Time-Series Domain Adaptation via Sparse Associative Structure Alignment: Learning Invariance and Variance

Zijian Li, Ruichu Cai, *Senior Member, IEEE,* Jiawei Chen, Yuguang Yan, Wei Chen, Keli Zhang, Junjian Ye

***Abstract*—Domain adaptation on time-series data is often encountered in the industry but received limited attention in academia. Most of the existing domain adaptation methods for time-series data borrow the ideas from the existing methods for non-time series data to extract the domain-invariant representation. However, two peculiar difficulties to time-series data have not been solved. 1) It is not a trivial task to model the domain-invariant and complex dependence among different timestamps. 2) The domain-variant information is important but how to leverage them is almost underexploited. Fortunately, the stableness of causal structures among different domains inspires us to explore the structures behind the time-series data. Based on this inspiration, we investigate the domain-invariant unweighted sparse associative structures and the domain-variant strengths of the structures. To achieve this, we propose Sparse Associative structure alignment by learning Invariance and Variance (SASA-IV in short), a model that simultaneously aligns the invariant unweighted spare associative structures and considers the variant information for time-series unsupervised domain adaptation. Technologically, we extract the domain-invariant unweighted sparse associative structures with a unidirectional alignment restriction and embed the domain-variant strengths via a well-designed autoregressive module. Experimental results not only testify that our model yields state-of-the-art performance on three real-world datasets but also provide some insightful discoveries on the knowledge transfer.**

***Index Terms*—Time-series Data, Time-series Domain Adaptation, Transfer Learning, Sparse Associative Structure.**

I. Introduction

Unsupervised domain adaptation (UDA) [1]–[4] is proposed to address the problem named "*domain shift*" [5], in which the source distribution and the target distribution are different. Because of the great success in the non-time series data, many researchers have extended the existing works for non-time series data to the scenario of time-series data. The existing methods [6]–[8] for time-series UDA usually combine the neural architectures for time-series data like recurrent neural networks [9], [10] and adversarial learning methodology like gradient reversal layer (GRL) [1] to extract the domain-invariant representation. Recently, Liu et.al [8] use the Fourier spectral theory and propose the adversarial spectral kernel matching method to extract the domain-invariant information for time-series data.

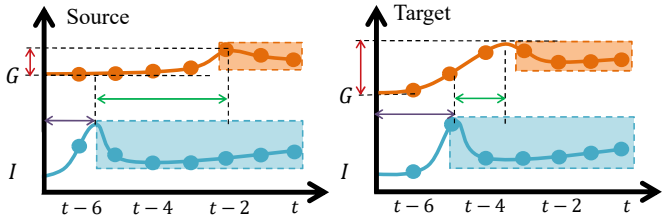


Fig. 1. The illustration of the relationships between "Growth Rate (G)↓" and "Irrigation Volume (I)↓", the decrease of G results in the decrease of I. The different lengths of red double-head arrows denote different value ranges. The different lengths of blue double-head arrows denote different offsets. The different lengths of green double-head arrows denote different response times. Different response time means different time lags. In growth rate prediction scenario, the conditional distribution $P(G_t|G_{1:t-1}, I_{t-1})$ are influenced by value ranges offsets and time lags. (*Best view in color.*)

Though taking the first step, two core challenges of time-series domain adaptation have not been addressed: (1) How to extract the domain-invariant information for time-series data? (2) How to use the domain-variant factors for model prediction for time-series data? As for the first question, the existing methods assume that the domain-invariant information can be extracted via a single feature extractor, but this assumption is hard to be satisfied in time-series data because these feature extractors are hard to capture how the variables relate to others and even the first-order Markov dependence between any two timestamps results in the variance of conditional distributions of different domains, i.e., $P_S(y_t|\phi(\boldsymbol{x}_1, \cdots, \boldsymbol{x}_t)) \neq P_T(y_t|\phi(\boldsymbol{x}_1, \cdots, \boldsymbol{x}_t))$. As shown in Figure 1, even small discrepancies between the source and the target domains may lead to sharp changes in conditional distributions. As for the second question, domain-variant information like the strength of the association is rarely considered but plays an important role in model prediction. For example, the influence degree between "Irrigation Volume" and "Growth Rate" varies with different plants, which should be taken into consideration in

Zijian Li, Jiawei Chen, Yuguang Yan are with the School of Computing, Guangdong University of Technology, Guangzhou China, 510006. E-mail: {leizigin, chenjiawei952}@gmail.com, ygyan@gdut.edu.cn

Ruichu Cai, Wei Chen are with the School of Computer Science, Guangdong University of Technology, Guangzhou, China, 510006 and Peng Cheng Laboratory, Shenzhen, China, 518066. E-mail: {cairuichu, chenweidelight}@gmail.com

Keli Zhang, Junjian Ye is with Huawei Noah's Ark Lab, Huanwei, Shenzhen, China, 518116 E-mail: {zhangkeli1, yejunjian}@huawei.com

Manuscript received XX; revised XX; accepted XX. Date of publication XX XX, 2019; date of current version XX XX, 2019. This research was supported in part by National Key R&D Program of China (2021ZD0111501), National Science Fund for Excellent Young Scholars (6212200101) and Natural Science Foundation of China (61876043, 61976052). Wei Chen was supported by China Postdoctoral Science Foundation (2021M690734). (*Ruichu Cai is the Corresponding author.)

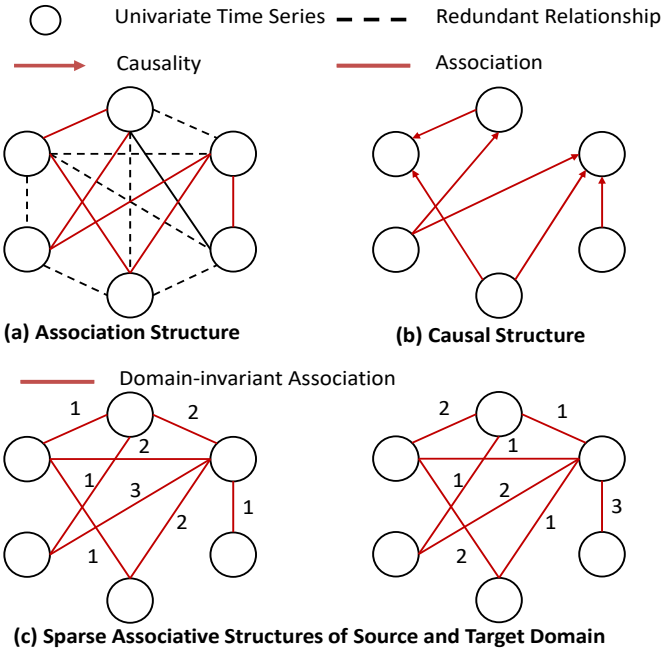


Fig. 2. The illustration of various structures among six time series. (a) The existing methods take all the relationships into account and lead to redundancy. (b) The causal structure of variables. (c) Inspired by the stability of the causal mechanism, our method considers both the domain-invariant sparse associative structure and the domain-variant strengths.

plant growth rate prediction. In summary, the existing methods, which essentially consider both the associations and the redundant relationships like Figure 2 (a), neither capture the domain-invariant sparse association nor leverage the domain-variant strength of association.

Fortunately, as shown in Figure 2(b), the causal mechanisms among different domains are stable (e.g., Irrigation Volume (I) and the Groth Rate (G) shown in Figure 1), which inspires us to explore the structures behind the time-series data. Excitingly, we find that considering the sparse associative structures can simultaneously bypass the difficulty of causal discovery and address the aforementioned two core challenges in time-series domain adaptation. Figure 2(c) illustrates that the unweighted sparse associative structures from different domains usually coincide with the causal structures and are invariant, while the strengths of the associative structures usually vary with different domains. Therefore, we can solve the time-series domain adaptation problem under the framework of sparse associative structures.

Following the aforementioned intuition, we propose the **S**parse **A**ssociative **S**tructure **A**lignment by Learning **I**nvariance and **V**ariance (SASA-IV in short) for time-series unsupervised domain adaptation. Technologically, we first propose the adaptive segment summarization to mitigate the obstacle of offset. Second, we extract the unweighted sparse associative structures with the help of intra-variables and the inter-variables attention mechanisms. Third, we encode the domain-variant strengths with the help of a well-designed autoregressive module. Finally, we propose the unidirectional alignment restriction to guarantee the correct transformation direction. Moreover, we also provide theoretical analysis for the proposed methods, in which the generalization risk on the target domain depends on the heuristic structural generation distance. Extensive experimental studies demonstrate that the proposed **SASA-IV** approach outperforms the several state-of-the-art time-series UDA methods on three real-world datasets.

This paper involves a substantial extension to its conference version [11], the extra contributions can be shown as follows:

- Compared with the conference version, we discuss what the invariant factor is and what the variant factor is in time-series unsupervised domain adaptation, which provide the more novel insight for time-series unsupervised domain adaptation.
- We improve the previous SASA model in the following three folds. First, we align the domain-invariant unweighted sparse associative structures for knowledge transfer; Second, we propose the unidirectional alignment restriction to guarantee for correct transformation; Third, we encode the domain-variant strengths with the help of well-designed autoregressive autoencoder.
- As for the theoretical analysis, we first propose a novel distribution metric for time-series data then provide a generalization bound for time-series unsupervised domain adaptation for the proposed methods, where the generalization risk on the target domain depends on the risk on the source domain and the distribution metric.
- We compare our method with the latest state-of-the-art on 3 real-world datasets and validate the effectiveness with a series of ablation experiments. Moreover, we provide some insightful visualization results, showing what knowledge can be transferred.

The rest of the paper is organized as follows. Section II reviews existing studies on domain adaptation for non-time series data and time-series data. In section III, we first provide the problem definition of unsupervised domain adaptation for time-series data. Then we elaborate on the details of the proposed SASA-IV. We also provide the theoretical analysis in section IV. Section V presents the experiment results on three real-world datasets, including ablation analysis and the visualization. Section VI concludes the paper.

II. Related Works

In this section, we first review the existing techniques about unsupervised domain adaptation for non-time series and time-series data, then we review the works about time-series relational reasoning.

