# Disentangled Representation Learning for Causal Inference With Instruments

Debo Cheng, Jiuyong Li, *Member, IEEE*, Lin Liu, Ziqi Xu, Weijia Zhang, Jixue Liu, and Thuc Duy Le

***Abstract*— Latent confounders are a fundamental challenge for inferring causal effects from observational data. The instrumental variable (IV) approach is a practical way to address this challenge. Existing IV-based estimators need a known IV or other strong assumptions, such as the existence of two or more IVs in the system, which limits the application of the IV approach. In this article, we consider a relaxed requirement, which assumes there is an IV proxy in the system without knowing which variable is the proxy. We propose a variational autoencoder (VAE)-based disentangled representation learning method to learn an IV representation from a dataset with latent confounders and then utilize the IV representation to obtain an unbiased estimation of the causal effect from the data. Extensive experiments on synthetic and real-world data have demonstrated that the proposed algorithm outperforms the existing IV-based estimators and VAE-based estimators.**

***Index Terms*— Causal inference, disentangled representation learning, instrumental variable (IV), latent variables, observational data.**

## I. Introduction

ESTIMATING the causal effect of a treatment (also known as intervention, or exposure) on an outcome is a fundamental task in many areas [1], [2], such as policy-making and new drug evaluation. Randomized controlled trials (RCTs) are the gold standard for inferring causal effects, but they are often impractical in real-world applications due to time or ethical constraints [3]. Thus, causal effect estimation with observational data has become an alternative to RCTs.

However, estimating causal effects using observational data suffers from confounding bias, due to the spurious association caused by confounders that affect both the treatment and outcome variables. Unmeasured confounders make the situation even worse [4], [5], [6]. As shown in Fig. 1(a), if there is an unmeasured confounder ($U$) between the treatment ($W$) and the outcome ($Y$), the causal effect of $W$ on $Y$ cannot be estimated with observational data except there is an instrumental variable (IV) [1], [7], [8].

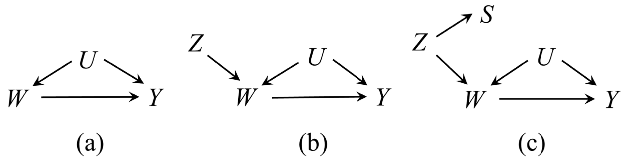


Fig. 1. Causal graphs with latent variables to show the problem of causal effect estimation from observational data. (a) In DAG, the causal effect of $W$ on $Y$ cannot be estimated from observational data. (b) In DAG, there is a valid IV $Z$. (c) In DAG, $Z$ is an unmeasured IV and $S$ is an SIV of $Z$. The causal effect of $W$ on $Y$ in both DAGs [(b) and (c)] can be recovered from observational data.

The IV approach is a commonly used way to estimate causal effects from data when the unconfoundedness assumption is violated [1], [9]. A valid IV (denoted as $Z$) must satisfy the following three conditions [7], [10], [11]: 1) $Z$ influences the treatment (i.e., *relevance condition*); 2) the causal effect of $Z$ on $Y$ is only through $W$ (also known as *exclusion restriction*); and 3) $Z$ and $Y$ do not have any common causes (i.e., *unconfounded instrument*). The three conditions of a valid IV can only be verified using domain knowledge or based on the underlying causal graph of the system but not from data [12]. It is known that domain knowledge of an IV or a causal graph is rarely available in many real applications [7]. Therefore, it is desirable to explore an effective data-driven method to discover a valid IV directly from data.

A few data-driven IV-based methods have been proposed for causal effect estimation without assuming a known IV, but they often have other constraints. For example, IV.Tetrad [13] requires that at least a pair of IVs exist in the system and the set of all the remaining variables excluding the pair of IVs is a conditioning set of the IV with respect to the treatment and the outcome. sisVIVE [14] requires that at least half of the variables (the set of candidate IVs) are valid IVs.

Some research has proposed the necessary conditions of IVs for obtaining a bound estimation of a causal effect, i.e., a multiset of possible estimates from data, instead of a unique estimate. For instance, Pearl [12] proposed an instrumental inequality to find a set of candidate IVs from data with discrete variables, and Kuroki and Cai [15] extended the instrumental inequality to linear structural equation models with continuous variables. Assuming a linear non-Gaussian acyclic causal model, Xie et al. [16] proposed a necessary condition based on a generalized independent noise condition for identifying continuous variables as the candidate valid IVs, but the condition only produces a bound estimation.

Received 6 March 2023; revised 13 October 2023 and 26 July 2024; accepted 3 December 2024. This work was supported by the Australian Research Council under Grant DP230101122. *(Corresponding author: Debo Cheng.)*

Debo Cheng, Jiuyong Li, Lin Liu, Jixue Liu, and Thuc Duy Le are with UniSA STEM, University of South Australia, Adelaide, SA 5095, Australia (e-mail: chedy055@mymail.unisa.edu.au; Jiuyong.Li@unisa.edu.au; Lin.Liu@unisa.edu.au; Jixue.Liu@unisa.edu.au; Thuc.Le@unisa.edu.au).

Ziqi Xu is with the School of Computing Technologies, RMIT University, Melbourne, VIC 3000, Australia (e-mail: ziqi.xu@rmit.edu.au).

Weijia Zhang is with the School of Information and Physical Sciences, University of Newcastle, Callaghan, NSW 2308, Australia (e-mail: weijia.zhang@newcastle.edu.au).

Digital Object Identifier 10.1109/TNNLS.2024.3512790

Therefore, the existing data-driven IV methods either rely on strong assumptions or only provide a necessary condition for determining candidate IVs. In order to develop a more effective and practical data-driven IV-based causal effect estimator, in this article, we consider a relaxed requirement, which assumes there exists at least one IV proxy (also known as surrogate IV, i.e., SIV) in the system without knowing which variable is the proxy. The assumption of an IV proxy is practical since there exist many proxy variables for latent confounders [17], [18].

It is challenging to determine IVs (or SIVs) from the set of measured covariates since IVs (or SIVs) and measured confounders are statistically inseparable. To address this challenge, we propose a data-driven method, disentangled IV based on variational autoencoder (DIV.VAE), using disentangling techniques [19], [20] to learn the latent representation $\mathbf{\Phi}$ of the set of pretreatment variables $\mathbf{X}$, which are measured before applying the treatment $W$ and observing the outcome $Y$ [8], and disentangle $\mathbf{\Phi}$ into $(\mathbf{Z}, \mathbf{C})$, where $\mathbf{Z}$ represents IV and $\mathbf{C}$ represents confounders in the latent space. To the best of authors' knowledge, DIV.VAE is the first work using the VAE model to infer the IV representation from observed pretreatment variables when the unconfoundedness assumption is not satisfied.

Our main contributions are summarized as follows.

1) We address the challenging problem of causal effect estimation from data in the presence of latent confounders.
2) We propose a novel disentangled representation learning method, DIV.VAE, to learn the latent IV representation and the latent confounding representation for achieving unbiased causal effect estimation.
3) We empirically evaluate the effectiveness of DIV.VAE on synthetic and real-world datasets, in comparison with the state-of-the-art causal effect estimators. The results show that the DIV.VAE outperforms baseline estimators.

