6570, jun 2019.
- Muandet, K., Balduzzi, D., and Schölkopf, B. Domain generalization via invariant feature representation. In *International Conference on Machine Learning*, pp. 10–18, 2013.
- Nguyen, A. T., Tran, T., Gal, Y., and Baydin, A. G. Domain invariant representation learning with domain density transformations. *arXiv preprint arXiv:2102.05082*, 2021.
- Parascandolo, G., Neitz, A., ORVIETO, A., Gresele, L., and Schölkopf, B. Learning explanations that are hard to vary. In *International Conference on Learning Representations*, 2021.
- Park, S. W., Shu, D. W., and Kwon, J. Generative adversarial networks for markovian temporal dynamics: Stochastic continuous data generation. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pp. 8413–8421, 18–24 Jul 2021.
- Pesaranghader, A. and Viktor, H. L. Fast hoeffding drift detection method for evolving data streams. In *ECML/PKDD*, 2016.
- Rosenfeld, E., Ravikumar, P. K., and Risteski, A. The risks of invariant risk minimization. In *International Conference on Learning Representations*, 2021.
- Shankar, S., Piratla, V., Chakrabarti, S., Chaudhuri, S., Jyothi, P., and Sarawagi, S. Generalizing across domains via cross-gradient training. In *International Conference on Learning Representations*, 2018.
- Shi, Y., Seely, J., Torr, P. H., Siddharth, N., Hannun, A., Usunier, N., and Synnaeve, G. Gradient matching for domain generalization. *arXiv preprint arXiv:2104.09937*, 2021.
- Sugiyama, M., Yamada, M., and du Plessis, M. C. Learning under nonstationarity: covariate shift and class-balance change. *Wiley Interdisciplinary Reviews: Computational Statistics*, 5(6):465–477, 2013.
- Sun, B. and Saenko, K. Deep coral: Correlation alignment for deep domain adaptation. In Hua, G. and Jégou, H. (eds.), *Computer Vision – ECCV 2016 Workshops*, pp. 443–450, Cham, 2016.
- Tahmasbi, A., Jothimurugesan, E., Tirthapura, S., and Gibbons, P. B. Driftsurf: Stable-state/reactive-state learning under concept drift. In *International Conference on Machine Learning*, pp. 10054–10064. PMLR, 2021.
- Torralba, A. and Efros, A. A. Unbiased look at dataset bias. In *CVPR 2011*, pp. 1521–1528. IEEE, 2011.
- Vapnik, V. Statistical learning theory wiley. *New York*, 1998.

- Vergara, A., Vembu, S., Ayhan, T., Ryan, M. A., Homer, M. L., and Huerta, R. Chemical gas sensor drift compensation using classifier ensembles. *Sensors and Actuators B: Chemical*, 166:320–329, 2012.
- Volpi, R., Namkoong, H., Sener, O., Duchi, J. C., Murino, V., and Savarese, S. Generalizing to unseen domains via adversarial data augmentation. In *Advances in Neural Information Processing Systems*, volume 31, 2018.
- Wang, H., He, H., and Katabi, D. Continuously indexed domain adaptation. In III, H. D. and Singh, A. (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119, pp. 9898–9907, 13–18 Jul 2020.
- Wang, Y., Li, H., Chau, L.-P., and Kot, A. C. Variational disentanglement for domain generalization. *arXiv preprint arXiv:2109.05826*, 2021.
- Wulfmeier, M., Bewley, A., and Posner, I. Incremental adversarial domain adaptation for continually changing environments. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 4489–4495. IEEE, 2018.
- Yan, S., Song, H., Li, N., Zou, L., and Ren, L. Improve unsupervised domain adaptation with mixup training. *arXiv preprint arXiv:2001.00677*, 2020.
- Yingzhen, L. and Mandt, S. Disentangled sequential autoencoder. In *International Conference on Machine Learning*, pp. 5670–5679. PMLR, 2018.
- Zhou, K., Yang, Y., Qiao, Y., and Xiang, T. Domain generalization with mixstyle. *arXiv preprint arXiv:2104.02008*, 2021.

## A. Proofs

### A.1. Derivation of ELBO for Stationary Environments

As we introduce two latent variables to account for the two types of distribution shift, the data generating procedure for one domain can be expressed as:

$$\begin{aligned} p(\mathbf{x}, \mathbf{y}, \mathbf{z}^c, \mathbf{z}^w, \mathbf{z}^v) &= p(\mathbf{z}^c)p(\mathbf{z}^w)p(\mathbf{z}^v)p(\mathbf{x}|\mathbf{z}^c, \mathbf{z}^w)p(\mathbf{y}|\mathbf{z}^c, \mathbf{z}^v) \\ &= p(\mathbf{z}^w)p(\mathbf{z}^v) \prod_{i=1}^N p(z_i^c)p(x_i|z_i^c, \mathbf{z}^w)p(y_i|z_i^c, \mathbf{z}^v) \end{aligned} \quad (14)$$

We let  $(\mathbf{z}^c, \mathbf{z}^w)$  be the latent variables for  $\mathbf{x}$ , and  $\mathbf{z}^v$  be latent variable for  $\mathbf{y}$ , thus the distribution of these three latent variables can be inferred from the observable datapoints as  $p(\mathbf{z}^c|\mathbf{x}), p(\mathbf{z}^w|\mathbf{x}), p(\mathbf{z}^v|\mathbf{y})$ , respectively. The joint distribution of latent variables is:

$$p(\mathbf{z}^c, \mathbf{z}^w, \mathbf{z}^v|\mathbf{x}, \mathbf{y}) = p(\mathbf{z}^c|\mathbf{x})p(\mathbf{z}^w|\mathbf{x})p(\mathbf{z}^v|\mathbf{y}) \quad (15)$$

For our approximating distribution in Eq 15, we can choose the form  $q(\mathbf{z}^c, \mathbf{z}^w, \mathbf{z}^v|\mathbf{x}, \mathbf{y}) = q(\mathbf{z}^c|\mathbf{x})q(\mathbf{z}^w|\mathbf{x})q(\mathbf{z}^v|\mathbf{y})$ . Thus, we can get:

$$\begin{aligned} \mathbb{D}_{KL}(q, p) &= \mathbb{E}_q \log \frac{q(\mathbf{z}^c, \mathbf{z}^w, \mathbf{z}^v|\mathbf{x}, \mathbf{y})}{p(\mathbf{z}^c, \mathbf{z}^w, \mathbf{z}^v|\mathbf{x}, \mathbf{y})} \\ &= \mathbb{E}_q \log \frac{q(\mathbf{z}^c, \mathbf{z}^w, \mathbf{z}^v|\mathbf{x}, \mathbf{y})}{p(\mathbf{x}, \mathbf{y}, \mathbf{z}^c, \mathbf{z}^w, \mathbf{z}^v)} p(\mathbf{x}, \mathbf{y}) \\ &= \mathbb{E}_q \log \frac{q(\mathbf{z}^c, \mathbf{z}^w, \mathbf{z}^v|\mathbf{x}, \mathbf{y})}{p(\mathbf{x}, \mathbf{y}, \mathbf{z}^c, \mathbf{z}^w, \mathbf{z}^v)} + \mathbb{E}_q \log p(\mathbf{x}, \mathbf{y}) \\ &= \mathbb{E}_q \log \frac{q(\mathbf{z}^c, \mathbf{z}^w, \mathbf{z}^v|\mathbf{x}, \mathbf{y})}{p(\mathbf{x}, \mathbf{y}, \mathbf{z}^c, \mathbf{z}^w, \mathbf{z}^v)} + \log p(\mathbf{x}, \mathbf{y}) \end{aligned} \quad (16)$$