A. Unsupervised Domain Adaptation on Non-Time Series Data.

To handle the "*domain shift*" issue between the source and the target domains, unsupervised domain adaptation has been proposed and applied in various fields [12]–[18]. Most of unsupervised domain adaptation methods follow the covariate shift assumption and aim to extract the domain-invariant representation [19]. They can be categorized into the *Maximum Mean Discrepancy based* methods and *adversarial training based*

methods. Recently, motivated by the stableness of causality [20], Zhang et.al [21]–[24] consider other more challenging scenarios and model how the data distribution changes across different domains.

1) Maximum Mean Discrepancy based methods.: The Maximum Mean Discrepancy (MMD) based [25]–[27] methods used the maximum mean discrepancy to measure and reduce the distance of extracted feature. Tzeng et.al [2] introduce the adaptation layer with an additional domain confuse loss to learn the domain-invariant representation. Long et.al [28] propose the deep adaptation networks to extract the domain-invariant representation with the help of multiple kernel maximum mean discrepancy. Assuming that the source and target classifiers differ by a small residual function, Long et.al [4] further propose the residual transfer network to explicitly learn the residual function with reference to the target classifier. Yan et.al [29] consider the changes of class prior distributions and propose the weighed MMD, where the class-specific auxiliary weights are brought into the original MMD.

2) Adversarial training based methods.: The adversarial training based methods borrow the ideas of generative adversarial networks [30] and extract the domain-invariant representation with the help of the domain classifier. Considering that the domain-invariant representation should be similar, Ganin et.al [1] employ the gradient reversal layer to address the UDA problem. Aiming to minimize the intra-class feature distance, Xie et.al [31] propose the moving average centroid alignment method by combining the adversarial training and pseudo label technique. To design a more effective invariant feature space, Gu et.al [32] raise the adversarial domain adaptation method that is defined in the spherical feature space. Recently, Long et.al [33] introduce the margin disparity discrepancy as a new measurement to generalization bound and implement it with the help of gradient reversal layer.

3) Causality based methods.: Since the covariate shift assumption might be not satisfied, Zhang et.al [23] propose the target shift, conditional shift and generalized target shift. Based on the causal generation process, Cai et.al [21] propose the disentangled semantic representation framework, in which the semantic information and the domain information are disentangled with the help of the variational autoencoders (VAE) [34]. Ren et.al [35] study the generalized conditional domain adaptation problem and propose the propose transforming the class conditional probability matching to the marginal probability matching. Considering the domain adaptation as a graphical inference problem, Zhang et.al [22] model how the joint distribution changes by discovering the causal structures and inferring the label under the causal structures. Recently, Petar and Li et.al [24] find that domain-invariant representation can not be extracted with the help of a single encoder when the support overlap exists, so they take the domain-specific information into account and propose the domain-specific adversarial network under the causal generation process.

In this paper, we are also inspired by the stable causal structures assumption. Since discovering the causal structures behind the data is another challenging task that is usually hindered by other factors like hidden confounders, we relax the stable causal structures assumption to stable sparse associative structure assumption and apply it into the time-series UDA problem.

B. Unsupervised Domain Adaptation on Time-series Data.

Time series is one of the most familiar data that can be found in many applications. Time-series domain adaptation is an important problem but desiderates more concern. Recently, more and more attention has been paid to the time-series domain adaptation. Da Costa et al. [6] straightforwardly employ the feature extractors that are designed for time-series data like RNN [9] and VRNN [10] and reuse the conventional unsupervised adaptation frameworks [1]. Wilson et.al [36] propose the CoDATS that leverages the weak supervision in the form of target-domain label distribution. Ragab et.al [37] propose the self-supervised autoregressive domain adaptation framework for time-series data which introduces the autoregressive model to model the temporal dependency. Recently, considering that the low-order and local statistics have limited expression for time-series distribution, Liu et.al [8] propose the adversarial spectral kernel matching method with the help of Fourier transform.

Based on the previous sparse associative structure alignment model (SASA) [11], we make a substantial extension and propose SASA-IV. In this paper, we simultaneously consider the invariant unweighted sparse associative structures and the variant strength information.

C. Time-series Relational Reasoning

The proposed method also relates to the time-series relational reasoning problem. Aiming to mine the generalized and explainable representation between entities and their properties, time-series relational reasoning [38] aims to explore the inter-sample relation and intra-temporal relation of time-series data to learn the underlying the structures. In the past decades, several researchers [39], [40] pay lots of attention on this field. Cao et.al [41] represent both intra-series and inter-series correlations in the spectral domain for time-series forecasting task. Li et.al [42] propose a dynamic mechanism to infer the evolving latent graph for trajectory forecasting. Fan et.al [43] infer the temporal relationships by sampling the time pieces from the anchor samples in the scenario of self-supervised learning. Kipf [44] propose the neural relation inference model to infer interactions while simultaneously learning the dynamics purely from observational data under the framework of variation auto-encoder [34].

In this paper, we reconstruct and align the relationships of time-series data for unsupervised domain-adaptation for time-series domain adaptation, which simultaneously considers the domain-invariant sparse associative structures and domain-variant weights.

III. Sparse Associative Structure Alignment

In this paper, we first provide the problem definition of time-series domain adaptation, then we introduce the proposed **SASA-IV**. Our **SASA-IV** method is motivated by the process from stable causality assumption to relaxed stable association

assumption. Under this intuition, we devise a unified model to align the invariant unweighted sparse associative structure and encode the variant strengths for time-series domain adaptation.

A. Problem Formulation and Model Overview

In this subsection, we first formulate the problem of time-series domain adaptation. Then we provide the overview of the proposed SASA-IV model.

We let $\boldsymbol{x} = \{\boldsymbol{x}_{t-N+1}, \cdots, \boldsymbol{x}_{t-1}, \boldsymbol{x}_t\}$ denote a multivariate time series sample with N time steps, where $\boldsymbol{x}_t \in \mathbb{R}^M$, and $y \in \mathbb{R}$ is the corresponding label. We assume that $P_S(\boldsymbol{x}, y)$ and $P_T(\boldsymbol{x}, y)$ are different distributions from the source and the target domains but are generated from a shared causal mechanism. Since the two variable sets generated by the same causal structure should share the same associative structure, $P_S(\boldsymbol{x}, y)$ and $P_T(\boldsymbol{x}, y)$ share the same associative structure. $(\mathcal{X}_S, \mathcal{Y}_S)$ and $(\mathcal{X}_T, \mathcal{Y}_T)$, which are sampled from $P_S(\boldsymbol{x}, y)$ and $P_T(\boldsymbol{x}, y)$ respectively, denote the source and target domain dataset. In unsupervised domain adaptation, each source domain sample $\boldsymbol{x}_S$ comes with y_S, while the target domain has no labeled sample. Our goal is to devise a predictive model that can predict y_T given time series sample $\boldsymbol{x}_T$ from the target domain.

Based on the aforementioned problem definition, we aim to extract the domain-invariant structure information and encode the domain-variant strength information under a unified framework of sparse associative structures. The solution is inspired by the intuition that the causal mechanism is invariant across different domains. Due to the complexity of discovering causal structures, we relax the causal structures to the sparse associative structures. Considering that the offsets vary with different domains and hinder the model from extracting the domain-invariant associative structures, we first elaborate on how to obtain the fine-grain segments of time-series data to ease the obstacle of the offsets. Second, considering time lags from different domains, we reconstruct the unweighted associative structures with the help of the intra-variables and the inter-variables attention mechanisms. Different from the existing works that align the feature from different domains, the proposed method employs the unidirectional sparse associative structure alignment restriction to obtain the common associative structures from different domains to indirectly extract the domain-invariant representation. We encode the domain-variant strength information with the help of domain-variant autoencoder.

B. Adaptive Segment Summarization

In this subsection, we will elaborate on how to obtain the candidate segments to remove the obstacle of offsets. As shown in Figure 1, the orange blocks, whose duration varies with different domains, denote the segment of the change of variable 'G'. Existing methods, which take the whole time-series data as input, can not accurately capture when a segment stars and when a variable affects the others, i.e., the sphere of influence of any variables. Therefore, these methods can not address the obstruction of offsets (i.e., the duration between the start point of time-series and the start point of a segment).

To address the this problem, we first propose the adaptive segment summarization, which is shown in Figure 3(a). To obtain the candidate segments of i-th time-series $\tilde{\boldsymbol{x}}^i$, we construct multiple segments with different length for each variable $\boldsymbol{x}^i$. Hence, we let $\tilde{\boldsymbol{x}}^i$ be the segment set of $\boldsymbol{x}^i$ shown as Equation (1):

$$\tilde{\boldsymbol{x}}^i = \{\boldsymbol{x}^i_{t:t}, \boldsymbol{x}^i_{t-1:t}, \cdots, \boldsymbol{x}^i_{t-\tau+1:t}, \cdots, \boldsymbol{x}^i_{t-N+1:t}\}, \tag{1}$$

Motivated by RIM [45], we allocate an independent LSTM for each variable. In detail, given a segment of i-th variable with τ timestamps, we have:

$$\boldsymbol{h}^i_\tau = f_i(\boldsymbol{x}^i_{t-\tau+1:t}; \boldsymbol{\theta}^i), \tag{2}$$

in which $f_i(\cdot)$ denote the i-th long term short term memory (LSTM) for $\boldsymbol{x}^i$ and $\boldsymbol{\theta}^i$ denote the parameters of $f_i(\cdot)$. For convenience, we let $\Theta = \{\theta^1, \cdots, \theta^i, \cdots, \theta^M\}$ be parameters of all the LSTM. Note that the segments in the same segment set $\tilde{\boldsymbol{x}}^i$ share the same LSTM. And finally we can obtain the segments representation set shown as follow:

$$\boldsymbol{h}^i = \{\boldsymbol{h}^i_1, \cdots, \boldsymbol{h}^i_2, \cdots, \boldsymbol{h}^i_N\}. \tag{3}$$

Since it is almost impossible to consider all the exact segments from the multivariable time-series data, we first obtain the representation of all candidate segments via the aforementioned processing. The most suitable segment representation are selected and used to reconstruct the associative structure, which will be described in the following subsections.

C. Sparse Associative Structure Discovery

In this section, we will introduce how to generalize the most exact segment representation and how to reconstruct the associative structure with the help of intra-variable attention mechanism and inter-variable attention mechanism respectively.