## II. Preliminaries

### A. Notations

We represent variables and their values with uppercase and lowercase letters, respectively. A set of variables and a value assignment of the set are denoted by bold-faced uppercase and lowercase letters, respectively.

Let $\mathcal{G} = (\mathbf{V}, \mathbf{E})$ be a directed acyclic graph (DAG), where $\mathbf{V} = \{V_1, \dots, V_p\}$ are the set of nodes representing $p$ random variables and $\mathbf{E} \subseteq \mathbf{V} \times \mathbf{V}$ are the set of edges representing the relationships between nodes. In DAG $\mathcal{G}$, two nodes are *adjacent* when there exists a directed edge → between them. A path $\pi$ from $V_i$ to $V_j$ is a directed or causal path if all edges along it are directed toward $V_j$. If there is a directed path $\pi$ from $V_i$ to $V_j$, $V_i$ is known as an ancestor of $V_j$ and $V_j$ is a descendant of $V_i$. The sets of ancestors and descendants of a node $V$ are denoted as $\text{An}(V)$ and $\text{De}(V)$, respectively.

A DAG is causal if the directed edge $V_i \to V_j$ between $V_i$ and $V_j$ indicates that $V_i$ is a direct cause of $V_j$. In a DAG $\mathcal{G}$, a path $\pi$ between $V_i$ and $V_j$ comprises a sequence of distinct nodes $\langle V_i, \dots, V_j \rangle$ with every pair of successive nodes being adjacent, and $V_i$ and $V_j$ are end nodes of $\pi$.

The definitions of Markov property and faithfulness are introduced in the following.

*Definition 1 (Markov Property [1]):* Given a DAG $\mathcal{G} = (\mathbf{V}, \mathbf{E})$ and the joint probability distribution of $\mathbf{V}$ $(P(\mathbf{V}))$, $\mathcal{G}$ satisfies the Markov property; if for $\forall V_i \in \mathbf{V}$, $V_i$ is probabilistically independent of all of its nondescendants, given the parent nodes of $V_i$.

*Definition 2 (Faithfulness [4]):* A DAG $\mathcal{G} = (\mathbf{V}, \mathbf{E})$ is faithful to a joint distribution $P(\mathbf{V})$ over the set of variables $\mathbf{V}$ if and only if every independence present in $P(\mathbf{V})$ is entailed by $\mathcal{G}$ and satisfies the Markov property. A joint distribution $P(\mathbf{V})$ over the set of variables $\mathbf{V}$ is faithful to the DAG $\mathcal{G}$ if and only if the DAG $\mathcal{G}$ is faithful to the joint distribution $P(\mathbf{V})$.

When the faithfulness assumption is satisfied between a joint distribution $P(\mathbf{V})$ and a DAG of a set of variables $\mathbf{V}$, the dependence/independence relations among the variables can be read from the DAG [1], [4]. In a DAG, d-separation is a well-known graphical criterion that is used to read off the conditional independence between variables entailed in the DAG when the Markov property and faithfulness are satisfied [1], [4].

*Definition 3 (d-Separation [1]):* A path $\pi$ in a DAG $\mathcal{G} = (\mathbf{V}, \mathbf{E})$ is said to be d-separated (or blocked) by a set of nodes $\mathbf{M}$ if and only if: 1) $\pi$ contains a chain $V_i \to V_k \to V_j$ or a fork $V_i \leftarrow V_k \to V_j$ such that the middle node $V_k$ is in $\mathbf{M}$ or 2) $\pi$ contains a collider $V_k$ such that $V_k$ is not in $\mathbf{M}$ and no descendant of $V_k$ is in $\mathbf{M}$. A set $\mathbf{M}$ is said to d-separate $V_i$ from $V_j$ $(V_i \perp\!\!\!\perp V_j | \mathbf{M})$ if and only if $\mathbf{M}$ blocks every path between $V_i$ and $V_j$. Otherwise, they are said to be d-connected by $\mathbf{M}$, denoted as $V_i \not\perp\!\!\!\perp V_j | \mathbf{M}$.

The back-door criterion is a well-known graphical criterion for determining an adjustment set in a given DAG $\mathcal{G}$. The back-door criterion can be used directly to find an adjustment set $\mathbf{M} \subseteq \mathbf{X}$ relative to an ordered pair of variables $(V_i, V_j)$ in the given $\mathcal{G}$.

*Definition 4 (Back-Door Criterion [1]):* In a DAG $\mathcal{G} = (\mathbf{V}, \mathbf{E})$, for an ordered pair of variables $(V_i, V_j) \in \mathbf{V}$, a set of variables $\mathbf{M} \subseteq \mathbf{V} \setminus \{V_i, V_j\}$ is said to satisfy the back-door criterion in the given DAG $\mathcal{G}$ if: 1) $\mathbf{M}$ does not contain a descendant node of $V_i$ and 2) $\mathbf{M}$ blocks every back-door path between $V_i$ and $V_j$ (the paths between $V_i$ and $V_j$ starting with an arrow into $V_i$). A set $\mathbf{M}$ is referred to as a *back-door set* relative to $(V_i, V_j)$ in $\mathcal{G}$ if $\mathbf{M}$ satisfies the back-door criterion relative to $(V_i, V_j)$ in $\mathcal{G}$.

### B. Instrumental Variables

We follow the convention and definitions of IVs used in [7], [13], and [21]. We assume a causal DAG $\mathcal{G}$ with the set of variables $\mathbf{V} = \mathbf{X} \cup \mathbf{U} \cup \{W, Y\}$, where $W$ is a binary treatment indicator ($w = 1$ for being treated and $w = 0$ for control), $Y$ is the outcome of interest, $\mathbf{X}$ is a set of pretreatment variables,[1] i.e., $\forall X \in \mathbf{X}$, $X \notin \text{De}(W \cup Y)$ where $\text{De}(W \cup Y)$ is a shorthand of $\text{De}(W) \cup \text{De}(Y)$, and $\mathbf{U}$ is the set of latent confounders. The

[1] A variable is measured before applying $W$ and observing $Y$ in a study or experiment. Pretreatment variables can be distinguished from other variables in a real-world application by domain experts [8].

goal of this work is to estimate the average causal effect of $W$ on $Y$ from observational data with latent confounders.

A valid IV facilitates the identification of the causal effect of $W$ on $Y$ from data with latent confounders [9], [12]. A valid IV $Z$ [7], [10] satisfies the three conditions as described in the Introduction. Given a valid IV $Z$, the causal effect of $W$ on $Y$ (referred to as $\beta_{wy}$) can be calculated as $\sigma_{zy}/\sigma_{zw}$, where $\sigma_{zy}$ and $\sigma_{zw}$ are the estimated causal effects of $Z$ on $Y$ and $Z$ on $W$, respectively.