We can get:

$$\log p(\mathbf{x}, \mathbf{y}) = \mathbb{D}_{KL}(q, p) + \mathbb{E}_q \log \frac{p(\mathbf{x}, \mathbf{y}, \mathbf{z}^c, \mathbf{z}^w, \mathbf{z}^v)}{q(\mathbf{z}^c, \mathbf{z}^w, \mathbf{z}^v|\mathbf{x}, \mathbf{y})} \quad (17)$$

As  $\mathbb{D}_{KL}(q, p) \geq 0$ , the variational lower bound for  $\log p(\mathbf{x}, \mathbf{y})$  is:

$$\mathcal{L} = \mathbb{E}_q \log \frac{p(\mathbf{x}, \mathbf{y}, \mathbf{z}^c, \mathbf{z}^w, \mathbf{z}^v)}{q(\mathbf{z}^c, \mathbf{z}^w, \mathbf{z}^v|\mathbf{x}, \mathbf{y})}. \quad (18)$$

This formulation can be reorganized as:

$$\begin{aligned} \mathcal{L} &= \mathbb{E}_q \log \frac{p(\mathbf{x}, \mathbf{y}|\mathbf{z}^c, \mathbf{z}^w, \mathbf{z}^v)p(\mathbf{z}^c)p(\mathbf{z}^w)p(\mathbf{z}^v)}{q(\mathbf{z}^c|\mathbf{x})q(\mathbf{z}^w|\mathbf{x})q(\mathbf{z}^v|\mathbf{y})} \\ &= \mathbb{E}_q \log p(\mathbf{x}, \mathbf{y}|\mathbf{z}^c, \mathbf{z}^w, \mathbf{z}^v) + \mathbb{E}_q \log \frac{p(\mathbf{z}^c)}{q(\mathbf{z}^c|\mathbf{x})} + \mathbb{E}_q \log \frac{p(\mathbf{z}^w)}{q(\mathbf{z}^w|\mathbf{x})} + \mathbb{E}_q \log \frac{p(\mathbf{z}^v)}{q(\mathbf{z}^v|\mathbf{y})} \\ &= \mathbb{E}_q \log p(\mathbf{x}, \mathbf{y}|\mathbf{z}^c, \mathbf{z}^w, \mathbf{z}^v) - \mathbb{D}_{KL}(q(\mathbf{z}^c|\mathbf{x}), p(\mathbf{z}^c)) - \mathbb{D}_{KL}(q(\mathbf{z}^w|\mathbf{x}), p(\mathbf{z}^w)) - \mathbb{D}_{KL}(q(\mathbf{z}^v|\mathbf{y}), p(\mathbf{z}^v)). \end{aligned} \quad (19)$$

As  $p(\mathbf{x}, \mathbf{y}|\mathbf{z}^c, \mathbf{z}^w, \mathbf{z}^v) = p(\mathbf{x}|\mathbf{z}^c, \mathbf{z}^w)p(\mathbf{y}|\mathbf{z}^c, \mathbf{z}^v)$ , the above formulation can be rewrote as:

$$\mathcal{L} \geq \mathbb{E}_q \log p(\mathbf{x}|\mathbf{z}^c, \mathbf{z}^w)p(\mathbf{y}|\mathbf{z}^c, \mathbf{z}^v) - \mathbb{D}_{KL}(q(\mathbf{z}^c|\mathbf{x}), p(\mathbf{z}^c)) - \mathbb{D}_{KL}(q(\mathbf{z}^w|\mathbf{x}), p(\mathbf{z}^w)) - \mathbb{D}_{KL}(q(\mathbf{z}^v|\mathbf{y}), p(\mathbf{z}^v)). \quad (20)$$

### A.2. Derivation of ELBO for Non-stationary Environments

We assume the prior distribution of latent variables  $\mathbf{z}^w$  and  $\mathbf{z}^v$  satisfy Markov property. This means each of them relies on the value of their previous states:

$$p(\mathbf{z}^w) = p(\mathbf{z}_t^w|\mathbf{z}_{<t}^w), p(\mathbf{z}^v) = p(\mathbf{z}_t^v|\mathbf{z}_{<t}^v), \quad (21)$$

The joint distribution of data and latent variables is:

$$\begin{aligned}
 & p(\mathbf{x}_{1:T}, \mathbf{y}_{1:T}, \mathbf{z}^c, \mathbf{z}_{1:T}^w, \mathbf{z}_{1:T}^v) \\
 &= \prod_{t=1}^T p(\mathbf{x}_t, \mathbf{y}_t | \mathbf{z}^c, \mathbf{z}_t^w, \mathbf{z}_t^v) p(\mathbf{z}^c) p(\mathbf{z}_t^w | \mathbf{z}_{<t}^w) p(\mathbf{z}_t^v | \mathbf{z}_{<t}^v) \\
 &= \prod_{t=1}^T p(\mathbf{z}_t^w | \mathbf{z}_{<t}^w) p(\mathbf{z}_t^v | \mathbf{z}_{<t}^v) \prod_{i=1}^{N_t} p(z_i^c) p(x_i | z_i^c, \mathbf{z}_t^w) p(y_i | z_i^c, \mathbf{z}_t^v)
 \end{aligned} \tag{22}$$

where  $p(\mathbf{z}_1^w) = p(\mathbf{z}_1^w | \mathbf{z}_0^w)$  and  $p(\mathbf{z}_1^v) = p(\mathbf{z}_1^v | \mathbf{z}_0^v)$ .

For a non-stationary environment, we assume that  $(\mathbf{z}^c, \mathbf{z}_t^w)$  are the latent variables for  $\mathbf{x}_t$  and  $\mathbf{z}_t^v$  is the latent variable for  $\mathbf{y}_t$  at  $t$ -th time stamp. Thus the distribution of the latent variables for  $t$ -th time stamp can be express as  $p(\mathbf{z}^c | \mathbf{x}_t)$ ,  $p(\mathbf{z}_t^w | \mathbf{x}_t)$  and  $p(\mathbf{z}_t^v | \mathbf{y}_t)$ , respectively. We employ  $q(\mathbf{z}^c | \mathbf{x}_t)$ ,  $q(\mathbf{z}_t^w | \mathbf{z}_{<t}^w, \mathbf{x}_t)$  and  $q(\mathbf{z}_t^v | \mathbf{z}_{<t}^v, \mathbf{y}_t)$  to approximate the prior distributions here. Similar to Eq 18, we can draw the variational lower bound for  $\log p(\mathbf{x}_{1:T}, \mathbf{y}_{1:T})$  as:

$$\mathcal{L} = \mathbb{E}_q \log \frac{p(\mathbf{x}_{1:T}, \mathbf{y}_{1:T}, \mathbf{z}^c, \mathbf{z}_{1:T}^w, \mathbf{z}_{1:T}^v)}{q(\mathbf{z}^c, \mathbf{z}_{1:T}^w, \mathbf{z}_{1:T}^v | \mathbf{x}_{1:T}, \mathbf{y}_{1:T})}. \tag{23}$$