1) Segments Representation Selection via Intra-Variables Attention Mechanism.: In order to get rid of the obstacle brought from the offsets, we need to pay more attention to the exact segment representation among all the candidate segment representations with the help of the self attention mechanism [46]. Formally, we calculate the weights of each segment of $\boldsymbol{x}^i$ as follow:

$$\begin{aligned} \boldsymbol{\alpha}^i =& [\alpha^i_1, \cdots, \alpha^i_\tau, \cdots, \alpha^i_N] \\ =& \text{sparsemax}\left([u^i_1, \cdots, u^i_\tau, \cdots, u^i_N]\right), \\ u^i_\tau =& \frac{1}{N}\sum_{k=1}^{N} \frac{\left(\boldsymbol{W}^Q \boldsymbol{h}^i_\tau\right)^\top \left(\boldsymbol{W}^K \boldsymbol{h}^i_\tau\right)}{\sqrt{d_h}}, \end{aligned} \tag{4}$$

in which $\boldsymbol{W}^Q$ and $\boldsymbol{W}^K$ are the trainable projection parameters and $\sqrt{d_h}$ is the scaling factor. In order to obtain the sparse weights that represent specific segment representation clearly, we employ sparsemax [47] to calculate the weights. The sparsemax is defined as:

$$\text{sparsemax}(\boldsymbol{z}) = \arg\min_{\boldsymbol{p} \in \Delta^{K-1}} ||\boldsymbol{p} - \boldsymbol{z}||^2, \tag{5}$$

which returns the Euclidean projection of vector $\boldsymbol{z} \in \mathbb{R}^K$ onto probability simplex Δ^{K-1}.

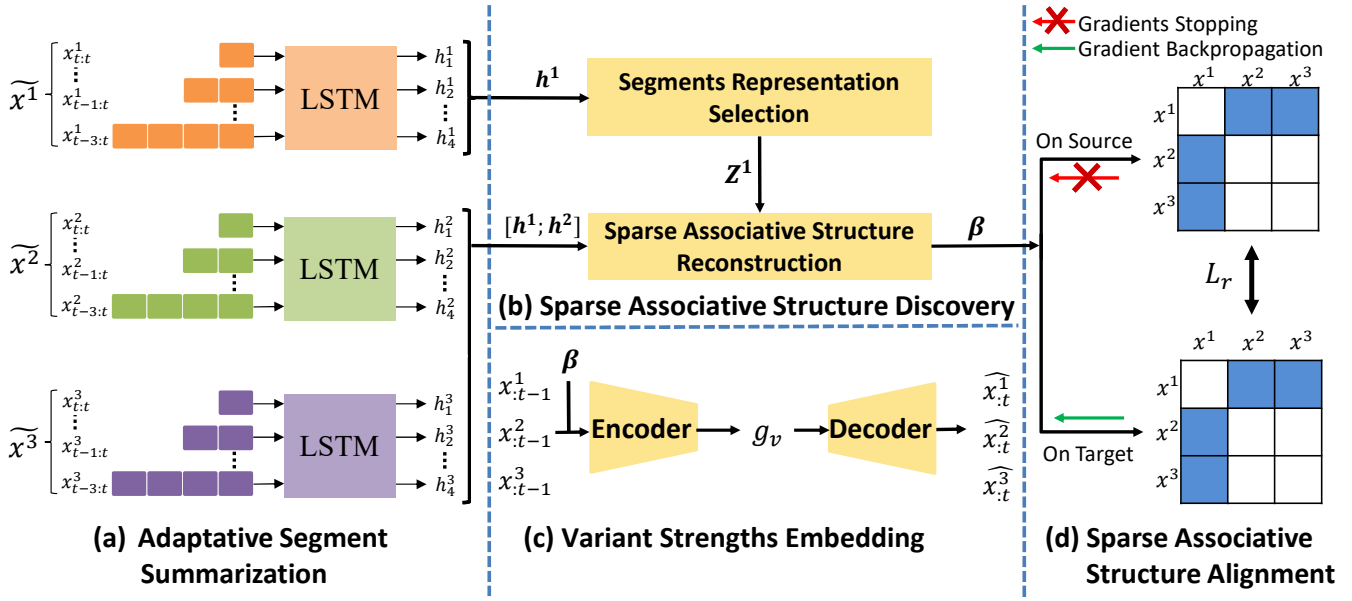


Fig. 3. The framework of the SASA-IV model. (a) Adaptative segment summarization process with variable-specific LSTM. (b) Unweighted sparse associative structure discovery via intra-variables and inter-variables attention mechanism. (c) Variant Strength Embedding module to encode the domain-variant strength information. (d) Unidirectional sparse associative structure alignment between the source and the target domain. (*Best view in color.*)

2) Sparse Associative Structure Reconstruction via Inter-variables Attention Mechanism.: With the help of the intra-variables attention mechanism, we can extract the weighted segment representations despite the obstacle of offsets, which is shown as follow:

$$\boldsymbol{Z}^i = \sum_{\tau=1}^{N} \alpha_\tau^i \left(\boldsymbol{W}^V \boldsymbol{h}_\tau^i \right), \tag{6}$$

in which $\boldsymbol{W}^V$ is trainable projection parameter. Note that α also denotes the probability of the length of a segment.

Then we leverage these weighted segment representation to reconstruct the sparse associative structure among variables. So we propose the inter-variables attention mechanism to mine the associative structure among variables.

In this part, our goal is to reconstruct the associative structure among variables. Technologically, we employ the standard attention mechanism [48]. One of the most straightforward methods to calculate the degree of correlation of variable i and variable j is shown as follow:

$$\boldsymbol{e}^{ij} = \frac{\boldsymbol{Z}^i \cdot \boldsymbol{Z}^j}{||\boldsymbol{Z}^i|| \cdot ||\boldsymbol{Z}^j||}. \tag{7}$$

However, the associative structure calculated by Equation (7) ignores the time lags from different domains between i-th and j-th variables, which may result in the false estimation for the associative structures. In order to take the time lags into consideration, we calculate the degrees of association between i-th variable and j-th variable by:

$$\begin{aligned} e_\tau^{ij} &= \frac{\boldsymbol{Z}^i \cdot \boldsymbol{h}_\tau^j}{||\boldsymbol{Z}^i|| \cdot ||\boldsymbol{h}_\tau^j||}, \\ \boldsymbol{e}^{ij} &= \{e_1^{ij}, \cdots, e_\tau^{ij}, \cdots, e_N^{ij}\}, \end{aligned} \tag{8}$$

Then we normalized these degrees of association with Sparsemax [47]. Formally, we have:

$$\begin{aligned} \boldsymbol{\beta}^i &= [\boldsymbol{\beta}^{i1}, \cdots, \boldsymbol{\beta}^{ij}, \boldsymbol{\beta}^{iM}] \\ &= \text{sparsemax}\left([\boldsymbol{e}^{i1}, \cdots, \boldsymbol{e}^{ij}, \cdots, \boldsymbol{e}^{iM}] \right) (j \neq i). \end{aligned} \tag{9}$$

Note that $\beta_\tau^{ij} \in \boldsymbol{\beta}^i$ denotes the associative strength between i-th variables and j-th variables with regard to segment duration of τ.

Similar to Equation (6), we calculate the associative structure representation of the i-th variable as follows:

$$\begin{aligned} \boldsymbol{U}^{ij} &= \sum_{\tau=1}^{N} \beta^{ij} \cdot \boldsymbol{h}_\tau^j, \\ \boldsymbol{U}^i &= \sum_{m=1, m\neq i}^{M} \boldsymbol{U}^{im}. \end{aligned} \tag{10}$$

D. Sparse Associative Structure Alignment

1) Unweighted Sparse Associative Structure Alignment.: Based on the associative structures that are extracted in Equation (9), we need to restrict the distance of the structure between the source and the target domains for extracting the domain-invariant associative structures. The conference version borrows the idea of domain confuse network and restricts the distance of $\boldsymbol{\beta}$ with the help of maximum mean discrepancy (MMD) [28], which is shown in Equation 11.

$$\mathcal{L}_\beta = \textbf{MMD}(\boldsymbol{\beta}_S, \boldsymbol{\beta}_T). \tag{11}$$

Since the time lags between any two variables may be similar but different, the duration of segments might change over different domains. Therefore, in order to reconstruct the associative structure more precisely, we minimize the **MMD** between $\boldsymbol{\alpha}$ from the source and the target domain to align the

offsets. It restricts the duration of the segment from different domains to be similar, which contributes to extracting structure for transfer. Formally, we have:

$$\mathcal{L}_\alpha = \text{MMD}(\boldsymbol{\alpha}_S, \boldsymbol{\alpha}_T). \tag{12}$$

Though the associative structures are stable, there are some domain-variant factors like the weights of the associative structures. Furthermore, domain-variant weights play an important role in forecasting. For example, we let gravity and mass be composed of an associative structure, the value of gravity also depends on the gravitational acceleration, which denotes the weights of the associative structure. Straightforwardly minimizing the distance between the structures with strengths will sacrifice the domain-variant information, which further degenerates the model performance. In order to address the aforementioned problem, we separate the weights from the associative structures and minimize the discrepancy between the unweighted structures from different domains, hence we modify Equation (11) and (12) to the Equation (13):

$$\begin{aligned} \mathcal{L}_\beta &= ||\mathbb{1}(\boldsymbol{\beta}_S > \mu) - \mathbb{1}(\boldsymbol{\beta}_T > \mu)||_1, \\ \mathcal{L}_\alpha &= ||\mathbb{1}(\boldsymbol{\alpha}_S > \mu) - \mathbb{1}(\boldsymbol{\alpha}_T > \mu)||_1, \end{aligned} \tag{13}$$

in which $||\cdot||_1$ denotes the L1 norm and the indicator function $\mathbb{1}(\cdot)$ is used to choose the edges with weights greater than μ. We also find that changing the value μ can control the sparseness of the generated associative structures, so we can remove more redundant relationships by tuning μ.

2) Unidirectional alignment restriction.: As mentioned in Equation (13), we extract the unweighted associative structures with the help of attention mechanisms, which are essential profited from the labeled source data. And the target associative structures might be wrongly extracted without any supervised signal. If we directly employ Equation (13) to align the structures, the wrong structures from the target domain might be aligned to the graph of source domain. In the worst case, the source structures would be totally wrong, which might result in the negative transfer.