In this work, we employ the orthogonal IVs approach (Ortho.IV) [22], [23] to calculate $\beta_{wy}$ from data with latent confounders when an IV $Z$ is available. The Ortho.IV is to optimize the minimization problem of a loss function that satisfies a Neyman orthogonality criterion with the aid of a known IV by solving the moment equation [22]: $\mathbb{E}[(Y - \mathbb{E}[Y|\mathbf{X}] - \theta(\mathbf{X}) \times (W - \mathbb{E}[W|\mathbf{X}]))(Z - \mathbb{E}[Z|\mathbf{X}])] = 0$, where $\theta(.)$ is a function of $\mathbf{X}$. Once we have a valid IV, the Ortho.IV method can be used to obtain $\beta_{wy}$ unbiasedly from observational data with latent variables.

An IV, such as $Z$ in $Z \rightarrow W$ in Fig. 1(b), is often unmeasured in many real-world applications. An effect variable of an IV, such as $S$ in Fig. 1(c), is more likely to be measured in real-world cases and is called an SIV [7], [24], [25].

*Definition 5 SIV]:* In a causal DAG $\mathcal{G} = (\mathbf{X} \cup \mathbf{U} \cup \{W, Y\}, \mathbf{E})$, a variable $S \in \mathbf{X}$ is said to be an SIV with respect to $W \rightarrow Y$, if: 1) $S$ and $W$ share a latent IV $Z$ (i.e., $S \leftarrow Z \rightarrow W$); 2) $S$ and $Y$ are associated only through $W$ (i.e., exclusion restriction); and 3) $S$ does not share common causes with $Y$ (i.e., unconfoundedness instrument).

An SIV is a proxy variable of a standard IV [7], [25]. An SIV can be used as a valid IV. However, in practice, it is often difficult to identify which variable is a valid IV (standard IV or SIV) even if it is measured.

### C. Problem Setup

We assume that the data are generated from a causal DAG $\mathcal{G} = (\mathbf{X} \cup \mathbf{U} \cup \{W, Y\}, \mathbf{E})$ containing the treatment variable $W$, the outcome variable $Y$, a set of pretreatment variables $\mathbf{X}$, and a set of latent confounders $\mathbf{U}$. There exists at least one SIV in $\mathbf{X}$, but we do not know which $X$s are SIV(s). Or the SIV information is embedded in a number of $X$s. We will query the causal effect of $W$ on $Y$, i.e., $\beta_{wy}$ from observational data. We allow the existence of multiple back-door paths between $W$ and $Y$ with some latent variables in $\mathbf{U}$ lying on some of these back-door paths between $W$ and $Y$.

We do not assume that an IV or an SIV is known, and it is difficult to identify an IV or an SIV from data. For example, assume that the causal DAG in Fig. 2 represents a data generation mechanism. $\mathbf{S}$ is a set of SIVs. In the DAG, we can infer that $\mathbf{S}$ is independent of each variable in $\{\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_Y, \mathbf{X}_C\}$. In data, we can obtain a set of independent pairs $(\mathbf{S}, \mathbf{X}_1)$, $(\mathbf{S}, \mathbf{X}_2)$, $(\mathbf{S}, \mathbf{X}_C)$, $(\mathbf{X}_C, \mathbf{X}_1)$, $(\mathbf{X}_C, \mathbf{X}_2)$, $(\mathbf{X}_1, \mathbf{X}_2)$. This does not help us to find which variable is an SIV. One may wish to learn a partial ancestral graph (PAG) [26] from data for determining the set of SIVs $\mathbf{S}$. However, two variables in the above pair (e.g., $\mathbf{S}$ and $\mathbf{X}_Y$) cannot be separated since the spurious association caused by the latent confounder $U$ [10], [12]. Hence, a causal discovery algorithm using conditional independence tests such as fast causal inference (FCI) [4] does not help us find an SIV from data. Even worse, the IV information might be scattered in other variables.

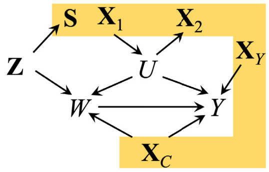


Fig. 2. Example causal DAG representing the data generation mechanism. The shaded area indicates all the measured pretreatment variables, and among them, $\mathbf{S}$ is a set of SIVs, $\mathbf{Z}$ is a set of latent IVs, and $U$ is a latent confounder affecting both $W$ and $Y$.

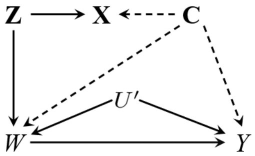


Fig. 3. Disentanglement scheme of DIV.VAE, represented as a causal graph. The dotted arrows indicate possible ancestral relationships between nodes. $W$, $Y$, and $U'$ are the treatment variable, the outcome, and the latent confounder of $W$ and $Y$, respectively. $\mathbf{X}$ is the set of measured pretreatment variables and contains at least one SIV. $\mathbf{\Phi} = (\mathbf{Z}, \mathbf{C})$ is the latent representation of $\mathbf{X}$, where $\mathbf{Z}$ and $\mathbf{C}$ are the sets of disentangled IV representation and confounding representation, respectively.

We will use the disentangling techniques [19], [20] to learn the latent IV representation through disentangling the latent representation of $\mathbf{X}$. We aim at learning a latent representation $\mathbf{\Phi} = (\mathbf{Z}, \mathbf{C})$ of $\mathbf{X}$, where $\mathbf{Z}$ represents IVs in $\mathbf{X}$ and $\mathbf{C}$ represents the remaining information in $\mathbf{X}$. Our problem setting is given in the following.

*Problem 1:* Given a joint distribution $P(\mathbf{X}, W, Y)$ generated from an underlying DAG $\mathcal{G} = (\mathbf{X} \cup \mathbf{U} \cup \{W, Y\}, \mathbf{E})$, $W$ and $Y$ are the treatment and outcome variables, respectively. $\mathbf{X}$ is pretreatment variables. $\mathbf{U}$ contains unobserved variables including unobserved confounders of $W$ and $Y$. Suppose that there exists at least one SIV (i.e., a set of SIVs $\mathbf{S} \subseteq \mathbf{X}$ with $|\mathbf{S}| \geq 1$). Our goal is to learn a latent IV representation $\mathbf{Z}$ through the disentanglement of the latent representation $\mathbf{\Phi}$ of $\mathbf{X}$ into two disjoint sets $(\mathbf{Z}, \mathbf{C})$ for recovering the causal effect of $W$ on $Y$.

## III. Proposed DIV.VAE Method

### A. Proposed Disentanglement Scheme

Following the literature [19], [20], [27], we propose the causal structure in Fig. 3 to represent the causal relationships among $W$, $Y$, $U'$, $\mathbf{X}$, $\mathbf{Z}$, and $\mathbf{C}$, where the set $\mathbf{X}$ is generated from the set of latent variables $\mathbf{\Phi} = (\mathbf{Z}, \mathbf{C})$, where $\mathbf{Z}$ is the latent IV representation, and $\mathbf{C}$ captures the remaining information in $\mathbf{X}$.

We first show that the proposed disentanglement scheme can estimate the causal effect of $W$ on $Y$ in presenting the following theorem.