When incorporating with Eq 22, this can be reorganized as:

$$\begin{aligned}
 \mathcal{L} &= \mathbb{E}_q \log \frac{\prod_{t=1}^T p(\mathbf{x}_t, \mathbf{y}_t | \mathbf{z}^c, \mathbf{z}_t^w, \mathbf{z}_t^v) p(\mathbf{z}^c) p(\mathbf{z}_t^w | \mathbf{z}_{<t}^w) p(\mathbf{z}_t^v | \mathbf{z}_{<t}^v)}{\prod_{t=1}^T q(\mathbf{z}^c | \mathbf{x}_t) q(\mathbf{z}_t^w | \mathbf{z}_{<t}^w, \mathbf{x}_t) q(\mathbf{z}_t^v | \mathbf{z}_{<t}^v, \mathbf{y}_t)} \\
 &= \mathbb{E}_q \left[ \log \frac{\prod_{t=1}^T p(\mathbf{z}_t^w | \mathbf{z}_{<t}^w)}{\prod_{t=1}^T q(\mathbf{z}_t^w | \mathbf{z}_{<t}^w, \mathbf{x}_t)} + \log \frac{\prod_{t=1}^T p(\mathbf{z}_t^v | \mathbf{z}_{<t}^v)}{\prod_{t=1}^T q(\mathbf{z}_t^v | \mathbf{z}_{<t}^v, \mathbf{y}_t)} + \log \frac{\prod_{t=1}^T p(\mathbf{z}^c)}{\prod_{t=1}^T q(\mathbf{z}^c | \mathbf{x}_t)} \right. \\
 &\quad \left. + \log \prod_{t=1}^T p(\mathbf{x}_t | \mathbf{z}_t^c, \mathbf{z}_t^w) p(\mathbf{y}_t | \mathbf{z}_t^c, \mathbf{z}_t^v) \right] \\
 &= \mathbb{E}_q \left[ - \sum_{t=1}^T \log \frac{q(\mathbf{z}_t^w | \mathbf{z}_{<t}^w, \mathbf{x}_t)}{p(\mathbf{z}_t^w | \mathbf{z}_{<t}^w)} - \sum_{t=1}^T \log \frac{q(\mathbf{z}_t^v | \mathbf{z}_{<t}^v, \mathbf{y}_t)}{p(\mathbf{z}_t^v | \mathbf{z}_{<t}^v)} - \sum_{t=1}^T \log \frac{q(\mathbf{z}_t^c | \mathbf{x}_t)}{p(\mathbf{z}_t^c)} \right. \\
 &\quad \left. + \sum_{t=1}^T \log p(\mathbf{x}_t | \mathbf{z}^c, \mathbf{z}_t^w) p(\mathbf{y}_t | \mathbf{z}^c, \mathbf{z}_t^v) \right]
 \end{aligned} \tag{24}$$

By Jensen's inequality, the above formulation can be rewrote as:

$$\begin{aligned}
 \mathcal{L} &\geq \mathbb{E}_q \left[ \sum_{t=1}^T \log p(\mathbf{x}_t | \mathbf{z}^c, \mathbf{z}_t^w) p(\mathbf{y}_t | \mathbf{z}^c, \mathbf{z}_t^v) - \mathbb{D}_{KL}(q(\mathbf{z}^c | \mathbf{x}_t), p(\mathbf{z}^c)) - \mathbb{D}_{KL}(q(\mathbf{z}_t^w | \mathbf{z}_{<t}^w, \mathbf{x}_t), p(\mathbf{z}_t^w | \mathbf{z}_{<t}^w)) \right. \\
 &\quad \left. - \mathbb{D}_{KL}(q(\mathbf{z}_t^v | \mathbf{z}_{<t}^v, \mathbf{y}_t), p(\mathbf{z}_t^v | \mathbf{z}_{<t}^v)) \right]
 \end{aligned} \tag{25}$$

This complements the proof.

## B. Probabilistic Generative Model of LSSAE

In Fig. 6, we present the complete probabilistic generative graph of our LSSAE. The left part is the DAG of LSSAE presented in Fig. 1(d) where we did not include  $\mathbf{z}^c$  in this figure as the main focus of Fig. 1 is the dynamic factors (*i.e.*,  $W$  and  $V$ ) for presenting our main idea of evolving dynamics at a high level instead of DG. In the right part, we illustrate the relationship between our DAG and probabilistic generative model (with  $\mathbf{z}^c$ ). Here, the dependence of  $X_t$  on  $Y_t$  (the dotted blue line) is substituted with  $\mathbf{z}^c$  (the solid red line) which mainly captures the static category information in data sample space.

## C. Additional Details on the Experimental Setup

### C.1. Datasets

Our experiments are conducted on 2 synthetic and 4 real-world datasets presented in Table 4. More detailed description are given below.

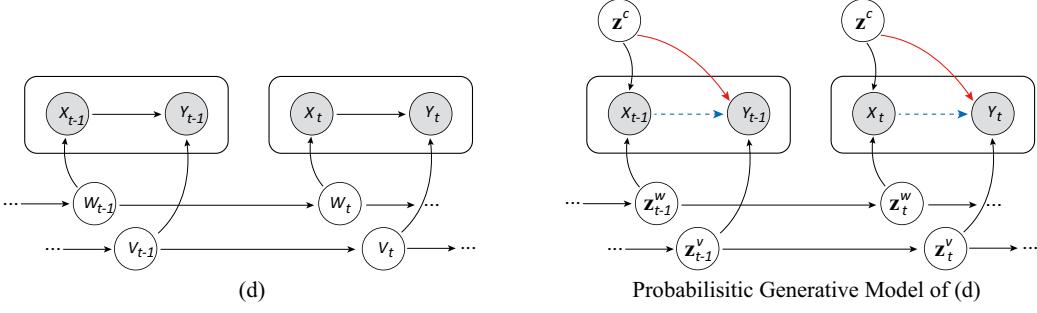


Figure 6. The probabilistic generative graph of LSSAE with  $\mathbf{z}^c$  presented. Here, the dependence of  $X_t$  on  $Y_t$  in Fig. 1(d) is substituted with  $\mathbf{z}^c$  which mainly captures the static category information in data sample space.

Table 4. Brief description of employed benchmarks in this work.

Dataset	Type	Number	Source Domains	Intermediate Domains	Target Domains	Total Domains
Circle-/C	Digital	3,000	15	5	10	30
Sine-/C	Digital	2,280	12	4	8	24
RMNIST	Image	70,000	10	3	6	19
Portraits	Image	37,921	19	5	10	34
CalTran	Image	5,450	19	5	10	34
PowerSupply	Digital	29,928	15	5	10	30