In order to address the this issue, we provide a simple but effective solution that prevent the knowledge transfer from target to source. In detail, we apply the gradient stopping operation $\mathcal{C}(\cdot)$ on the source associative structures, which is shown as follows:

$$\begin{aligned} \mathcal{L}_\beta &= ||\mathcal{C}(\mathbb{1}(\boldsymbol{\beta}_S > \mu)) - \mathbb{1}(\boldsymbol{\beta}_T > \mu)||_1, \\ \mathcal{L}_\alpha &= ||\mathcal{C}(\mathbb{1}(\boldsymbol{\alpha}_S > \mu)) - \mathbb{1}(\boldsymbol{\alpha}_T > \mu)||_1, \end{aligned} \tag{14}$$

Note that L1 Norm of a matrix is calculated by $||A||_1 = \sum_{ij} |A_{ij}|$, where A_{ij} is the i-th row j-th column element in matrix A.

E. Domain-variant information Extraction

In subsection III-C, we have discussed that the domain-invariant unweighted associative structures are domain-invariant while some domain-variant factors play an important role in model prediction.

However, it is not a simple task to extract the domain-variant information from the source to the target domain, one major difficulty comes from the unlabeled target domain data, since we can only extract the source-specific information with the help of source labeled data. Fortunately, the stationary time-series data, which are generated via a stable causal structure, contain the inherent autoregressive property. This inspiration enlightens us to extract the domain-variant factors in a autoregressive paradigm. Hence, we devise the domain-variant factor extraction module, which is a graph attention networks [49]–[51] based autoregressive autoencoder. Formally, the encoder and the decoder are respectively represented as:

$$\boldsymbol{g}_v = \textbf{GNN}(\mathbb{1}(\boldsymbol{\beta} > \mu)), \psi_e(\boldsymbol{x}_{1:t-1}); \boldsymbol{W}^e), \tag{15}$$

$$\hat{\boldsymbol{x}}_t = \psi_d(\boldsymbol{g}_v; \boldsymbol{W}^d), \tag{16}$$

in which $\boldsymbol{W}^g$ and $\boldsymbol{W}^d$ are the trainable parameters.

In the encoder, we first use the LSTM-based feature extractor $\psi_e(\cdot)$ to extract the feature for each variable. Note that we allocate an independent LSTM for each variable. Then we take the unweighted sparse associative structures and the $\boldsymbol{x}_{1:t-1}$ and the extracted feature as the input of graph attention networks and obtain domain-variant representation $\boldsymbol{g}_v$. In the decoder, we use $\boldsymbol{g}_v$ to generate the predicted result $\hat{\boldsymbol{x}}_t$. $\hat{\boldsymbol{x}}_t$ are the predicted future values. Finally, we use mean squared error (MSE) to optimize the autoregressive model, which is shown as follows:

$$\mathcal{L}_r = MSE(\hat{\boldsymbol{x}}_t, \boldsymbol{x}_t). \tag{17}$$

In summary, the GNN work as the encoder which approximately imitate the the data generation process under the invariant sparse associative structures; $\psi_d(\cdot)$ works as the decoder and predict the value in the final timestamp to guarantee the the domain-variant factors are preserved. So we can extract the source and target domain-variant factors by respectively using the source and target training data.

F. Model Summary

1) Task based Label Predictor.: Finally, we can obtain the final representation that contains the domain-invariant associative structure information and the domain-variant strength information as shown in Equation (18):

$$\begin{aligned} \boldsymbol{H}^i &= \boldsymbol{Z}^i \oplus \boldsymbol{U}^i, \\ \hat{\boldsymbol{H}} &= \boldsymbol{H}^1 \oplus \boldsymbol{H}^2 \oplus \cdots \oplus \boldsymbol{H}^M, \\ \boldsymbol{H} &= \hat{\boldsymbol{H}} \oplus \boldsymbol{g}_v, \end{aligned} \tag{18}$$

in which $\oplus$ denotes the concatenate operation.

For convenience, we describe the above process as :

$$\boldsymbol{H} = G_f(\boldsymbol{x}; \Theta, \boldsymbol{W}^Q, \boldsymbol{W}^K, \boldsymbol{W}^V, \boldsymbol{W}^e, \boldsymbol{W}^d). \tag{19}$$

After obtaining the final representation, we take $\boldsymbol{H}$ as the input of label classifier $G_y(\cdot; \boldsymbol{\phi})$ whose loss function is $\mathcal{L}_y$. For the classification problems, we employ cross-entropy as the label loss. For the regression problems, we employ RMSE as the label loss.

2) Objective Function.: The total loss of the proposed structure alignment model for time series domain adaptation is formulated as:

$$\begin{aligned} \mathcal{L}\left(\Theta, \boldsymbol{W}^Q, \boldsymbol{W}^K, \boldsymbol{W}^V, \boldsymbol{W}^e, \boldsymbol{W}^d, \boldsymbol{\phi}\right) = \\ \mathcal{L}_y + \omega(\mathcal{L}_\alpha + \mathcal{L}_\beta) + \gamma \mathcal{L}_r, \end{aligned} \tag{20}$$

in which ω and γ are hyper-parameters.

Under the above objective function, our model is trained on the source and target domain using the following procedure:

$$\arg \min_{\boldsymbol{\Theta}, \boldsymbol{W}^Q, \boldsymbol{W}^K, \boldsymbol{W}^V, \boldsymbol{W}^e, \boldsymbol{W}^d, \boldsymbol{\phi}} \mathcal{L}\left(\boldsymbol{\Theta}, \boldsymbol{W}^Q, \boldsymbol{W}^K, \boldsymbol{W}^e, \boldsymbol{W}^d, \boldsymbol{W}^V, \boldsymbol{\phi}\right). \tag{21}$$

IV. Theoretical Analysis

Several works have focused on the generalization theory of domain adaptation [33], [52], [53], [53]–[55], which mainly depends on the distance of distribution between the source and the target domains. In this paper, we first provide the straightforward definition of time-series structural distance based on the generation process of time-series data, then establish the generalization bound for time-series unsupervised domain adaptation, in which the generalization risk on the target domain depends on not only the risk on the source data but also the time-series structural distance between the source and the target domains.

A. Structural Generation Distance for Time-Series Data.

In this subsection, we provide the heuristic definition of Structural Distance (**SD**) for time-series data. We first introduce the data generation process for time-series data based on structural causal models (**SCM**), which is shown as follows:

$$x_t^i = F_i(\boldsymbol{pa}(x_t^i), N_i), \tag{22}$$

in which F_i denotes any flexible types of function for i-th time-series data x^i; $\boldsymbol{pa}(x_t^i)$ denotes the parents of x_t^i; and N_i denotes the independent noise terms. According to Equation (22), we can easily find that the distribution of x_t^i depends on the distribution of $\boldsymbol{pa}(x_t^i)$ and the causal mechanism.

Inspired by the aforementioned structural causal models, it is straightforward to define the distance for time-series data. As an effective metric for time-series data, the new distance should satisfy all necessary axioms for a general metric.

Definition 1. *(**Structural Distance, SD**.) We let $\boldsymbol{x}$ be the multivariate time series, and $\boldsymbol{x}_0$ be the values of the first timestamps. We further assume that $\boldsymbol{x}$ is generated by causal structures A with the strength $\mathcal{W}$. Given $\boldsymbol{x}_S$ and $\boldsymbol{x}_T$ from the source and the target domain, the structural distance between $\boldsymbol{x}_S$ and $\boldsymbol{x}_T$ can be formalized as follows:*

$$\begin{aligned}\text{dis}_{SD}^{\mathcal{S}\leftrightarrow\mathcal{T}}(\boldsymbol{x}_S, \boldsymbol{x}_T) = dist(P_S(\boldsymbol{x}_0), P_T(\boldsymbol{x}_0)) + ||A_S - A_T||_1 \\ + ||\mathcal{W}_S - \mathcal{W}_T||_1.\end{aligned} \tag{23}$$

Note that $dist(\cdot, \cdot)$ in Equation (23) can be any distance metric for distribution, and we employ the total variation distance in the following generalization bound; A_S and A_T can be the sparse associative structures corresponding to their causal structures.

According to Equation (23), we can find that the distance of time-series distribution is affected by the following three factors:

- The first term $dist(P_S(\boldsymbol{x}_0), P_T(\boldsymbol{x}_0))$ denotes the distribution distance of the start points $\boldsymbol{x}_0$. An comprehensible example is that the value ranges of time-series data influence the distribution. Note that $dist(\cdot, \cdot)$ denotes any type of distribution metric for static data.
- The second term $||A_S - A_T||_1$ denotes the distance between the unweighted sparse associative structures from the source and target domains.
- The third term $||\mathcal{W}_S - \mathcal{W}_T||_1$ denotes the distance between the strengths of causal structures from the source and target domains.

Moreover, we find that the structural generation distance satisfies the three axioms for a general metric:

Theorem 1. *The Structural Distance (**SD**) satisfies the three axioms for a general metric, to be specific, it satisfies the following conditions:*

(1) $\text{dis}_{SD}^{\mathcal{S}\leftrightarrow\mathcal{T}} \geq 0$ *and* $\text{dis}_{SD}^{\mathcal{S}\leftrightarrow\mathcal{T}} = 0$ *if and only if* $\mathcal{S} = \mathcal{T}$*;*

(2) $\text{dis}_{SD}^{\mathcal{S}\leftrightarrow\mathcal{T}}(\boldsymbol{x}^S, \boldsymbol{x}^T) = \text{dis}_{SD}^{\mathcal{S}\leftrightarrow\mathcal{T}}(\boldsymbol{x}^T, \boldsymbol{x}^S)$ *(symmetric);*

(3) $\text{dis}_{SD}^{\mathcal{S}\leftrightarrow\mathcal{T}} \leq \text{dis}_{SD}^{\mathcal{S}\leftrightarrow\mathcal{D}} + \text{dis}_{SD}^{\mathcal{D}\leftrightarrow\mathcal{T}}$ *(triangle inequality).*

Theorem 1 are easy to be proved, it is not hard to find that time-series structurally generation distance can be used to measure the distance between the distribution of time-series data.