*Theorem 1:* Given a joint distribution $P(\mathbf{X}, W, Y)$ generated from a causal DAG $\mathcal{G} = (\mathbf{X} \cup \mathbf{U} \cup \{W, Y\}, \mathbf{E})$, $\mathcal{G}$ contains $W \rightarrow Y$ and $W \leftarrow U' \rightarrow Y$ in $\mathcal{G}$, and $\forall X \in \mathbf{X}$,

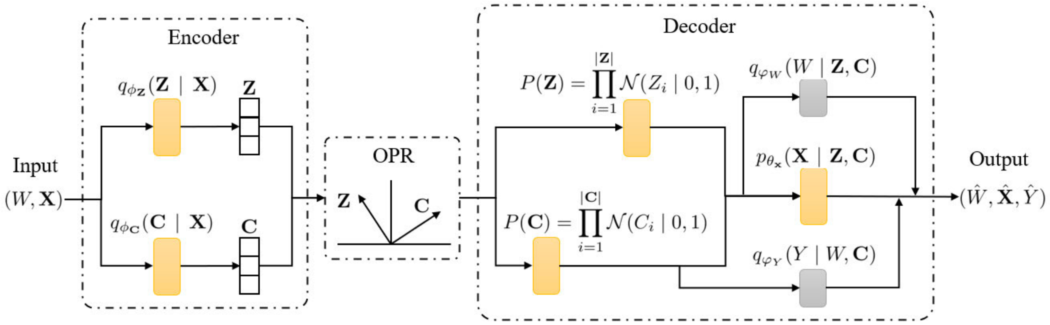

Encoder
$q_{\phi_{\mathbf{Z}}}(\mathbf{Z} \mid \mathbf{X})$
$\mathbf{Z}$
$q_{\phi_{\mathbf{C}}}(\mathbf{C} \mid \mathbf{X})$
$\mathbf{C}$
Input
$(W, \mathbf{X})$
OPR
$\mathbf{Z}$
$\mathbf{C}$
Decoder
$P(\mathbf{Z}) = \prod_{i=1}^{|\mathbf{Z}|} \mathcal{N}(Z_i \mid 0,1)$
$P(\mathbf{C}) = \prod_{i=1}^{|\mathbf{C}|} \mathcal{N}(C_i \mid 0,1)$
$q_{\varphi_W}(W \mid \mathbf{Z}, \mathbf{C})$
$p_{\theta_{\mathbf{X}}}(\mathbf{X} \mid \mathbf{Z}, \mathbf{C})$
$q_{\varphi_Y}(Y \mid W, \mathbf{C})$
Output
$(\hat{W}, \hat{\mathbf{X}}, \hat{Y})$

Fig. 4. DIV.VAE architecture. The input $\mathbf{X}$ is encoded by $q_{\phi_{\mathbf{Z}}}(\mathbf{Z}|\mathbf{X})$ and $q_{\phi_{\mathbf{C}}}(\mathbf{C}|\mathbf{X})$ into the parameters of the latent representation. The middle dashed box is the OPR for ensuring $\mathbf{Z} \perp\!\!\!\perp \mathbf{C}$. Samples are drawn from each of the latent representations using the reparametrized trick. The samples are then concatenated and decoded through $p_{\theta_{\mathbf{X}}}(\mathbf{X}|\mathbf{Z}, \mathbf{C})$. The two gray boxes indicate the two auxiliary predictors $q_{\varphi_W}(W|\mathbf{Z}, \mathbf{C})$ and $q_{\varphi_Y}(Y|W, \mathbf{C})$.

$X \notin \mathrm{De}(W \cup Y)$ in $\mathcal{G}$. There exists at least one SIV (i.e., a set of SIVs $\mathbf{S} \subseteq \mathbf{X}$ with $|\mathbf{S}| \geq 1$). If we learn and disentangle simultaneously the latent representation $\mathbf{\Phi}$ of $\mathbf{X}$ into two disjoint sets $(\mathbf{Z}, \mathbf{C})$, where $\mathbf{Z}$ is a common cause of $W$ and $\mathbf{X}$, and $\mathbf{C}$ is a common cause of $\mathbf{X}$, $W$, and $Y$, respectively, then $\mathbf{Z}$ is a valid IV for estimating the causal effect of $W$ on $Y$ from data of $P(\mathbf{X}, W, Y)$.

*Proof:* We prove that the IV representation $\mathbf{Z}$ is a valid IV based on the disentanglement causal model. First of all, the clause $|\mathbf{S}| \geq 1$ is to ensure that there is information of valid IVs in the set of covariates $\mathbf{X}$. In the causal DAG: 1) $\mathbf{Z}$ is a set of causes of $W$, so $\mathbf{Z}$ satisfies the first condition (1) of an IV as described in Introduction; 2) there is only a causal path from $\mathbf{Z}$ to $Y$, i.e., $\mathbf{Z} \to W \to Y$, so $\mathbf{Z}$ affects $Y$ only through $W$, i.e., $\mathbf{Z}$ satisfies the second condition (2) of an IV; and 3) there are four back-door paths from $\mathbf{Z}$ to $Y$, i.e., $\mathbf{Z} \to \mathbf{X} \leftarrow \mathbf{C} \to Y$, $\mathbf{Z} \to \mathbf{X} \leftarrow \mathbf{C} \to W \leftarrow U \to Y$, $\mathbf{Z} \to W \leftarrow \mathbf{C} \to Y$, and $\mathbf{Z} \to W \leftarrow U \to Y$, and all four back-door paths are blocked by $\emptyset$ according to the back-door criterion, i.e., $\mathbf{Z}$ does not share common causes with $Y$, so $\mathbf{Z}$ satisfies the last condition of an IV. Therefore, $\mathbf{Z}$ is a set of valid IVs for estimating the causal effect of $W$ on $Y$ from data with latent confounders. □

Theorem 1 states that the soundness of the proposed disentangled representation learning method relies on the ability to learn correct representations. The conditional clause "If we learn and disentangle simultaneously the latent representation $\mathbf{\Phi}$ of $\mathbf{X}$ into two disjoint sets $(\mathbf{Z}, \mathbf{C})$" in the theorem is an assumption that is unfortunately untestable in data. Such an assumption is used in previous VAE-based causal inference works [19], [20], [27]. Once $\mathbf{Z}$ is correctly learned, the causal effect of $W$ on $Y$ can be unbiasedly estimated from data with latent confounders. In real-world applications, $U$ may affect the learned representation of $\mathbf{C}$, but $\mathbf{C}$ is not used in the causal effect estimation and thus, the uncertainty in $\mathbf{C}$ does not affect the unbiasedness of causal effect $W$ on $Y$.

The establishment of Theorem 1 relies on the correctness of the disentanglement $\mathbf{\Phi} = (\mathbf{Z}, \mathbf{C})$. In this work, we leverage VAEs to optimize a variational lower bound on likelihood, enabling the learning of $\mathbf{\Phi}$. This approach requires substantially weaker assumptions about the data-generating process and the latent variable structure as in [27], [28], and [29]. In Section III-B, we will introduce our proposed DIV.VAE for learning $\mathbf{\Phi}$ and disentangling $\mathbf{\Phi}$ into two disjoint sets, $(\mathbf{Z}, \mathbf{C})$.