- **Circle** (Pesaranghader & Viktor, 2016): Each data in this dataset owns two attributes  $(x, y)$ ,  $x, y \in [0, 1]$ . The label is assigned using a circle curve as the decision boundary following  $(x - x_0)^2 + (y - y_0)^2 \leq r^2$ , where  $(x_0, y_0)$  are the location of the center and  $r$  is the radius of this circle. To generate Circle-C, we inject gradual shift via modifying the value of  $x_0$  continuously throughout the domains.
- **Sine** (Pesaranghader & Viktor, 2016): Each data in this dataset owns two attributes  $(x, y)$ ,  $x, y \in [0, 1]$ . The label is assigned using a sine curve as the decision boundary following  $y \leq \sin(x)$ . To simulate abrupt concept drift for Sine-C, labels are reversed (*i.e.*, from 0 to 1 or from 1 to 0) from the 6-th domain to the last one.
- **RMNIST** (Ghifary et al., 2015): This dataset is composed of MNIST digits of various rotations. We generate 19 domains via applying the rotations with degree of  $\mathcal{R} = \{0^\circ, 15^\circ, 30^\circ, \dots, 180^\circ\}$  on each domain. Note that each image is seen at exactly one angle, so the training procedure cannot track a single image across different angles.
- **Portraits** (Ginosar et al., 2015): This is a real-world dataset of photos collected in American high school seniors. The portraits are taken over 108 years (1905-2013) across 26 states. The goal is to classify the gender for each photo. We split the dataset into 34 domains by a fixed internal along time.
- **CalTran** (Hoffman et al., 2014): This dataset contains real-world images captured by a fixed traffic camera deployed in an intersection over time. Frames were updated at 3 minute intervals each with a resolution  $320 \times 320$ . We divide it into 34 domains by time. This is a scene classification task to determine whether one or more cars are present in, or approaching the intersection. The challenge mainly raise from the continually evolving domain shift as changes include time, illumination, weather, *etc.*
- **PowerSupply** (Dau et al., 2019): This dataset is comprised of records of hourly power supply collected by an Italy electricity company. We form 30 domains according to days. Each data is assigned by a binary class label which represents which time of day the current power supply belongs to (*i.e.*, am or pm). The concept shift may results from the change in season, weather, price or the differences between working days and weekend.

## C.2. Model Architecture & Hyperparameters

Neural network architectures used for each dataset:

Table 5. Model architectures for different datasets.

Dataset	Encoder	Decoder
Circle/-C		
Sine/-C	Non-linear Encoder	Non-linear Decoder
PowerSupply		
RMNIST	MNIST ConvNet	MNIST ConvTranNet
Portraits	ResNet-18	ConvTranNet
CalTran		

Neural network architecture for digital experiments (Circle/-C, Sine/-C, PowerSupply):

Table 6. Implementation of Non-linear Encoder.

#	Layer
1	Linear(in= $d$ , output=512)
2	ReLU
3	Linear(in=512, output=512)
4	ReLU
5	Linear(in=512, output=512)
6	ReLU
7	Linear(in=512, output=512)

Table 7. Implementation of Non-linear Decoder.

#	Layer
1	Linear(in= $d$ , output=16)
2	BatchNorm
3	LeakyReLU(0.2)
4	Linear(in=16, output=64)
5	BatchNorm
6	LeakyReLU(0.2)
7	Linear(in=64, output=128)
8.	BatchNorm
9	LeakyReLU(0.2)
10	Linear(in=128, output= $d$ )

Neural network architecture for RMNIST experiments:

Table 8. Implementation of MNIST ConvNet.

#	Layer
1	Conv2D(in= $d$ , output=64)
2	ReLU
3	GroupNorm(groupds=8)
4	Conv2D(in=64, output=128, stride=2)
5	ReLU
6	GroupNorm(groupds=8)
7	Conv2D(in=128, output=128)
8	ReLU
9	GroupNorm(groupds=8)
10	Conv2D(in=128, output=128)
11	ReLU
12	GroupNorm(groupds=8)
13	Global average-pooling

Table 9. Implementation of MNIST ConvTranNet.

#	Layer
1	Linear(in= $d$ , output=1024)
2	BatchNorm
3	ReLU
4	Upsample(8)
5	ConvTransposed2D(in=64, output=128, kernel=5)
6	BatchNorm
7	ReLU
8	Upsample(24)
9	ConvTransposed2D(in=128, output=256, kernel=5)
10	BatchNorm
11	ReLU
12	Conv2D(in=256, output=1, kernel=1)
13	Sigmoid

Neural network architecture for Portraits and CalTran experiments:

*Table 10.* Implementation of ConvTranNet.

#	Layer
1	Linear( $\text{in}=d$ , $\text{output}=1024$ )
2	BatchNorm
3	ReLU
4	Upsample(16)
5	ConvTransposed2D( $\text{in}=64$ , $\text{output}=128$ , $\text{kernel}=5$ )
6	BatchNorm
7	ReLU
8	Upsample(40)
9	ConvTransposed2D( $\text{in}=128$ , $\text{output}=256$ , $\text{kernel}=5$ )
10	BatchNorm
11	ReLU
12	Upsample(80)
13	ConvTransposed2D( $\text{in}=256$ , $\text{output}=3$ , $\text{kernel}=5$ )
14	BatchNorm
15	ReLU
16	Sigmoid

The model architecture of the encoder is ResNet-18, we replace the final softmax layer of the official version following [Gulrajani & Lopez-Paz \(2021\)](#). Besides, a dropout layer before the final linear layer is inserted. This network is initialized by random rather than loading the pretrained parameters on ImageNet.

We list the values of hyperparameters for different datasets below. All models are optimized by Adam (Kingma & Ba, 2015). In our experiments, we found that keeping the balance of the three KL divergence terms for  $\mathbf{z}^c$ ,  $\mathbf{z}^w$  and  $\mathbf{z}^v$  via adjusting the value of  $\lambda_1$ ,  $\lambda_2$  and  $\lambda_3$  is beneficial for the final results.

Table 11. Hyperparametes and their default values.

Dataset	Parameters	Value
Circle	learning rate for $E^c, D, C$	5e-5
	learning rate for $E^w, F^w, E^v, F^v$	5e-6
	batch size	24
	$\lambda_1, \lambda_2, \lambda_3$	1.0, 1.0, 1.0
	$\alpha$	0.05
Circle-C	learning rate for $E^c, D, C$	1e-5
	learning rate for $E^w, F^w, E^v, F^v$	1e-6
	batch size	24
	$\lambda_1, \lambda_2, \lambda_3$	1.0, 2.0, 1.0
	$\alpha$	0.05
Sine	learning rate for $E^c, D, C$	5e-5
	learning rate for $E^w, F^w, E^v, F^v$	5e-6
	batch size	24
	$\lambda_1, \lambda_2, \lambda_3$	2.0, 1.0, 1.0
	$\alpha$	0.05
Sine-C	learning rate for $E^c, D, C$	1e-5
	learning rate for $E^w, F^w, E^v, F^v$	1e-6
	batch size	24
	$\lambda_1, \lambda_2, \lambda_3$	2.0, 1.0, 1.0
	$\alpha$	0.05
RMNIST	learning rate for $E^c, D, C$	1e-3
	learning rate for $E^w, F^w, E^v, F^v$	1e-4
	batch size	48
	$\lambda_1, \lambda_2, \lambda_3$	2.0, 1.0, 1.0
	$\alpha$	0.05
Portraits	learning rate for $E^c, D, C$	1e-5
	learning rate for $E^w, F^w, E^v, F^v$	1e-6
	batch size	24
	$\lambda_1, \lambda_2, \lambda_3$	0.5, 1.0, 1.0
	$\alpha$	0.05
CalTran	learning rate for $E^c, D, C$	5e-5
	learning rate for $E^w, F^w, E^v, F^v$	5e-6
	batch size	24
	$\lambda_1, \lambda_2, \lambda_3$	1.0, 1.0, 1.0
	$\alpha$	0.1
PowerSupply	learning rate for $E^c, D, C$	1e-5
	learning rate for $E^w, F^w, E^v, F^v$	1e-6
	batch size	48
	$\lambda_1, \lambda_2, \lambda_3$	2.0, 1.0, 1.0
	$\alpha$	0.1

## D. Additional Experimental Results

In this section, we provide more experimental results for our proposed evolving domain generalization task. As we can see, in most of the cases, we can achieve the state-of-the-art performance compared with other domain generalization baselines. However, in some cases, our proposed method cannot achieve desired results. For example, in Sine-C, ERM and some domain generalization baselines can achieve a desired classification performance especially for domain index 17, 18 and 19, which we conjecture the reason that their decision boundaries overfit to the abrupt concept shift, which further leads to their poor performance on the following domains after domain 19. However, our proposed method aims to learn the evolving pattern starting from  $t = 0$ , which may not be optimal to suit to this abrupt concept shift case. One possible future direction is only to learn the evolving pattern in a certain time duration (i.e., time duration in  $[t - T_0, t]$ , where  $t$  is the current time stamp). Nevertheless, our proposed method shows its effectiveness in dynamic modeling for evolving domain generalization task. We will leave the discussion how to find a suitable time duration (i.e., a suitable  $T_0$ ) in our future work.