B. Generalization Bounds

In this section, we provide the generalization bound for time-series unsupervised domain adaptation. We first formalize some notations that will be used in the following statement. Suppose $\mathcal{X}$ be an instance set of time-series and $\{0, 1\}$ be the label set for binary classification. We let $\mathcal{H}$ be a hypothesis space that maps $\boldsymbol{x}$ to $\mathbb{R}$ and $\forall h \in \mathcal{H}, h : \mathcal{X} \to \{0, 1\}$. We further let $\eta : \mathcal{X} \to \{0, 1\}$ be the labeling function. The probability according to the distribution $P_{\mathcal{S}}$ is defined as $\epsilon_{\mathcal{S}}(h, f) = \mathbb{E}_{\boldsymbol{x}\sim P_{\mathcal{S}}}[|h(\boldsymbol{x}) - \eta_S(\boldsymbol{x})|]$. We use the shorthand $\epsilon_S(h) = \epsilon(h, \eta_S)$ and $\epsilon_T(h)$ is defined the same. Based on the aforementioned definition, we make the following assumption:

Assumption 1. ***Time-series distribution bound assumption:*** *Suppose that P_S and P_T are the source and the target distributions and A_S and A_T are sparse associative structures with strengths $\mathcal{W}_S, \mathcal{W}_T$, then a positive value K exists that makes the following inequality hold:*

$$\begin{aligned}|P_{\mathcal{S}}(\boldsymbol{x}) - P_{\mathcal{T}}(\boldsymbol{x})| \leq & K(|P_{\mathcal{S}}(\boldsymbol{x}_0) - P_{\mathcal{T}}(\boldsymbol{x}_0)| + ||A_S - A_T||_1 \\ & + ||\mathcal{W}_S - \mathcal{W}_T||_1) \\ = & K\text{dis}_{SD}^{\mathcal{S}\leftrightarrow\mathcal{T}}(\boldsymbol{x}_S, \boldsymbol{x}_T),\end{aligned} \tag{24}$$

Note that the third line establishes when $dist(P_S(\boldsymbol{x}_0), P_T(\boldsymbol{x}_0))$ is the total variation distance.

Based on the aforementioned definition and assumption, we propose the generalization bound of the propose **SASA-IV**, which is shown as follows.

Theorem 2. ***(Generalization Bound for Time-Series Unsupervised Domain Adaptation.)*** *Given Assumption 1, we have:*

$$\begin{aligned}\epsilon_T(h) \leq & \epsilon_S(h) + K(|P_{\mathcal{S}}(\boldsymbol{x}_0) - P_{\mathcal{T}}(\boldsymbol{x}_0)| + ||A_S - A_T||_1 \\ & + ||\mathcal{W}_S - \mathcal{W}_T||_1) + \lambda, \\ = & \epsilon_S(h) + K\text{dis}_{SD}^{\mathcal{S}\leftrightarrow\mathcal{T}}(\boldsymbol{x}_S, \boldsymbol{x}_T) + \lambda\end{aligned} \tag{25}$$

in which K, λ are constants and $h \in \mathcal{H}$ a any hypothesis.

Proof.

$$\begin{aligned}
&\epsilon_T(h) = \epsilon_T(h) + \epsilon_S(h) - \epsilon_S(h) + \epsilon_S(h,\eta_T) - \epsilon_S(h,\eta_T) \\
\leq &\epsilon_S(h) + \epsilon_S(h,\eta_T) - \epsilon_S(h,\eta_S) + |\epsilon_T(h) - \epsilon_S(h,\eta_T)| \\
\leq &\epsilon_S(h) + \epsilon_S(\eta_S,\eta_T) + |\epsilon_T(h) - \epsilon_S(h,\eta_T)| \\
\leq &\epsilon_S(h) + \epsilon_S(\eta_S,\eta_T) + \int |P_S(\boldsymbol{x}) - P_T(\boldsymbol{x})||h(\boldsymbol{x}) - \eta_T(\boldsymbol{x})| d\boldsymbol{x} \\
\leq &\epsilon_S(h) + \epsilon_S(\eta_S,\eta_T) + \int |P_S(\boldsymbol{x}) - P_T(\boldsymbol{x})| d\boldsymbol{x} \\
\leq &\epsilon_S(h) + K(|P_{\mathcal{S}}(\boldsymbol{x}_0) - P_{\mathcal{T}}(\boldsymbol{x}_0)| + ||A_S - A_T||_1 \\
&+ ||\mathcal{W}_S - \mathcal{W}_T||_1) + \epsilon(\eta_S,\eta_T) \\
= &\epsilon_S(h) + K\mathrm{dis}_{\mathrm{SD}}^{\mathcal{S}\leftrightarrow\mathcal{T}}(\boldsymbol{x}_S,\boldsymbol{x}_T) + \lambda
\end{aligned} \tag{26}$$

in which $\lambda = \epsilon(\eta_S, \eta_T)$ is a constant. □

According to this generalization bound, we can find that the expected risk $\epsilon_T(h)$ is not only controlled by $\epsilon_S(h)$ but also the distance between the unweighted associative structures, the distance between the strengths as well as the structural generation distance between the source and the target domains. It inspires us to extract the domain-invariant unweighted structures and weights, which is essentially employed in the conference version. In the meanwhile, the proposed **SASA-IV** takes the domain-variant factors into consideration, so the strengths alignment term $||\mathcal{W}_S - \mathcal{W}_T||_1$ and the start point distribution alignment term $|P_{\mathcal{S}}(\boldsymbol{x}_0) - P_{\mathcal{T}}(\boldsymbol{x}_0)|$ can be removed. Hence we can obtain a tighter bound, which is derived Equation (25) as follows:

$$\epsilon_T(h) \leq \epsilon_S(h) + K||A_S - A_T||_1 + \lambda. \tag{27}$$

V. Experiments

A. Dataset

1) Air Quality Forecast Dataset.: The air quality forecast dataset [56] is collected in the Urban Air project[1] from 2014/05/01 to 2015/04/30, which contains air quality data, meteorological data, weather forecast data, etc. The dataset covers 4 major Chinese cities: Beijing (B), Tianjin (T), Guangzhou(G), and Shenzhen(S). We employ air quality data as well as meteorological data to predict PM2.5. We choose the air quality station with the least missing value and take each city as a domain. We use this dataset because the air quality data is common and the sensors in the smart city systems usually contain complex causality. The associations among sensors are often sparse, which is suitable for our model.

2) In-hospital Mortality Prediction Dataset.: MIMIC-III [57], [58][2] is another published dataset with de-identified health-related data associated with more than forty thousand patients who stayed in critical care units of the Beth Israel Deaconess Medical Center between 2001 and 2012. It's the benchmark of time series domain adaptation in VRADA [7]. Similar to Purushotham et al. [59], we choose 12 time series (such as Heart Rate, Temperature, Systolic blood pressure, etc) and five static feature from 35637 records. In order to prepare the in-hospital mortality prediction dataset for time series domain adaptation, we split the patients into 4 groups according to their age (Group1: 20-45, Group2: 46-65, Group3: 66-85, Group4: >85).

3) Boiler Fault Detection Dataset.: The boiler data consists of sensor data from three boilers from 2014/3/24 to 2016/11/30. There are 3 boilers in this dataset and each boiler is considered as one domain. The learning task is to predict the *faulty blowdown valve* of each boiler. Since the fault data is very rare. It's hard to obtain the fault samples in the mechanical system. So it's important to utilize the labeled source data and unlabeled target data to improve the model generalization.

B. Compared Methods

We consider as many as possible compared methods for time-series domain adaptation, which are described as follows:

- **LSTM_S2T.** LSTM_S2T uses the source domain data to train a vanilla LSTM model and applies it to the target domain without any adaptation (S2T stands for source to target). It's expected to provide the lower bound performance.
- **R-DANN.** R-DANN [6] is an unsupervised domain adaptation architecture proposed in [1] with GRL (Gradient Reversal Layer) on LSTM, which is a straightforward solution for time series domain adaptation.
- **RDC.** Deep domain confusion is an unsupervised domain adaptation method proposed in [2] which minimizes the distance between the source and target distributions by employing Maximum Mean Discrepancy (MMD). Similar to the aforementioned R-DANN, we use LSTM as the feature extractor for time series data.
- **VRADA.** VRADA [7] is a time series unsupervised domain adaptation method which combines the GRL and VRNN [10].
- **AdvSKM** [8] is one of the latest approaches for time-series domain adaptation. The **AdvSKM** uses the hybrid spectral kernel network to reform the MMD metric.
- **CODATS** [36] leverages the convolutional neural networks and weak supervision in the form of target-domain label distribution.

C. Model Variants

In order to verify the effectiveness of each component of our model, we further devise the following model variants.

- SASA: We take the previous SASA in conference as a model variant.
- SASA-$\boldsymbol{\alpha}$: We remove $\mathcal{L}_\alpha$ to verify the usefulness of the segment length restriction loss from the SASA model.
- SASA-$\boldsymbol{\beta}$: We remove $\mathcal{L}_\beta$ to verify the usefulness of the sparse associative structure alignment loss from the SASA model.
- SASA-IV-$\boldsymbol{\alpha}$: We remove $\mathcal{L}_\alpha$ to verify the usefulness of the segment length restriction loss.
- SASA-IV-$\boldsymbol{\beta}$: We remove $\mathcal{L}_\beta$ to verify the usefulness of the sparse associative structure alignment loss.

[1] https://www.microsoft.com/en-us/research/project/urban-air/
[2] https://mimic.physionet.org/gettingstarted/demo/

- SASA-IV-$\boldsymbol{\gamma}$: We remove the $\mathcal{L}_\gamma$ to verify the usefulness of the domain-variant factors.
- SASA-IV-$\boldsymbol{C}$: We remove the gradient stopping operation $\mathcal{C}(\cdot)$ to verify the usefulness of the unidirectional alignment restriction.

D. Implementation Details

In our experiments, we follow the standard protocol of unsupervised domain adaptation and leverage the labeled source data and the unlabeled target data. All the experiments are conducted on NVIDIA GeForce RTX 2070S GPU. We use a batch size of 1024 for all the datasets. We further let $\epsilon = 0.08$ for all the datasets. Note that the presented experiment results are averaged over several replicated with different random seeds, so the values might be slightly different from the conference version that reports the best results.