### B. Finding IV Representation by Disentangled Representation Learning

In this section, we introduce the details of the proposed VAE-based disentangled representation learning architecture of DIV.VAE (as shown in Fig. 4) to learn a valid IV representation $\mathbf{Z}$ following the proposed scheme in Fig. 3. Then, we can use the learned IV representation to obtain unbiased causal effect estimation from data with latent confounders.

The goal of our designed architecture for DIV.VAE is to learn the latent representation $\mathbf{\Phi}$ of $\mathbf{X}$ and disentangle $\mathbf{\Phi}$ into $(\mathbf{Z}, \mathbf{C})$ simultaneously, following the proposed causal structure in Fig. 3. It is worth noting that $\mathbf{C}$ plays a critical role as a set of auxiliary variables used to capture the information from the set of $\mathbf{X} \setminus \{\mathbf{S}\}$ in the representation learning and disentanglement process but it is not used for the causal effect estimation.

The proposed DIV.VAE architecture in Fig. 4 uses the inference model and the generation model to approximate the posterior $p_{\theta_{\mathbf{X}}}(\mathbf{X}|\mathbf{Z}, \mathbf{C})$ where $\theta$ is a set of generative model parameters. For the inference model, we develop two separate encoders $q_{\phi_{\mathbf{Z}}}(\mathbf{Z}|\mathbf{X})$ and $q_{\phi_{\mathbf{C}}}(\mathbf{C}|\mathbf{X})$ that serve as variational posteriors over the latent variables. For the generative model, the two latent representations $(\mathbf{Z}, \mathbf{C})$ are obtained from the two separate encoders used by a single decoder $p_{\theta_{\mathbf{X}}}(\mathbf{X}|\mathbf{Z}, \mathbf{C})$ to reconstruct $\mathbf{X}$. Following the VAE literature [28], [29], the prior distributions of $P(\mathbf{Z})$ and $P(\mathbf{C})$ are drawn from Gaussian distributions.

In the inference model, the variational approximations of the posteriors are defined as

$$
\begin{aligned}
q_{\phi_{\mathbf{Z}}}(\mathbf{Z}|\mathbf{X}) &= \prod_{i=1}^{|\mathbf{Z}|} \mathcal{N}\big(\mu = \hat{\mu}_{Z_i}, \sigma^2 = \hat{\sigma}^2_{Z_i}\big) \\
q_{\phi_{\mathbf{C}}}(\mathbf{C}|\mathbf{X}) &= \prod_{i=1}^{|\mathbf{C}|} \mathcal{N}\big(\mu = \hat{\mu}_{C_i}, \sigma^2 = \hat{\sigma}^2_{C_i}\big) \qquad (1)
\end{aligned}
$$

where $\hat{\mu}_{Z_i}, \hat{\mu}_{C_i}$ and $\hat{\sigma}^2_{Z_i}, \hat{\sigma}^2_{C_i}$ are the means and variances of the Gaussian distributions parameterized by neural networks, respectively. Note that, since one IV is sufficient for obtaining unbiased causal effect estimation, we use $|\mathbf{Z}| = 1$ in the experiments on real-world datasets. However, in the algorithm design, we keep $\mathbf{Z}$ as multidimensional for a general solution.

The prior distributions of $(\mathbf{Z}, \mathbf{C})$ are defined as

$$P(\mathbf{Z}) = \prod_{i=1}^{|\mathbf{Z}|} \mathcal{N}(Z_i|0, 1); \quad P(\mathbf{C}) = \prod_{i=1}^{|\mathbf{C}|} \mathcal{N}(C_i|0, 1). \tag{2}$$

The generative models for $W$ and $\mathbf{X}$ are defined as

$$\begin{aligned} p_{\theta_W}(W|\mathbf{Z}, \mathbf{C}) &= \text{Bern}(\sigma(g_1(\mathbf{Z}, \mathbf{C}))) \\ p_{\theta_{\mathbf{x}}}(\mathbf{X}|\mathbf{Z}, \mathbf{C}) &= \prod_{i=1}^{|\mathbf{X}|} P(X_i|\mathbf{Z}, \mathbf{C}) \end{aligned} \tag{3}$$

where $P(X_i|\mathbf{Z}, \mathbf{C})$ is the distribution for the $i$th measured variable, $g_1(\cdot)$ is the function parameterized by neural networks, and $\sigma(\cdot)$ is the logistic function.

The generative model for $Y$ depends on the data type of $Y$. For continuous $Y$, we sample it from a Gaussian distribution with its mean and variance given by the mutually exclusive neural networks that define $P(Y|W = 0, \mathbf{Z}, \mathbf{C})$ and $P(Y|W = 1, \mathbf{Z}, \mathbf{C})$, respectively, and the generative model of $Y$ is defined as

$$\begin{aligned} P(Y|W, \mathbf{C}) &= \mathcal{N}\big(\mu = \hat{\mu}_Y, \sigma^2 = \hat{\sigma}_Y^2\big) \\ \hat{\mu}_Y &= W \cdot g_2(\mathbf{C}) + (1 - W) \cdot g_3(\mathbf{C}) \\ \hat{\sigma}_Y^2 &= W \cdot g_4(\mathbf{C}) + (1 - W) \cdot g_5(\mathbf{C}) \end{aligned} \tag{4}$$

where $g_2(\cdot)$, $g_3(\cdot)$, $g_4(\cdot)$, and $g_5(\cdot)$ are the functions parameterized by neural networks. For binary $Y$, we parameterize it with a Bernoulli distribution, and the model is defined as

$$p_{\theta_Y}(Y|W, \mathbf{C}) = \text{Bern}(\sigma(g_6(W, \mathbf{C}))) \tag{5}$$

where $g_6(\cdot)$ is a neural network with its own parameters. Given the joint distribution $P(\mathbf{X}, W, Y)$, the parameters can be optimized by maximizing the evidence lower bound (ELBO) $\mathcal{M}$ [28]

$$\begin{aligned} \mathcal{M} = \; & \mathbb{E}_{q_{\phi_{\mathbf{Z}}} q_{\phi_{\mathbf{C}}}}\big[\log p_{\theta_{\mathbf{x}}}(\mathbf{X}|\mathbf{Z}, \mathbf{C})\big] \\ & - D_{\text{KL}}\big[q_{\phi_{\mathbf{Z}}}(\mathbf{Z}|\mathbf{X}) || P(\mathbf{Z})\big] \\ & - D_{\text{KL}}\big[q_{\phi_{\mathbf{C}}}(\mathbf{C}|\mathbf{X}) || P(\mathbf{C})\big] \end{aligned} \tag{6}$$

where $D_{\text{KL}}[\cdot||\cdot]$ is a Kullback-Leibler (KL) divergence term.

To learn the latent IV representation $\mathbf{Z}$ from the set of SIVs $\mathbf{S}$ and the latent representation $\mathbf{C}$ from the remaining variables $\mathbf{X}\backslash\{\mathbf{S}\}$, we add two auxiliary predictors to the above variational ELBO to ensure that the treatment variable $W$ and the outcome variable $Y$ can be estimated from $\mathbf{Z}$ and $\mathbf{C}$ as designed. Thus, we have the following objective function:

$$\begin{aligned} \mathcal{L}' = -\mathcal{M} + \alpha_W \mathbb{E}_{q_{\phi_{\mathbf{Z}}} q_{\phi_{\mathbf{C}}}}\big[\log q_{\varphi_W}(W|\mathbf{Z}, \mathbf{C})\big] \\ + \alpha_Y \mathbb{E}_{q_{\phi_{\mathbf{C}}}}\big[\log q_{\varphi_Y}(Y|W, \mathbf{C})\big] \end{aligned} \tag{7}$$

where $\alpha_W$ and $\alpha_Y$ are the weights for the two auxiliary predictors.