Table 12. Circle. We show the results on each target domain by domain index.

Algorithm	21	22	23	24	25	26	27	28	29	30	Avg
ERM	53.9 $\pm$ 3.5	55.8 $\pm$ 4.8	53.9 $\pm$ 5.2	44.7 $\pm$ 6.3	56.9 $\pm$ 4.4	47.8 $\pm$ 5.8	41.9 $\pm$ 7.5	41.7 $\pm$ 5.9	54.2 $\pm$ 3.0	47.8 $\pm$ 5.7	49.9
Mixup	48.6 $\pm$ 3.8	51.7 $\pm$ 4.0	49.4 $\pm$ 4.5	43.6 $\pm$ 5.8	56.9 $\pm$ 4.4	47.8 $\pm$ 5.8	41.9 $\pm$ 7.5	41.7 $\pm$ 5.9	54.2 $\pm$ 3.0	47.8 $\pm$ 5.7	48.4
MMD	50.0 $\pm$ 3.9	53.6 $\pm$ 4.4	55.0 $\pm$ 4.3	51.9 $\pm$ 6.0	60.8 $\pm$ 3.9	49.7 $\pm$ 7.2	41.9 $\pm$ 7.5	41.7 $\pm$ 5.9	54.2 $\pm$ 3.0	47.8 $\pm$ 5.7	50.7
MLDG	57.8 $\pm$ 3.6	57.7 $\pm$ 5.0	55.3 $\pm$ 4.9	46.4 $\pm$ 6.8	56.9 $\pm$ 4.4	47.8 $\pm$ 5.8	41.9 $\pm$ 7.5	41.7 $\pm$ 5.9	54.2 $\pm$ 3.0	47.8 $\pm$ 5.7	50.8
IRM	57.8 $\pm$ 3.9	59.4 $\pm$ 5.4	56.9 $\pm$ 4.9	48.1 $\pm$ 7.4	57.5 $\pm$ 4.3	47.8 $\pm$ 5.8	41.9 $\pm$ 7.5	41.7 $\pm$ 5.9	54.2 $\pm$ 3.0	47.8 $\pm$ 5.7	51.3
RSC	45.3 $\pm$ 3.6	51.4 $\pm$ 3.8	49.4 $\pm$ 4.5	43.6 $\pm$ 5.8	56.9 $\pm$ 4.4	47.8 $\pm$ 5.8	41.9 $\pm$ 7.5	41.7 $\pm$ 5.9	54.2 $\pm$ 3.0	47.8 $\pm$ 5.7	48.0
MTL	61.4 $\pm$ 2.2	57.2 $\pm$ 6.4	53.3 $\pm$ 5.1	48.3 $\pm$ 6.2	56.9 $\pm$ 4.8	49.2 $\pm$ 3.7	43.3 $\pm$ 5.0	45.8 $\pm$ 2.8	54.2 $\pm$ 5.7	42.2 $\pm$ 4.9	51.2
Fish	51.7 $\pm$ 3.7	53.1 $\pm$ 3.7	49.4 $\pm$ 4.5	43.6 $\pm$ 5.8	56.9 $\pm$ 4.4	47.8 $\pm$ 5.8	41.9 $\pm$ 7.5	41.7 $\pm$ 5.9	54.2 $\pm$ 3.0	47.8 $\pm$ 5.7	48.8
CORAL	65.3 $\pm$ 3.2	63.9 $\pm$ 4.4	60.0 $\pm$ 4.8	56.4 $\pm$ 6.0	60.2 $\pm$ 4.3	47.8 $\pm$ 5.8	41.9 $\pm$ 7.5	41.7 $\pm$ 5.9	54.2 $\pm$ 3.0	47.8 $\pm$ 5.7	53.9
AndMask	42.8 $\pm$ 3.4	50.6 $\pm$ 4.1	49.4 $\pm$ 4.5	43.6 $\pm$ 5.8	56.9 $\pm$ 4.4	47.8 $\pm$ 5.8	41.9 $\pm$ 7.5	41.7 $\pm$ 5.9	54.2 $\pm$ 3.0	49.7 $\pm$ 5.4	47.9
DIVA	81.3 $\pm$ 3.5	76.3 $\pm$ 4.2	74.7 $\pm$ 4.6	56.7 $\pm$ 5.1	67.0 $\pm$ 6.1	62.3 $\pm$ 5.1	62.0 $\pm$ 5.6	66.3 $\pm$ 4.1	70.3 $\pm$ 5.6	62.0 $\pm$ 4.2	67.9
LSSAE (Ours)	95.8 $\pm$ 1.9	95.6 $\pm$ 2.1	93.5 $\pm$ 2.9	96.3 $\pm$ 1.8	83.8 $\pm$ 5.2	74.3 $\pm$ 3.6	51.9 $\pm$ 5.6	52.3 $\pm$ 8.1	46.5 $\pm$ 9.2	48.4 $\pm$ 5.3	73.8

Table 13. Sine. We show the results on each target domain denoted by domain index.