E. Result

1) Results on Air Quality Forecast.: Then we further evaluate the transferability of our model on the time-series forecasting task. We consider the air quality forecast dataset since the meteorologic sensors in the smart cities system usually contain the stable causal mechanism, meaning that the associative structures are stable. So adaptively forecasting the air quality can benefit the environmental prediction. Experiment results on the air quality forecast dataset are shown in Table I.

Similar to the results in the boiler fault detection dataset, the proposed SASA-IV also achieves the best performance and outperforms all the comparison methods. According to the results, we can find that:

- Compared SASA-IV with the previous SASA, we can find that almost all the results of SASA-IV are much better than that of other methods. Note that the tasks of smaller geographical distance like $S \to B$ and $B \to S$ achieve more improvement, this is because the city pairs with closer geographical distance may not only share more common associative structures but also contain similar strengths.
- As for the city pairs with further geographical distance like Shenzhen and Tianjin, the improvement is a bit smaller, but the performance of SASA-IV on this task is still better than that of SASA, reflecting that the domain-invariant information can benefit the performance of time-series domain adaptation.
- However, the improvement is not so notable when we take Guangzhou and Shenzhen as the target domain, this is because the label value ranges of these cities are much lower than the others. And the CNN-based method like CODAT and AdvSKM do not achieve ideal performance, this is because the size of the air quality forecast dataset is small.

2) Results on In-hospital Mortality Prediction Dataset.: Finally, we also testify our method on the MIMIC-III dataset, which is chosen as the benchmark of unsupervised domain adaptation for time-series data in [7]. We use the MIMIC-III dataset for mortality prediction and split the patients into 4 groups according to their ages. Unsupervised domain adaptation on mortality prediction is another practically significant task since the hospital can easily collect the records of the old patient but the data of the young are hard to access.

We choose 12 modalities described in [7]. According to the experiment results shown in Table II, we can learn the following lessons:

- Similar to the other datasets, the SASA-IV overpasses the other comparison models on all the transfer tasks. Some domain adaptation tasks such as $1 \to 4$ and $2 \to 4$ even achieve 4.02 and 6.54 improvement respectively.
- We also find that the transfer task with large age pairs like $1 \to 4$ and $3 \to 1$ also achieve great improvement. This is because the domain discrepancy becomes larger when the age distance becomes bigger, and the proposed SASA-IV method not only aligns the domain-invariant associative structures but also considers the domain-variant information.

3) Results on Boiler Fault Detection.: Since the boilers are usually following stable physical rules, which are naturally considered to be transferred among different domains. Moreover, since boiler fault labeled data are very difficult to collect, it is significant to simultaneously leverage the limited labeled data and the massive unlabeled data collected from different boilers. Hence we take different boilers as different domains and consider boiler Fault detection as the time-series unsupervised domain adaptation problem.

In order to evaluate the performance of the proposed SASA-IV, we first illustrate the experimental results on the Boiler Fault Detection dataset, which are shown in Table III. According to the experiments, we can obtain the following observations:

- Both the SASA and SASA-IV outperform the other baselines with a large margin, which proves the superior transferability. Furthermore, it is worth mentioning that the performance of the SASA-IV is much better that of SASA, which reflects the advantages of the improved strategy of sparse associative structure alignment.
- We also find that the other latest baselines like AdvSKM and CODAT also perform better than the baselines in the conference version like VRADA.
- Similar to the SASA in conference version, SASA-IV promotes the AUC score substantially on tasks, e.g. $1 \to 2$ and $3 \to 2$, which are respectively improved by 4.99 and 6.9. On other easy tasks like $1 \to 3$ and $2 \to 3$, our method still achieves comparable results.

F. Ablation Study and Visualization

1) The study of the effectiveness of different model variant: In this subsection, we provide the results of different model variants to verify the effectiveness of our method. The experiment results on each dataset are respectively shown in Table IV, V, VI. According to these experiment results, we can learn the following lessons:

- Compared with SASA-IV and SASA-IV-α, we can find that the performance of SASA-IV-α drops, this is because

TABLE I
RMSE ON AIR QUALITY PREDICTION.

Method	B→T	G→T	S→T	T→B	G→B	S→B	B→G	T→G	S→G	B→S	T→S	G→S	Avg
LSTM_S2T	40.89	42.20	49.21	52.18	56.55	70.52	19.12	19.37	17.53	13.82	13.80	14.17	34.11
RDC	39.31	40.36	47.75	51.99	56.77	69.40	18.84	19.28	15.56	13.67	13.56	13.91	33.37
R-DANN	41.49	39.85	46.76	52.69	54.80	68.92	18.11	18.95	15.14	13.76	13.86	13.83	33.18
VRADA	38.56	39.12	46.12	52.74	54.57	65.53	17.84	18.55	14.75	13.84	14.22	13.85	32.50
AdvSKM	40.21	39.23	46.74	47.14	55.79	62.74	18.63	19.45	17.26	14.06	16.95	13.67	32.65
CODAT	38.47	38.70	48.19	48.17	55.17	57.65	17.86	18.38	17.61	14.76	17.58	14.16	32.29
SASA	35.52	34.44	40.74	49.41	53.98	57.42	16.45	15.84	14.27	13.52	13.47	**13.48**	29.88
SASA-IV	**35.14**	**33.76**	**40.51**	**46.28**	**53.06**	**55.88**	**15.26**	**15.01**	**14.01**	**13.00**	**13.01**	13.50	**29.03**

TABLE II
AUC SCORE(%) ON IN-HOSPITAL MORTALITY PREDICTION.

Method	2→1	3→1	4→1	1→2	3→2	4→2	1→3	2→3	4→3	1→4	2→4	3→4	Avg
LSTM_S2T	80.11	78.09	76.91	80.22	81.26	76.09	75.73	79.21	75.07	65.35	60.07	69.15	75.52
RDC	80.96	78.32	77.18	80.28	82.63	77.36	76.04	79.90	75.52	65.73	69.16	70.75	76.15
R-DANN	80.88	79.57	77.35	80.41	82.14	78.24	75.93	79.01	75.80	66.55	69.52	69.49	76.24
VRADA	80.94	80.81	77.08	81.52	83.09	78.25	75.57	79.24	75.67	68.14	69.23	69.94	76.62
AdvSKM	83.03	81.90	78.08	80.52	83.42	77.08	75.61	79.38	76.40	65.51	70.03	70.93	76.83
CODAT	81.11	78.09	76.46	80.24	83.00	77.23	75.98	79.09	75.95	66.14	69.81	72.31	76.28
SASA	84.21	82.68	80.20	83.14	84.22	81.87	77.41	80.56	78.58	70.84	72.04	73.20	79.08
SASA-IV	**85.80**	**86.25**	**81.32**	**83.48**	**84.74**	**82.42**	**78.67**	**80.43**	**79.04**	**74.86**	**78.58**	**79.03**	**81.22**

TABLE III
AUC SCORE(%) ON BOILER FAULT DETECTION.

Method	1→2	1→3	3→1	3→2	2→1	2→3	Avg
LSTM_S2T	67.04	94.50	93.23	56.06	84.71	91.14	81.17
RDC	67.17	94.65	93.17	57.30	85.59	92.45	81.71
R-DANN	67.26	94.88	93.57	58.14	85.54	92.41	81.97
VRADA	67.38	94.69	93.58	58.89	84.78	92.52	81.97
AdvSKM	68.68	94.99	93.25	59.35	86.51	91.14	82.32
CODAT	68.62	94.75	93.31	58.38	86.03	91.14	82.04
SASA	71.56	95.39	93.33	61.90	88.04	93.16	83.89
SASA-IV	**73.61**	**95.64**	**93.77**	**65.28**	**92.37**	**94.75**	**85.90**

$\mathcal{L}_\alpha$ can restrict the common segment length, so the sideeffect of different time lags will be removed.

- Compared with SASA-IV and SASA-IV-β, we can also find that the performance of SASA-IV-β drops. These experiment results indirectly reflect that the sparse associative structures vary with different domains and aligning the sparse associative structures can avoid the sideeffect of domain-specific association. We also find that SASA-IV-β still performs better than most of the baselines, reflecting that the extracted sparse associative structures can remove most of the redundant relationships and make the model robust.
- we also explore the effectiveness of domain-variant information. Compared with SASA-IV and SASA-IV-γ, we can find that the performance of SASA-IV-γ degenerates. For one thing, these experiment results indirectly show that the strengths of associative structures vary with different domains and for another, our SASA-IV can extract and leverage the domain-variant information to achieve more robust performance.
- In order to evaluate the effectiveness of the proposed unidirectional alignment restriction, we devise the SASA-IV-$\mathcal{C}$. According to the experiment results of SASA-IV-$\mathcal{C}$ in different datasets, we can find that: 1) The performance of SASA-IV-$\mathcal{C}$ is comparable with that of SASA on the air quality forecast datasets and better than SASA on the other two datasets, reflecting that the strategy of seperating structures and strengths is more superior than the alignment of weighted associative structures. 2) Compared with SASA-IV and SASA-IV-$\mathcal{C}$, we can find that the performance of SASA-IV-$\mathcal{C}$ is lower than that of SASA-IV, this circumstance indirectly proves that the bidirectional alignment will result in the wrong associative structure discovery, which further leads to the suboptimal performance. With the help of unidirectional alignment restriction, we can address this issue.
- We also consider the ablation experiments of SASA. Compared the result of SASA and SASA-α, we can find that the performance of SASA-α drops. This is because of α represents the probability of the length of a segment. And the duration of segments varies with different domains. With the restriction of α, we can exclude the influence of domain-specific segments duration.
- According to the experiment results of SASA-β, we can

TABLE IV
RMSE ON AIR QUALITY PREDICTION FOR ABLATION STUDY.