In practice, there may be some very weak associations between $\mathbf{Z}$ and $\mathbf{C}$. To encourage $\mathbf{Z} \perp\!\!\!\perp \mathbf{C}$ as specified in Fig. 3, we employ the orthogonality promoting regularization (OPR) [30] for our proposed DIV.VAE

$$\mathcal{L} = \mathcal{L}' + \frac{1}{b}\sum_{i=1}^{b} \text{CS}(\mathbf{Z}_i, \mathbf{C}_i) \tag{8}$$

where $b$ is the batch size of the neural network, and $\text{CS}(\mathbf{Z}_i, \mathbf{C}_i) = ((\mathbf{Z}_i^{\text{T}}\mathbf{C}_i)/(\|\mathbf{Z}_i\|_2\|\mathbf{C}_i\|_2))$ is the cosine similarity (CS).

After training DIV.VAE, we draw $\mathbf{Z}$ from the model and feed it into the function of Ortho.IV [22], [23] for calculating $\beta_{wy}$. When the learned distribution of $\mathbf{Z}$ is close to the true unmeasured IV distribution, the DIV.VAE method has the ability to obtain an unbiased estimate $\beta_{wy}$ as shown in the experimental results. Notably, the main advantage of our DIV.VAE is that it no longer requires domain knowledge or experts to provide a valid IV. Instead, it only requires the presence of an SIV in the data. This is a weaker assumption compared to those required by other methods such as two-stage least squares (TSLS), forest for IV regression (FIVR), deep ensemble method for the IV (DeepIV), and IV.Tetrad.

*Limitations:* The soundness of DIV.VAE relies on the ability of the proposed VAE architecture to learn $\mathbf{\Phi}$ and disentangle the latent variable $\mathbf{\Phi}$ into $(\mathbf{Z}, \mathbf{C})$. However, VAE-based methods are susceptible to the problem of unidentifiability in the VAE model [31], [32]. In other words, there is no theoretical guarantee that the learned IV representation $\mathbf{Z}$ can always approximate the true latent IV. Fortunately, as shown in our experiments, in the presence of SIV, the learned IV representation $\mathbf{Z}$ by DIV.VAE closely approximates the true latent IV. We note that identifiable VAE (iVAE) gives an identifiability guarantee but with more limitations [32]. iVAE assumes injective and linear relationships and a latent presentation learned by iVAE needs its child variable and parent variables to be observed. These additional requirements limit the application of a method based on iVAE and make it infeasible for iVAE to recover the IV representation from the error term of SIV. VAE does not guarantee the identifiability but has some advantages, such as not requiring linear and injective relationships or $\mathbf{Z}$'s parent to be observed for its representation learning. However, users should review their results by DIV.VAE with domain knowledge and perform sensitivity analyses [8] before taking the results.

# IV. Experiments

## A. Experimental Setup

*1) Baseline Causal Effect Estimators:* We compare DIV.VAE with four representative IV-based estimators and two VAE-based causal effect estimators. In three of the IV-based estimators, TSLS regression [9], causal random FIVR [33], and the popular DeepIV [34], each requires a given IV, whereas the other IV-based estimator, IV.Tetrad [13], does not require a given IV, but needs the majority of variables in $\mathbf{X}$ to be valid IVs. The two VAE-based estimators are causal effect VAE (CEVAE) [27] and treatment effect by disentangled VAE (TEDVAE) [20]. These two estimators assume no latent confounder between $(W, Y)$. They have been used in our experiments since our DIV.VAE is also based on the VAE model.

*2) Evaluation Metrics:* For synthetic datasets with the ground-truth $\beta_{wy}$, we use the estimation bias $|(\hat{\beta}_{wy} - \beta_{wy})/\beta_{wy}| * 100$ (%) to demonstrate the performance of all estimators. For the real-world datasets, we evaluate all estimators against the reference causal effects in the literature.

*3) Implementation Details:* DIV.VAE is implemented by using Python with packages including *PyTorch* [35], *Pyro* [36], and *EconML* [37]. The implementation of TSLS is based on the functions *glm* and *ivglm* in the R packages *stats* and *ivtools* [11]. FIVR is implemented using the function *instrumental forest* in the R package *grf* [33]. DeepIV is retrieved from the authors' GitHub.[2] IV.Tetrad is also retrieved from the authors' site.[3] CEVAE is implemented using the function *CEVAE* in the Python package *Pyro* [36] and TEDVAE is obtained from the authors' GitHub.[4]

In our experiments, we evaluate the performance of DIV.VAE against the above baselines on simulated and real-world data. We will provide the ablation experiments of our DIV.VAE and the empirical evaluation on the independence relation of $\mathbf{Z}$ and $\mathbf{C}$ in Sections IV-H and IV-I, respectively.

### B. Simulation Study

We utilize the true DAG over $\mathbf{X} \cup \mathbf{U} \cup \{W, Y\}$ as shown in Fig. 5 to generate the synthetic datasets with latent variables for the experiments by following the literature [38]. We generate datasets with a range of sample sizes: 0.5k, 2k, 4k, 6k, 8k, 10k, 50k, 100k, and 200k. The set of measured variables $\mathbf{X}$ consists of $\{S, X_1, X_2, X_3, X_4, X_5, X_6, X_7, W, Y\}$. The DAG also include three latent variables $\{U, U_1, U_2\}$ in the data, where $U$ affects both $W$ and $Y$. The details of data generation are introduced as follows.

The synthetic datasets are generated based on the DAG in Fig. 5, and the specifications are as follows: $Z \sim N(0, 1)$; $U_1, U_2 \sim N(0, 1)$; $X_1, X_3, X_5, X_7 \sim N(0, 1)$; $\epsilon_{1,2,3,4,S} \sim N(0, 0.5)$; $S \sim N(0, 1) + Z + \epsilon_S$; $U \sim N(0, 1) + 0.8 * X_1 + \epsilon_1$; $X_2 \sim N(0, 1) + 2 * U + \epsilon_2$; $X_4 \sim N(0, 1) + U_1 + \epsilon_3$; and$X_6 \sim N(0, 1) + 0.6 * U_2 + \epsilon_4$, where $N(,)$ denotes the normal distribution. The treatment assignment $W$ is generated from $n$ (where $n$ is the sample size) Bernoulli trials by using the assignment probability based on the measured variables $\{X_4, X_5\}$ and latent variables $\{U, U_2\}$ as $P(W = 1|U, Z, X_4, X_5, U_2) = [1 + \exp\{2 - 2 * U - 2 * Z - 3 * X_4 - X_5 - 3 * U_2\}]^{-1}$.