Algorithm	17	18	19	20	21	22	23	24	Avg
ERM	71.4 $\pm$ 6.1	91.0 $\pm$ 1.5	81.6 $\pm$ 2.4	53.4 $\pm$ 2.9	51.1 $\pm$ 6.7	54.3 $\pm$ 4.7	49.5 $\pm$ 4.8	51.7 $\pm$ 5.0	63.0
Mixup	63.1 $\pm$ 5.9	93.5 $\pm$ 1.7	80.6 $\pm$ 3.8	52.8 $\pm$ 2.9	60.3 $\pm$ 7.2	54.2 $\pm$ 2.7	49.5 $\pm$ 4.4	49.3 $\pm$ 8.0	62.9
MMD	57.0 $\pm$ 4.2	57.1 $\pm$ 4.1	47.6 $\pm$ 5.4	50.0 $\pm$ 1.8	55.1 $\pm$ 6.7	54.4 $\pm$ 4.7	49.5 $\pm$ 4.8	51.7 $\pm$ 5.0	55.8
MLDG	69.2 $\pm$ 4.2	67.7 $\pm$ 4.1	52.1 $\pm$ 5.4	50.7 $\pm$ 1.8	51.1 $\pm$ 6.7	54.3 $\pm$ 4.7	49.5 $\pm$ 4.8	51.7 $\pm$ 5.0	63.2
IRM	66.9 $\pm$ 6.2	81.1 $\pm$ 3.2	88.5 $\pm$ 3.0	56.6 $\pm$ 6.0	57.2 $\pm$ 5.8	53.7 $\pm$ 5.1	49.5 $\pm$ 2.2	51.7 $\pm$ 5.4	63.2
RSC	61.3 $\pm$ 6.6	83.5 $\pm$ 1.9	84.5 $\pm$ 2.6	52.8 $\pm$ 2.8	55.1 $\pm$ 6.7	54.4 $\pm$ 4.7	49.5 $\pm$ 4.8	51.7 $\pm$ 5.0	61.5
MTL	70.6 $\pm$ 6.6	91.6 $\pm$ 1.2	79.9 $\pm$ 3.4	51.0 $\pm$ 4.7	60.3 $\pm$ 7.6	53.6 $\pm$ 5.2	49.5 $\pm$ 5.3	46.9 $\pm$ 5.9	62.9
Fish	66.1 $\pm$ 6.9	82.0 $\pm$ 2.7	87.5 $\pm$ 2.4	55.2 $\pm$ 3.0	51.1 $\pm$ 6.7	54.3 $\pm$ 4.7	49.5 $\pm$ 4.8	51.7 $\pm$ 5.0	62.3
CORAL	60.0 $\pm$ 5.3	57.1 $\pm$ 4.2	48.6 $\pm$ 6.4	50.7 $\pm$ 1.8	49.7 $\pm$ 6.2	48.6 $\pm$ 4.6	46.3 $\pm$ 5.0	51.7 $\pm$ 5.0	51.6
AndMask	44.2 $\pm$ 5.1	42.9 $\pm$ 4.2	54.2 $\pm$ 7.0	71.9 $\pm$ 1.9	86.4 $\pm$ 3.2	90.4 $\pm$ 2.9	88.1 $\pm$ 3.4	76.4 $\pm$ 3.7	69.3
DIVA	79.0 $\pm$ 6.6	60.8 $\pm$ 1.9	47.6 $\pm$ 2.6	50.0 $\pm$ 2.8	55.1 $\pm$ 6.7	51.9 $\pm$ 4.7	38.6 $\pm$ 4.8	40.4 $\pm$ 5.0	52.9
LSSAE (Ours)	93.0 $\pm$ 1.7	86.9 $\pm$ 0.7	69.2 $\pm$ 1.5	63.8 $\pm$ 3.8	68.8 $\pm$ 2.5	76.8 $\pm$ 4.8	63.9 $\pm$ 1.3	49.0 $\pm$ 3.1	71.4

Table 14. RMNIST. We show the results on each target domain denoted by rotation angle.

Algorithm	130°	140°	150°	160°	170°	180°	Avg
ERM	56.8 ± 0.9	44.2 ± 0.8	37.8 ± 0.6	38.3 ± 0.8	40.9 ± 0.8	43.6 ± 0.8	43.6
Mixup	61.3 ± 0.7	47.4 ± 0.8	39.1 ± 0.7	38.3 ± 0.7	40.5 ± 0.8	42.8 ± 0.9	44.9
MMD	59.2 ± 0.9	46.0 ± 0.8	39.0 ± 0.7	39.3 ± 0.8	41.6 ± 0.7	43.7 ± 0.8	44.8
MLDG	57.4 ± 0.7	44.5 ± 0.9	37.5 ± 0.8	37.5 ± 0.8	39.9 ± 0.8	42.0 ± 0.9	43.1
IRM	47.7 ± 0.9	38.5 ± 0.7	34.1 ± 0.7	35.7 ± 0.8	37.8 ± 0.8	40.3 ± 0.8	39.0
RSC	54.1 ± 0.9	41.9 ± 0.8	35.8 ± 0.7	37.0 ± 0.8	39.8 ± 0.8	41.6 ± 0.8	41.7
MTL	54.8 ± 0.9	43.1 ± 0.8	36.4 ± 0.8	36.1 ± 0.8	39.1 ± 0.9	40.9 ± 0.8	41.7
Fish	60.8 ± 0.8	47.8 ± 0.8	39.2 ± 0.8	37.6 ± 0.7	39.0 ± 0.8	40.7 ± 0.7	44.2
CORAL	58.8 ± 0.9	46.2 ± 0.8	38.9 ± 0.7	38.5 ± 0.8	41.3 ± 0.8	43.5 ± 0.8	44.5
AndMask	53.5 ± 0.9	42.9 ± 0.8	37.8 ± 0.7	38.6 ± 0.8	40.8 ± 0.8	43.2 ± 0.8	42.8
DIVA	58.3 ± 0.8	45.0 ± 0.8	37.6 ± 0.8	36.9 ± 0.7	38.1 ± 0.8	40.1 ± 0.8	42.7
LSSAE (Ours)	64.1 ± 0.8	51.6 ± 0.8	43.4 ± 0.8	38.6 ± 0.7	40.3 ± 0.8	40.4 ± 0.8	46.4

Table 15. Portraits. We show the results on each target domain denoted by domain index.

Algorithm	25	26	27	28	29	30	31	32	33	34	Avg
ERM	75.5 ± 0.9	83.8 ± 0.9	88.5 ± 0.8	93.3 ± 0.7	93.4 ± 0.6	92.1 ± 0.7	90.6 ± 0.8	84.3 ± 0.9	88.5 ± 0.9	87.9 ± 1.4	87.8
Mixup	75.5 ± 0.9	83.8 ± 0.9	88.5 ± 0.8	93.3 ± 0.7	93.4 ± 0.6	92.1 ± 0.7	90.6 ± 0.8	84.3 ± 0.9	88.5 ± 0.9	87.9 ± 1.4	87.8
MMD	74.0 ± 1.0	83.8 ± 0.8	87.2 ± 0.8	93.0 ± 0.7	93.0 ± 0.6	91.9 ± 0.7	90.9 ± 0.7	84.7 ± 1.4	88.3 ± 0.9	85.8 ± 1.8	87.3
MLDG	76.4 ± 0.8	85.5 ± 0.9	90.1 ± 0.7	94.3 ± 0.6	93.5 ± 0.6	92.0 ± 0.7	90.8 ± 0.8	85.6 ± 1.1	89.3 ± 0.8	87.6 ± 1.6	88.5
IRM	74.2 ± 0.9	83.5 ± 0.9	88.5 ± 0.8	91.0 ± 0.8	90.4 ± 0.7	87.3 ± 0.8	87.0 ± 0.9	80.4 ± 1.5	86.7 ± 0.9	85.1 ± 1.8	85.4
RSC	75.2 ± 0.9	84.7 ± 0.8	87.9 ± 0.7	93.3 ± 0.7	92.5 ± 0.7	91.0 ± 0.7	90.0 ± 0.7	84.6 ± 1.2	88.2 ± 0.8	85.8 ± 1.9	87.3
MTL	78.2 ± 0.9	86.5 ± 0.8	90.9 ± 0.8	94.2 ± 0.7	93.8 ± 0.6	92.0 ± 0.7	91.2 ± 0.7	86.0 ± 1.2	89.3 ± 0.8	87.4 ± 1.4	89.0
Fish	78.6 ± 0.9	86.9 ± 0.8	89.5 ± 0.8	93.5 ± 0.7	93.3 ± 0.6	92.1 ± 0.6	91.1 ± 0.7	86.2 ± 1.3	88.7 ± 0.9	87.7 ± 1.6	88.8
CORAL	74.6 ± 0.9	84.6 ± 0.8	87.9 ± 0.8	93.3 ± 0.6	92.7 ± 0.7	91.5 ± 0.7	90.7 ± 0.7	84.6 ± 1.5	88.1 ± 0.9	85.9 ± 1.9	87.4
AndMask	62.0 ± 1.1	70.8 ± 1.1	67.0 ± 1.2	70.2 ± 1.1	75.2 ± 1.1	74.1 ± 1.0	72.7 ± 1.1	64.7 ± 1.6	77.3 ± 1.1	74.9 ± 2.1	70.9
DIVA	76.2 ± 1.0	86.6 ± 0.8	88.8 ± 0.8	93.5 ± 0.7	93.1 ± 0.6	91.6 ± 0.6	91.1 ± 0.7	84.7 ± 1.3	89.1 ± 0.8	87.0 ± 1.5	88.2
LSSAE (Ours)	77.7 ± 0.9	87.1 ± 0.8	90.8 ± 0.7	94.3 ± 0.6	94.3 ± 0.6	92.2 ± 0.6	91.2 ± 0.7	86.7 ± 1.1	89.6 ± 0.8	86.9 ± 1.4	89.1

Table 16. CalTran. We show the results on each target domain denoted by domain index.