Method	B→T	G→T	S→T	T→B	G→B	S→B	B→G	T→G	S→G	B→S	T→S	G→S	Avg
SASA-α	36.61	34.71	41.73	49.92	54.76	58.53	17.38	16.36	14.70	13.87	13.90	13.91	30.53
SASA-β	36.51	34.96	41.32	49.93	55.10	58.33	17.08	16.40	14.67	13.89	13.90	13.88	30.50
SASA-IV-α	37.60	34.84	41.67	48.06	56.83	59.21	16.30	15.92	14.37	14.09	14.11	13.99	30.58
SASA-IV-β	36.47	35.85	42.40	47.26	54.42	59.56	16.78	16.60	14.63	13.88	13.85	13.80	30.46
SASA-IV-γ	38.58	37.21	41.82	47.77	55.23	57.54	16.60	16.11	14.50	14.19	14.21	13.67	30.63
SASA-IV-C	36.53	34.88	41.24	47.18	55.24	57.87	15.99	15.85	14.54	13.83	13.87	13.87	30.07
SASA-IV	**35.14**	**33.76**	**40.51**	**46.28**	**53.06**	**55.88**	**15.26**	**15.01**	**14.01**	**13.00**	**13.01**	**13.50**	**29.03**

TABLE V
AUC SCORE(%) ON IN-HOSPITAL MORTALITY PREDICTION FOR ABLATION STUDY.

Method	2→1	3→1	4→1	1→2	3→2	4→2	1→3	2→3	4→3	1→4	2→4	3→4	Avg
SASA-α	83.95	81.92	79.85	82.96	83.94	80.99	77.05	80.11	78.33	68.65	70.62	72.23	78.38
SASA-β	83.93	81.47	78.82	81.88	83.70	80.98	76.73	79.75	77.98	68.23	70.35	74.98	78.23
SASA-IV-α	84.44	84.96	80.10	81.59	82.66	81.08	77.85	79.66	78.18	73.99	77.40	77.57	79.96
SASA-IV-β	84.96	85.27	80.29	82.75	83.13	81.63	77.27	79.18	78.14	73.90	76.60	77.15	80.02
SASA-IV-γ	84.95	85.63	80.52	82.72	84.18	81.27	77.57	79.93	78.02	74.28	76.73	76.22	80.17
SASA-IV-C	84.91	84.93	80.67	82.65	84.73	81.47	78.08	80.35	78.64	74.58	78.00	77.92	80.58
SASA-IV	**85.80**	**86.25**	**81.32**	**83.48**	**84.74**	**82.42**	**78.67**	**80.43**	**79.04**	**74.86**	**78.58**	**79.03**	**81.22**

TABLE VI
AUC SCORE(%) ON BOILER FAULT DETECTION FOR ABLATION STUDY.

Method	1→2	1→3	3→1	3→2	2→1	2→3	Avg
SASA-α	70.62	95.19	92.59	59.79	87.81	92.98	83.16
SASA-β	70.22	94.87	93.01	60.07	87.44	92.77	83.06
SASA-IV-α	71.97	95.07	93.26	61.68	91.48	93.14	84.43
SASA-IV-β	70.43	94.71	93.00	61.74	91.49	93.58	84.16
SASA-IV-γ	72.64	95.45	93.15	61.98	91.76	93.88	84.81
SASA-IV-C	72.05	94.87	92.78	63.69	91.41	93.63	84.74
SASA-IV	**73.61**	**95.64**	**93.77**	**65.28**	**92.37**	**94.75**	**85.90**

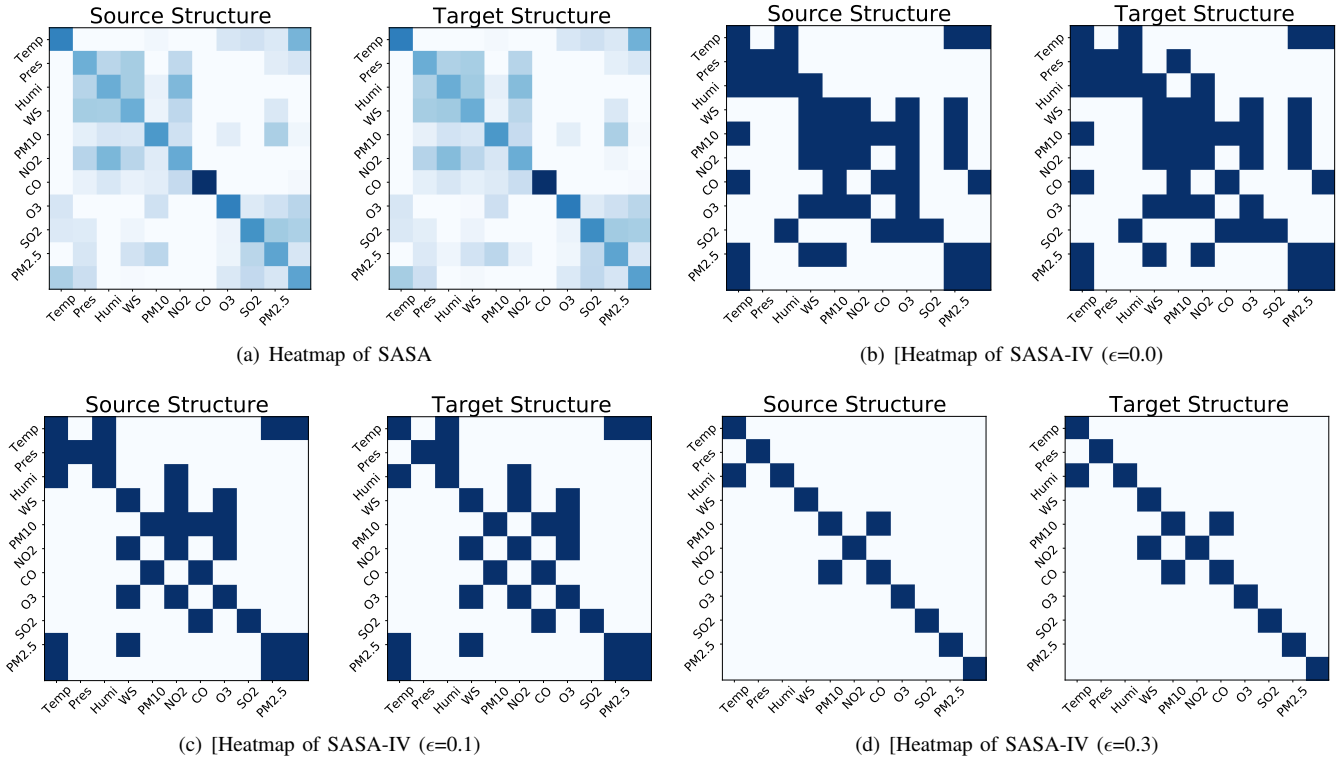
find that the performance of SASA-β is worse that the standard SASA. This is because the sparse associative structure have been extracted, which is also more robust than that of normal feature extractor. But the reserved domain-specific associative relationships lead to the sub-optimal results. Note that the SASA-β is still better than the baselines shown in Table I, II and III. This is because the $\mathcal{L}_\alpha$ aligns the offsets between different domains, which benefits to extracting sparse associative structure for adaptation.

2) Visualization of Aligned Associative Structures.: The SASA in the conference version simply leverages the Sparsemax and automatically generates the sparse associative structures. However, some redundant relationships might remain. In order to address the aforementioned issue, the proposed SASA-IV controls the sparseness of the learned associative structures by controlling the value of ϵ, which is shown in Figure 4. The visualization shows that:

- Figure 4(a) illustrates the heatmap of weighted associative structure of SASA. Deeper the color is , the stronger the relationships is. We can find that the color shades of the heatmap from different domain are almost the same, showing that the original SASA discard the domain-variant strengths.
- The structures from different domains have many shared associative relationships, which reflects the domain-invariant mechanisms.
- The associative structures are very sparse, even when $\epsilon = 0$, reflects that our methods can extract the sparse associative structures.
- The larger ϵ is, the more sparse the associative structures are. Hence our method can control the sparseness of the associative structures and remove more redundant relationships.

VI. CONCLUSION

This paper presents an improved sparse associative structure alignment model for time-series unsupervised domain adaptation. In our proposal, we explore what is invariant and variant in time-series data and provide insights into how to devise the model for time-series unsupervised domain adaptation. Technically, the weights for unweighted sparse associative structure alignment are embedded and the gradient stopping is employed for better common structure discovery. We further take domain-variance information into consideration with the help of autoregressive feature extraction. The success of the proposed sparse associative structure alignment method not only provides an effective and novel solution for time-series domain adaptation but also provides some insightful theorems



(a) Heatmap of SASA

(b) [Heatmap of SASA-IV (ϵ=0.0)

(c) [Heatmap of SASA-IV (ϵ=0.1)

(d) [Heatmap of SASA-IV (ϵ=0.3)

Fig. 4. The illustration of visualization of correlation structure adjacent matrices under different values of ϵ.

and results on what transfer to learn and how to achieve the ideal transfer.

ACKNOWLEDGMENT

The authors would like to thank Zhenjie Zhang and Xiaoyan Yang from the PVoice Technology as well as Zhuozhang Li from the Guangdong University of Technology for their help and support on this work.

REFERENCES

[1] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," in International conference on machine learning. PMLR, 2015, pp. 1180–1189.

[2] E. Tzeng, J. Hoffman, N. Zhang, K. Saenko, and T. Darrell, "Deep domain confusion: Maximizing for domain invariance," arXiv preprint arXiv:1412.3474, 2014.

[3] S. J. Pan and Q. Yang, "A survey on transfer learning," IEEE Transactions on knowledge and data engineering, vol. 22, no. 10, pp. 1345–1359, 2009.

[4] M. Long, H. Zhu, J. Wang, and M. I. Jordan, "Unsupervised domain adaptation with residual transfer networks," arXiv preprint arXiv:1602.04433, 2016.

[5] J. Li, K. Lu, Z. Huang, L. Zhu, and H. T. Shen, "Transfer independently together: A generalized framework for domain adaptation," IEEE Transactions on Cybernetics, vol. 49, no. 6, pp. 2144–2155, 2019.

[6] P. R. d. O. da Costa, A. Akçay, Y. Zhang, and U. Kaymak, "Remaining useful lifetime prediction via deep domain adaptation," Reliability Engineering & System Safety, vol. 195, p. 106682, 2020.

[7] S. Purushotham, W. Carvalho, T. Nilanon, and Y. Liu, "Variational recurrent adversarial deep domain adaptation," in 5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings. OpenReview.net, 2017. [Online]. Available: https://openreview.net/forum?id=rk9eAFcxg

[8] Q. Liu and H. Xue, "Adversarial spectral kernel matching for unsupervised time series domain adaptation," in Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21, Z.-H. Zhou, Ed. International Joint Conferences on Artificial Intelligence Organization, 8 2021, pp. 2744–2750, main Track. [Online]. Available: https://doi.org/10.24963/ijcai.2021/378

[9] T. Mikolov, M. Karafiát, L. Burget, J. Černockỳ, and S. Khudanpur, "Recurrent neural network based language model," in Eleventh annual conference of the international speech communication association, 2010.