In this work, we generate two types of potential outcomes, i.e., a linear function $Y_{\text{linear}}$ and a nonlinear function $Y_{\text{nonlinear}}$ for evaluating the ability of DIV.VAE in terms of causal effect estimation. $Y_{\text{linear}} = 2 + 2 * W + 2 * U + 3 * U_1 + 2 * X_3 + 2 * X_6 + 2 * X_7 + \epsilon_w$, where $\epsilon_w \sim N(0, 1)$, and $Y_{\text{nonlinear}} = 2 + 2 * W + 2 * U + 3 * U_1 + 2 * X_3 + 2 * X_6^2 + 2 * X_7 + \epsilon_w$. Note that the causal effect of $W$ on $Y$ is fixed to 2, i.e., $\beta_{wy} = 2$ on all synthetic datasets.

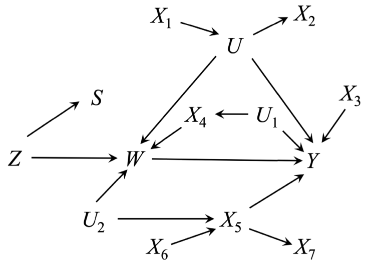


Fig. 5. True causal DAG with a latent confounder $U$ between $W$ and $Y$ is used to generate the synthetic datasets. $Z$ and $S$ are a latent IV and an SIV, respectively. $\{U_1, U_2\}$ are the two latent variables, and other measured variables are pretreatment variables of $(W, Y)$.

To avoid the random noises brought by data generation process, we repeatedly generated 30 datasets for each sample size. In our simulation experiments, we use the SIV $S$ in the underlying causal DAG as the known IV for the compared IV-based estimators, TSLS, FIVR, and DeepIV.

*1) Performance of DIV.VAE on Causal Effect Estimation:* The estimation biases of all estimators on synthetic datasets with $Y_{\text{linear}}$ and synthetic datasets $Y_{\text{nonlinear}}$ are visualized with boxplots in Figs. 6 and 7, respectively.

*Results:* From the experimental results, we have the following observations.

1) On all synthetic datasets, DIV.VAE consistently exhibits low bias and small variance, outperforming all compared estimators as the sample size increases.
2) DIV.VAE and FIVR both have low bias across all datasets with $Y_{\text{linear}}$, but the performance of DIV.VAE is more stable than FIVR. For smaller-sized datasets, DIV.VAE has a smaller variance and bias than FIVR. This is because FIVR uses SIV, which is a proxy of IV and this results in large variance with finite samples. DIV.VAE uses $\mathbf{Z}$, the representation of the IV. When the representation is learned properly, the estimation of DIV.VAE is unbiased.
3) For the two VAE-based estimators, CEVAE and TEDVAE, they have relatively low variances, but compared to DIV.VAE, their biases are much larger on both types of synthetic datasets since both methods do not allow a latent confounder $U$ between $(W, Y)$.
4) TSLS and DeepIV exhibit small biases on both types of datasets since they use the true SIV as a valid IV.
5) IV.Tetrad performs consistently poorly since the majority of valid IV assumptions does not hold on both types of datasets.

*2) Correctness of Learned IV Representation* $\mathbf{Z}$*:* We examine the quality of the learned representation $\mathbf{Z}$ by visualizing the probability density functions (pdfs) of the ground-truth IV and the learned IV representation $\mathbf{Z}$. We use the learned IV representation $\mathbf{Z}$ from the data with $Y_{\text{linear}}$ for the visualization in Fig. 8. We have the following observations from Fig. 8: 1) the learned IV representation $\mathbf{Z}$ approximates the ground-truth pdf very well even when the sample size is small and 2) as the sample size increases, the pdf of learned IV representation $\mathbf{Z}$ closely matches the ground-truth pdf. Thus, DIV.VAE can learn the correct IV representation from data with latent confounders.

[2] https://github.com/jhartford/DeepIV

[3] http://www.homepages.ucl.ac.uk/~ucgtrbd/code/iv_discovery

[4] https://github.com/WeijiaZhang24/TEDVAE

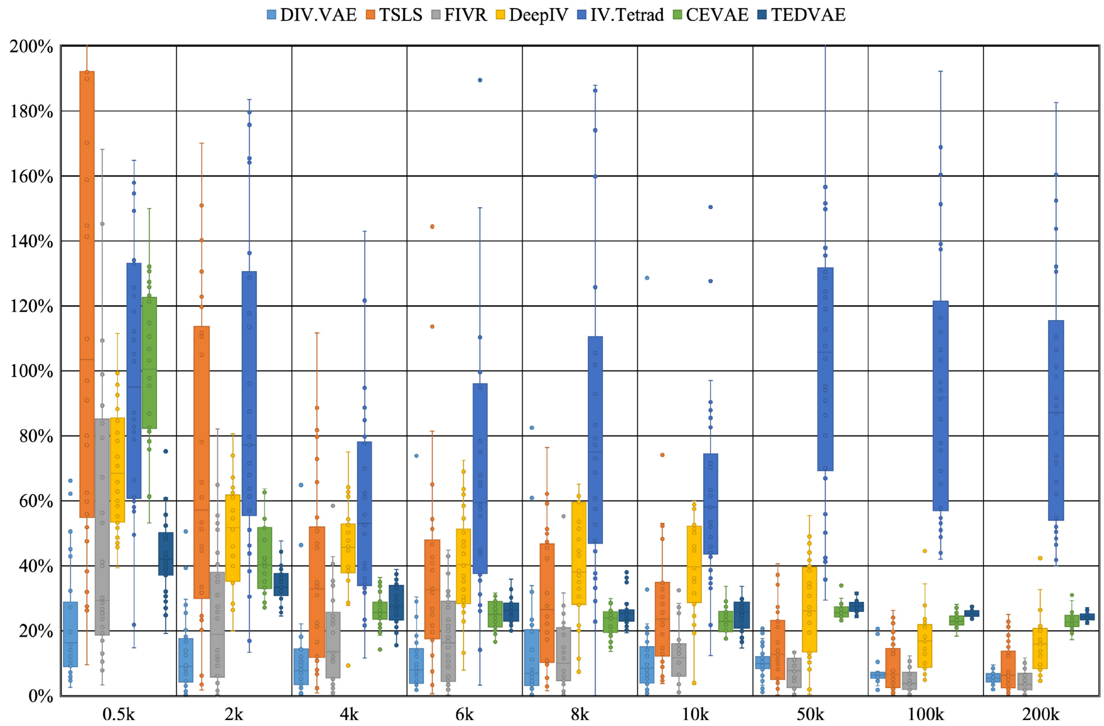


Fig. 6. Estimation biases over 30 synthetic datasets with $Y_{\text{linear}}$ for different estimators, where the horizontal axis represents the sample size and the vertical axis represents the estimation bias (%). FIVR performs competitively with DIV.VAE in large datasets. FIVR needs a given IV whereas DIV.VAE does not.