Algorithm	25	26	27	28	29	30	31	32	33	34	Avg
ERM	29.9 ± 3.5	88.4 ± 2.1	61.1 ± 3.5	56.3 ± 3.2	90.0 ± 1.6	60.1 ± 2.5	55.5 ± 3.5	88.8 ± 2.4	57.1 ± 3.5	50.5 ± 5.2	66.3
Mixup	53.6 ± 3.9	89.0 ± 2.0	61.8 ± 2.4	55.7 ± 2.9	88.2 ± 2.1	58.6 ± 3.0	52.3 ± 3.7	88.6 ± 2.7	57.1 ± 3.0	55.1 ± 4.3	66.0
MMD	30.2 ± 2.1	92.7 ± 1.7	56.4 ± 3.7	39.1 ± 3.2	93.6 ± 1.7	52.1 ± 3.2	42.8 ± 3.0	92.1 ± 2.2	42.1 ± 3.8	29.4 ± 3.8	57.1
MLDG	54.8 ± 4.1	88.6 ± 2.6	62.2 ± 3.6	55.1 ± 4.1	88.3 ± 1.7	60.9 ± 4.3	51.7 ± 2.6	89.0 ± 1.9	56.5 ± 3.4	55.3 ± 4.8	66.2
IRM	46.4 ± 3.7	90.8 ± 1.7	60.8 ± 3.4	52.9 ± 3.1	91.8 ± 1.7	56.6 ± 3.1	52.1 ± 2.9	90.9 ± 2.6	55.6 ± 3.9	43.1 ± 5.5	64.1
RSC	57.2 ± 3.0	88.4 ± 2.6	62.6 ± 3.0	56.5 ± 3.7	88.0 ± 2.4	59.4 ± 3.0	51.9 ± 2.9	90.0 ± 2.0	59.4 ± 2.9	56.0 ± 3.1	67.0
MTL	64.2 ± 3.0	87.2 ± 2.5	64.9 ± 3.9	60.0 ± 4.8	84.5 ± 2.2	60.6 ± 3.5	52.6 ± 3.7	83.9 ± 2.9	58.2 ± 4.1	65.7 ± 5.6	68.2
Fish	61.1 ± 3.5	88.2 ± 1.5	64.7 ± 4.0	57.9 ± 3.1	88.3 ± 2.2	59.9 ± 3.0	57.5 ± 2.7	87.4 ± 2.8	57.7 ± 3.7	63.0 ± 6.1	68.6
CORAL	50.4 ± 3.0	90.8 ± 2.0	61.2 ± 3.8	55.0 ± 2.5	92.0 ± 1.7	56.8 ± 3.8	52.0 ± 3.8	90.9 ± 1.6	56.8 ± 2.4	50.9 ± 5.6	65.7
AndMask	30.0 ± 2.2	92.7 ± 1.7	56.2 ± 3.8	39.1 ± 3.2	93.6 ± 1.7	51.6 ± 3.2	42.6 ± 2.9	92.1 ± 2.2	41.2 ± 3.7	29.9 ± 3.6	56.9
DIVA	60.6 ± 2.9	90.1 ± 1.7	67.5 ± 3.1	58.9 ± 3.5	88.4 ± 2.8	58.7 ± 3.3	53.8 ± 3.6	89.8 ± 1.7	61.8 ± 4.8	62.0 ± 3.4	69.2
LSSVAE	63.4 ± 3.4	92.1 ± 2.0	62.6 ± 4.7	58.8 ± 4.4	92.9 ± 1.6	62.0 ± 3.9	54.3 ± 3.0	92.1 ± 2.2	60.5 ± 3.8	67.4 ± 3.6	70.6

Table 17. Circle-C. We show the results on each target domain denoted by domain index.

Algorithm	21	22	23	24	25	26	27	28	29	30	Avg
ERM	40.6 ± 3.8	43.1 ± 4.7	39.7 ± 4.0	33.6 ± 5.0	44.7 ± 6.0	31.4 ± 6.4	27.8 ± 7.7	26.9 ± 6.0	27.8 ± 3.5	33.1 ± 7.1	34.9
Mixup	40.6 ± 3.8	42.5 ± 4.7	39.7 ± 4.0	33.6 ± 5.0	44.7 ± 6.0	31.4 ± 6.4	27.8 ± 7.7	26.9 ± 6.0	27.8 ± 3.5	33.1 ± 7.1	34.8
MMD	38.6 ± 3.7	42.5 ± 4.7	39.7 ± 4.0	33.6 ± 5.0	44.7 ± 6.0	31.4 ± 6.4	27.8 ± 7.7	26.9 ± 6.0	27.8 ± 3.5	33.1 ± 7.1	34.6
MLDG	44.4 ± 4.4	43.9 ± 5.0	39.7 ± 4.0	33.6 ± 5.0	44.7 ± 6.0	31.4 ± 6.4	27.8 ± 7.7	26.9 ± 6.0	27.8 ± 3.5	33.1 ± 7.1	35.3
IRM	63.6 ± 4.2	56.7 ± 7.0	48.1 ± 4.3	37.2 ± 4.8	44.7 ± 6.0	31.4 ± 6.4	27.8 ± 7.7	26.9 ± 6.0	27.8 ± 3.5	33.1 ± 7.1	39.7
RSC	38.6 ± 3.7	42.5 ± 4.7	39.7 ± 4.0	33.6 ± 5.0	44.7 ± 6.0	31.4 ± 6.4	27.8 ± 7.7	26.9 ± 6.0	27.8 ± 3.5	33.1 ± 7.1	34.6
MTL	41.7 ± 5.0	45.6 ± 7.1	36.9 ± 6.3	36.4 ± 6.2	44.7 ± 6.0	31.4 ± 3.2	27.8 ± 4.2	28.3 ± 2.8	31.9 ± 4.6	26.1 ± 4.1	35.1
Fish	42.5 ± 3.8	43.3 ± 4.8	39.7 ± 4.0	33.6 ± 5.0	44.7 ± 6.0	31.4 ± 6.4	27.8 ± 7.7	26.9 ± 6.0	27.8 ± 3.5	33.1 ± 7.1	35.1
CORAL	40.8 ± 3.7	43.3 ± 4.8	39.7 ± 4.0	33.6 ± 5.0	44.7 ± 6.0	31.4 ± 6.4	27.8 ± 7.7	26.9 ± 6.0	27.8 ± 3.5	33.1 ± 7.1	34.9
AndMask	53.6 ± 4.2	54.7 ± 5.4	51.7 ± 4.7	33.6 ± 5.1	44.7 ± 6.0	31.4 ± 6.4	27.8 ± 7.7	26.9 ± 6.0	27.8 ± 3.5	35.0 ± 6.7	38.9
DIVA	40.6 ± 3.8	42.5 ± 4.7	39.7 ± 4.0	33.6 ± 5.0	44.7 ± 6.0	31.4 ± 6.4	27.8 ± 7.7	26.9 ± 6.0	27.8 ± 3.5	33.1 ± 7.1	34.8
LSSAE (Ours)	74.5 ± 1.8	65.5 ± 3.9	55.5 ± 3.2	36.0 ± 3.5	45.0 ± 0.7	40.5 ± 4.6	35.0 ± 1.4	33.0 ± 1.4	34.0 ± 6.4	29.0 ± 2.8	44.8

Table 18. Sine-C. We show the results on each target domain denoted by domain index.