[10] J. Chung, K. Kastner, L. Dinh, K. Goel, A. C. Courville, and Y. Bengio, "A recurrent latent variable model for sequential data," in Advances in neural information processing systems, 2015, pp. 2980–2988.

[11] R. Cai, J. Chen, Z. Li, W. Chen, K. Zhang, J. Ye, Z. Li, X. Yang, and Z. Zhang, "Time series domain adaptation via sparse associative structure alignment," Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, no. 8, pp. 6859–6867, May 2021. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/16846

[12] Z. Wang, B. Du, and Y. Guo, "Domain adaptation with neural embedding matching," IEEE transactions on neural networks and learning systems, vol. 31, no. 7, pp. 2387–2397, 2019.

[13] F. Mahmood, R. Chen, and N. J. Durr, "Unsupervised reverse domain adaptation for synthetic medical images via adversarial training," IEEE transactions on medical imaging, vol. 37, no. 12, pp. 2572–2581, 2018.

[14] A. Ramponi and B. Plank, "Neural unsupervised domain adaptation in nlp—a survey," arXiv preprint arXiv:2006.00632, 2020.

[15] X. Glorot, A. Bordes, and Y. Bengio, "Domain adaptation for large-scale sentiment classification: A deep learning approach," in ICML, 2011.

[16] Y. Zhang, P. David, and B. Gong, "Curriculum domain adaptation for semantic segmentation of urban scenes," in Proceedings of the IEEE international conference on computer vision, 2017, pp. 2020–2030.

[17] Y. Zou, Z. Yu, B. Kumar, and J. Wang, "Unsupervised domain adaptation for semantic segmentation via class-balanced self-training," in Proceedings of the European conference on computer vision (ECCV), 2018, pp. 289–305.

[18] Z. Hao, D. Lv, Z. Li, R. Cai, W. Wen, and B. Xu, "Semi-supervised disentangled framework for transferable named entity recognition," Neural Networks, vol. 135, pp. 127–138, 2021.

[19] Z. Cui, W. Li, D. Xu, S. Shan, X. Chen, and X. Li, "Flowing on riemannian manifold: Domain adaptation by shifting covariance," IEEE Transactions on Cybernetics, vol. 44, no. 12, pp. 2264–2273, 2014.

[20] B. Schölkopf, F. Locatello, S. Bauer, N. R. Ke, N. Kalchbrenner, A. Goyal, and Y. Bengio, "Toward causal representation learning," Proceedings of the IEEE, vol. 109, no. 5, pp. 612–634, 2021.

[21] R. Cai, Z. Li, P. Wei, J. Qiao, K. Zhang, and Z. Hao, "Learning disentangled semantic representation for domain adaptation," in IJCAI: proceedings of the conference, vol. 2019. NIH Public Access, 2019, p. 2060.

[22] K. Zhang, M. Gong, P. Stojanov, B. Huang, Q. Liu, and C. Glymour, "Domain adaptation as a problem of inference on graphical models," arXiv preprint arXiv:2002.03278, 2020.

[23] K. Zhang, B. Schölkopf, K. Muandet, and Z. Wang, "Domain adaptation under target and conditional shift," in International Conference on Machine Learning. PMLR, 2013, pp. 819–827.

[24] P. Stojanov, Z. Li, M. Gong, R. Cai, J. Carbonell, and K. Zhang, "Domain adaptation with invariant representation learning: What transformations to learn?" Advances in Neural Information Processing Systems, vol. 34, 2021.

[25] X. Chen, S. Wang, J. Wang, and M. Long, "Representation subspace distance for domain adaptation regression," in International Conference on Machine Learning. PMLR, 2021, pp. 1749–1759.

[26] M. Long, Y. Cao, Z. Cao, J. Wang, and M. I. Jordan, "Transferable representation learning with deep adaptation networks," IEEE transactions on pattern analysis and machine intelligence, vol. 41, no. 12, pp. 3071–3085, 2018.

[27] Y. Chen, S. Song, S. Li, L. Yang, and C. Wu, "Domain space transfer extreme learning machine for domain adaptation," IEEE Transactions on Cybernetics, vol. 49, no. 5, pp. 1909–1922, 2019.

[28] M. Long, Y. Cao, J. Wang, and M. Jordan, "Learning transferable features with deep adaptation networks," in International conference on machine learning. PMLR, 2015, pp. 97–105.

[29] H. Yan, Y. Ding, P. Li, Q. Wang, Y. Xu, and W. Zuo, "Mind the class weight bias: Weighted maximum mean discrepancy for unsupervised domain adaptation," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 2017.

[30] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," Advances in neural information processing systems, vol. 27, 2014.

[31] S. Xie, Z. Zheng, L. Chen, and C. Chen, "Learning semantic representations for unsupervised domain adaptation," in International Conference on Machine Learning, 2018, pp. 5419–5428.

[32] X. Gu, J. Sun, and Z. Xu, "Spherical space domain adaptation with robust pseudo-label loss," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 9101–9110.

[33] Y. Zhang, T. Liu, M. Long, and M. Jordan, "Bridging theory and algorithm for domain adaptation," in International Conference on Machine Learning. PMLR, 2019, pp. 7404–7413.

[34] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," arXiv preprint arXiv:1312.6114, 2013.

[35] C.-X. Ren, X.-L. Xu, and H. Yan, "Generalized conditional domain adaptation: A causal perspective with low-rank translators," IEEE Transactions on Cybernetics, vol. 50, no. 2, pp. 821–834, 2020.

[36] G. Wilson, J. R. Doppa, and D. J. Cook, "Multi-source deep domain adaptation with weak supervision for time-series sensor data," in Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2020, pp. 1768–1778.

[37] M. Ragab, E. Eldele, Z. Chen, M. Wu, C.-K. Kwoh, and X. Li, "Self-supervised autoregressive domain adaptation for time series data," arXiv preprint arXiv:2111.14834, 2021.

[38] C. Kemp and J. B. Tenenbaum, "The discovery of structural form," Proceedings of the National Academy of Sciences, vol. 105, no. 31, pp. 10 687–10 692, 2008.

[39] S. Džeroski, L. De Raedt, and K. Driessens, "Relational reinforcement learning," Machine learning, vol. 43, no. 1, pp. 7–52, 2001.

[40] D. Koller, N. Friedman, S. Džeroski, C. Sutton, A. McCallum, A. Pfeffer, P. Abbeel, M.-F. Wong, C. Meek, J. Neville et al., Introduction to statistical relational learning. MIT press, 2007.

[41] D. Cao, Y. Wang, J. Duan, C. Zhang, X. Zhu, C. Huang, Y. Tong, B. Xu, J. Bai, J. Tong et al., "Spectral temporal graph neural network for multivariate time-series forecasting," Advances in Neural Information Processing Systems, vol. 33, pp. 17 766–17 778, 2020.

[42] J. Li, F. Yang, M. Tomizuka, and C. Choi, "Evolvegraph: Multi-agent trajectory prediction with dynamic relational reasoning," Advances in neural information processing systems, vol. 33, pp. 19 783–19 794, 2020.

[43] H. Fan, F. Zhang, and Y. Gao, "Self-supervised time series representation learning by inter-intra relational reasoning," arXiv preprint arXiv:2011.13548, 2020.

[44] T. Kipf, E. Fetaya, K.-C. Wang, M. Welling, and R. Zemel, "Neural relational inference for interacting systems," in International Conference on Machine Learning. PMLR, 2018, pp. 2688–2697.

[45] A. Goyal, A. Lamb, J. Hoffmann, S. Sodhani, S. Levine, Y. Bengio, and B. Schölkopf, "Recurrent independent mechanisms," arXiv preprint arXiv:1909.10893, 2019.

[46] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in Advances in neural information processing systems, 2017, pp. 5998–6008.

[47] A. Martins and R. Astudillo, "From softmax to sparsemax: A sparse model of attention and multi-label classification," in International conference on machine learning. PMLR, 2016, pp. 1614–1623.

[48] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings, Y. Bengio and Y. LeCun, Eds., 2015. [Online]. Available: http://arxiv.org/abs/1409.0473

[49] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Lio, and Y. Bengio, "Graph attention networks," arXiv preprint arXiv:1710.10903, 2017.

[50] X. Wang, H. Ji, C. Shi, B. Wang, Y. Ye, P. Cui, and P. S. Yu, "Heterogeneous graph attention network," in The World Wide Web Conference, 2019, pp. 2022–2032.

[51] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," arXiv preprint arXiv:1609.02907, 2016.

[52] M. Mohri, A. Rostamizadeh, and A. Talwalkar, Foundations of machine learning. MIT press, 2018.

[53] Y. Mansour, M. Mohri, and A. Rostamizadeh, "Domain adaptation: Learning bounds and algorithms," arXiv preprint arXiv:0902.3430, 2009.

[54] C. Cortes and M. Mohri, "Domain adaptation in regression," in International Conference on Algorithmic Learning Theory. Springer, 2011, pp. 308–323.

[55] S. Ben-David, J. Blitzer, K. Crammer, F. Pereira et al., "Analysis of representations for domain adaptation," Advances in neural information processing systems, vol. 19, p. 137, 2007.

[56] Y. Zheng, X. Yi, M. Li, R. Li, Z. Shan, E. Chang, and T. Li, "Forecasting fine-grained air quality based on big data," in Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2015, pp. 2267–2276.

[57] A. E. Johnson, T. J. Pollard, L. Shen, H. L. Li-wei, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. A. Celi, and R. G. Mark, "Mimic-iii, a freely accessible critical care database," Scientific data, vol. 3, p. 160035, 2016.

[58] Z. Che, S. Purushotham, K. Cho, D. Sontag, and Y. Liu, "Recurrent neural networks for multivariate time series with missing values," Scientific reports, vol. 8, no. 1, pp. 1–12, 2018.

[59] S. Purushotham, C. Meng, Z. Che, and Y. Liu, "Benchmarking deep learning models on large healthcare datasets," Journal of biomedical informatics, vol. 83, pp. 112–134, 2018.