Algorithm	17	18	19	20	21	22	23	24	Avg
ERM	64.2 ± 6.8	84.9 ± 3.2	83.7 ± 3.1	54.9 ± 2.1	51.1 ± 6.7	54.3 ± 4.7	49.5 ± 4.8	51.7 ± 5.0	61.8
Mixup	60.3 ± 7.8	77.4 ± 3.3	87.8 ± 1.8	57.3 ± 2.6	51.1 ± 6.7	54.3 ± 4.7	49.5 ± 4.8	51.7 ± 5.0	61.2
MMD	55.8 ± 5.1	57.1 ± 4.2	48.6 ± 6.4	50.7 ± 1.8	51.1 ± 6.7	54.3 ± 4.7	49.5 ± 4.8	51.7 ± 5.0	52.4
MLDG	65.6 ± 6.4	88.7 ± 2.3	84.4 ± 2.8	52.4 ± 2.6	51.1 ± 6.7	54.3 ± 4.7	49.5 ± 4.8	51.7 ± 5.0	62.2
IRM	61.4 ± 7.0	82.8 ± 3.2	86.5 ± 3.0	54.9 ± 1.8	55.1 ± 6.7	54.3 ± 4.7	49.5 ± 4.8	51.7 ± 5.0	61.5
RSC	65.0 ± 6.7	83.5 ± 3.1	85.4 ± 3.1	53.8 ± 2.6	51.1 ± 6.7	54.3 ± 4.7	49.5 ± 4.8	51.7 ± 5.0	61.8
MTL	61.9 ± 7.3	82.0 ± 2.3	83.7 ± 4.2	54.2 ± 5.4	60.3 ± 7.6	53.6 ± 5.2	49.5 ± 5.3	46.9 ± 5.9	61.5
Fish	69.4 ± 6.2	94.5 ± 1.5	79.5 ± 3.1	52.4 ± 2.6	51.1 ± 6.7	54.3 ± 4.7	49.5 ± 4.8	51.7 ± 5.0	62.8
CORAL	70.3 ± 5.0	77.2 ± 3.6	67.0 ± 4.2	52.1 ± 2.5	51.1 ± 6.7	54.3 ± 4.7	49.5 ± 4.8	51.7 ± 5.0	59.2
AndMask	55.8 ± 5.1	57.1 ± 4.2	48.6 ± 6.4	50.7 ± 1.8	51.1 ± 6.7	54.3 ± 4.7	49.5 ± 4.8	51.7 ± 5.0	52.3
DIVA	76.9 ± 6.1	61.1 ± 5.9	47.6 ± 3.4	48.6 ± 4.2	51.1 ± 7.0	52.9 ± 6.1	38.5 ± 5.0	36.5 ± 6.8	51.7
LSSAE (Ours)	62.3 ± 3.1	63.2 ± 6.7	57.6 ± 2.6	66.4 ± 2.6	63.5 ± 3.9	59.5 ± 4.0	52.6 ± 2.2	61.3 ± 1.9	60.8

Table 19. PowerSupply. We show the results on each target domain denoted by domain index.

Algorithm	21	22	23	24	25	26	27	28	29	30	Avg
ERM	69.8 ± 1.4	70.0 ± 1.4	69.2 ± 1.3	64.4 ± 1.5	85.8 ± 1.0	76.0 ± 1.3	70.1 ± 1.5	69.8 ± 1.5	69.0 ± 1.3	65.5 ± 1.5	71.0
Mixup	69.6 ± 1.4	69.5 ± 1.5	68.3 ± 1.5	64.3 ± 1.5	87.1 ± 1.0	76.6 ± 1.3	70.1 ± 1.4	69.2 ± 1.3	68.1 ± 1.5	65.0 ± 1.6	70.8
MMD	70.0 ± 1.3	69.7 ± 1.4	68.7 ± 1.4	64.8 ± 1.5	85.6 ± 1.0	76.1 ± 1.3	70.0 ± 1.5	69.5 ± 1.4	68.7 ± 1.3	65.6 ± 1.5	70.9
MLDG	69.7 ± 1.4	69.7 ± 1.5	68.6 ± 1.5	64.6 ± 1.5	86.4 ± 1.1	76.3 ± 1.4	70.1 ± 1.4	69.4 ± 1.3	68.4 ± 1.5	65.6 ± 1.5	70.8
IRM	69.8 ± 1.4	69.5 ± 1.4	68.3 ± 1.4	64.1 ± 1.4	87.2 ± 0.9	76.5 ± 1.3	70.0 ± 1.5	69.1 ± 1.5	68.2 ± 1.3	65.0 ± 1.4	70.8
RSC	69.9 ± 1.4	69.6 ± 1.4	68.6 ± 1.4	64.4 ± 1.5	86.6 ± 1.0	76.3 ± 1.3	70.0 ± 1.5	69.4 ± 1.4	68.4 ± 1.3	65.4 ± 1.5	70.9
MTL	69.6 ± 1.4	69.4 ± 1.5	68.2 ± 1.6	64.2 ± 1.5	87.4 ± 1.2	76.6 ± 1.3	69.9 ± 1.5	69.1 ± 1.5	68.2 ± 1.5	64.6 ± 1.4	70.7
Fish	69.7 ± 1.4	69.4 ± 1.4	68.2 ± 1.4	64.2 ± 1.4	87.3 ± 1.0	76.6 ± 1.3	69.9 ± 1.5	69.2 ± 1.5	68.2 ± 1.3	65.2 ± 1.5	70.8
CORAL	69.9 ± 1.4	69.7 ± 1.4	68.9 ± 1.4	64.6 ± 1.4	86.1 ± 1.0	76.3 ± 1.3	70.0 ± 1.5	69.5 ± 1.5	68.8 ± 1.3	65.7 ± 1.5	71.0
AndMask	69.9 ± 1.4	69.4 ± 1.4	68.2 ± 1.3	64.0 ± 1.4	87.4 ± 0.9	76.7 ± 1.3	70.0 ± 1.5	69.1 ± 1.5	68.0 ± 1.3	64.7 ± 1.5	70.7
DIVA	69.7 ± 1.4	69.5 ± 1.3	68.2 ± 1.4	63.9 ± 1.5	87.5 ± 1.0	76.5 ± 1.3	69.9 ± 1.5	69.1 ± 1.5	68.1 ± 1.3	64.7 ± 1.5	70.7
LSSAE (Ours)	70.0 ± 1.4	69.8 ± 1.4	69.0 ± 1.5	65.4 ± 1.4	85.1 ± 1.1	76.0 ± 1.4	70.1 ± 1.7	69.9 ± 1.3	69.0 ± 1.6	66.3 ± 1.4	71.1