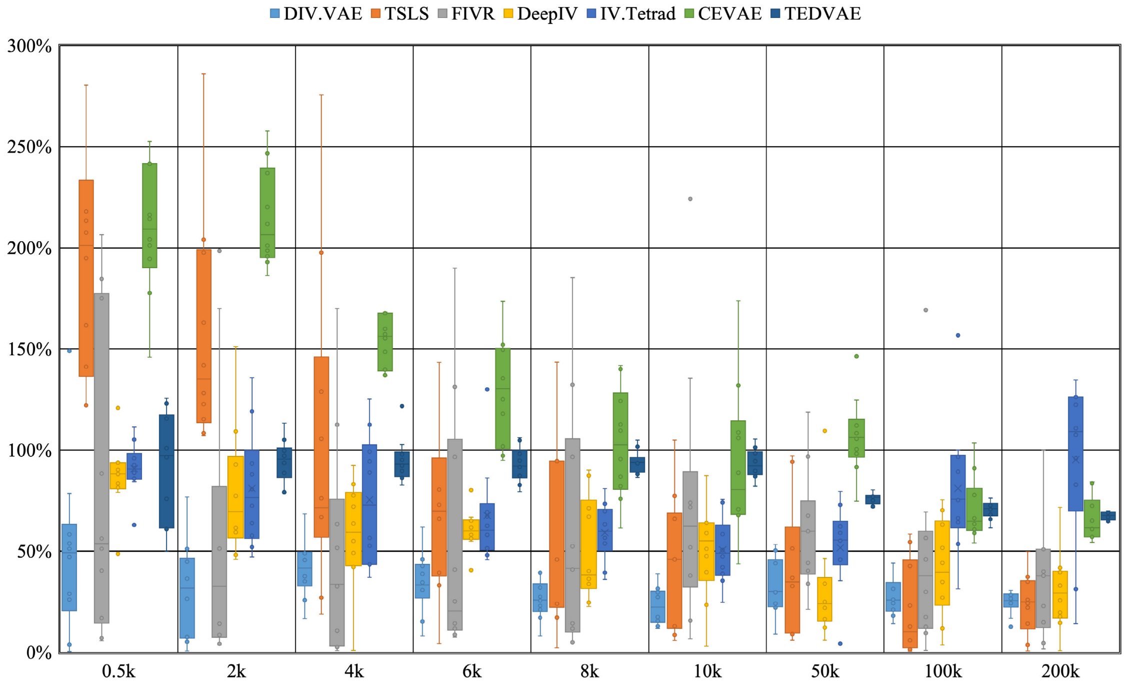


Fig. 7. Estimation biases over 30 synthetic datasets with $Y_{\text{nonlinear}}$ for different estimators, where the horizontal axis represents the sample size and the vertical axis represents the estimation bias (%). DeepIV performs competitively with DIV.VAE in large datasets. DeepIV needs a given IV whereas DIV.VAE does not.

To sum up, DIV.VAE is effective in achieving accurate and stable causal effect estimation from data without giving an IV.

### C. Experiments on Three Real-World Datasets

In this section, we conduct experiments on two commonly used benchmark datasets with known IVs, Vitamin D (VitD) [39] and Schooling Returns [40]. The empirical estimates of the causal effects of the two datasets are widely accepted. Another dataset, Sachs, is from a real application [41]. There is no nominated IV variable for Sachs.

*1) VitD Data:* VitD is a cohort study of VitD on mortality reported in [39]. The data contain 2571 individuals and five variables: age, filaggrin (a binary variable indicating filaggrin mutations), vitd [a continuous variable measured as serum

TABLE I

Results of All Estimators on the Three Real-World Datasets. The Estimated Causal Effects Within the 95% Confidence Interval Are Highlighted. "-" Indicates That a Method Is Not Applicable Since No IV Is Given on the Sachs Dataset

| Estimators | DIV.VAE | TSLS | FIVR | DeepIV | IV.Tetrad | CEVAE | TEDVAE |
|---|---|---|---|---|---|---|---|
| VitD | **3.543** | **3.940** | 0.205 | 0.012 | **2.150** | 0.006 | -0.042 |
| Schoolingreturns | **0.175** | 0.504 | 1.151 | -0.027 | **0.064** | **0.068** | -0.021 |
| Sachs | **1.303** | - | - | - | **1.430** | 0.239 | 0.254 |

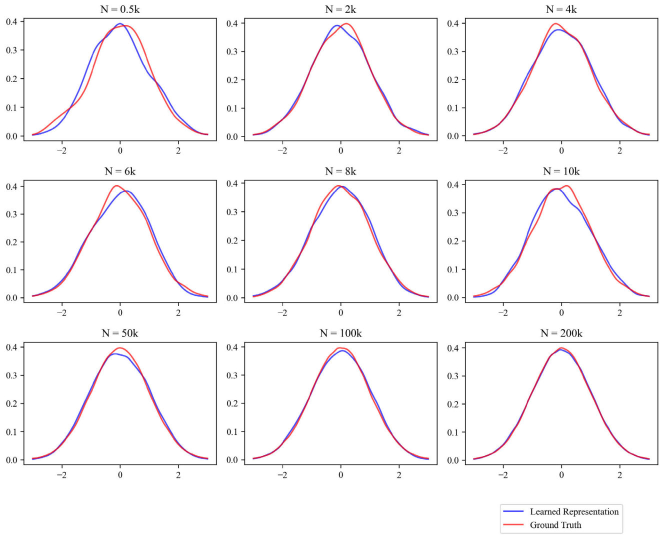


Fig. 8. PDFs of the ground-truth IV and the learned representation $Z$, where the horizontal axis represents the value and the vertical axis represents the density.

25-OH-D (nmol/L)], time (follow-up time), and death (binary outcome indicating whether an individual died during follow-up) [11]. The measured value of VitD less than 30 nmol/L implies VitD deficiency. The indicator of filaggrin is used as an instrument [39]. We take the estimated $\hat{\beta}_{wy} = 2.01$ with 95% conditional interval (0.96, 4.26) from the work [39] as the reference causal effect.

2) *Schoolreturning:* The data are from the National Longitudinal Survey of Youth (NLSY), a well-known dataset of U.S. young employees, aged range from 24 to 34 [40]. The treatment is the education of employees, and the outcome is raw wages in 1976 (in cents per hour). The data contain 3010 individuals and 19 covariates. The covariates include experience (years of labor market experience), ethnicity (factor indicating ethnicity), resident information of an individual, age, nearcollege (whether an individual grew up near a 4-year college?), marital status, father's educational attainment, and mother's educational attainment. A goal of the studies on this dataset is to investigate the causal effect of education on earnings. Card [40] used geographical proximity to a college, i.e., the covariate *nearcollege* as an instrument variable. We take $\hat{\beta}_{wy} = 0.1329$ with 95% conditional interval (0.0484, 0.2175) from [42] as the reference causal effect.

3) *Sachs:* This data is collected from cell activity measurements for single-cell data under a variety of conditions [41]. Following the work [13], we focus on a single condition, i.e., simulation with anti-CD3 and anti-CD28. The data contain 853 records and 11 variables [41]. The treatment is the manipulation of concentration levels of molecule Erk. The outcome is the concentration of Akt. The other nine cell products are pretreatment variables [41]. The data have some weak correlations among variables, but we assume that there are no conditional independencies held between Erk and the remaining ten variables. Note that there is not a given IV. We take the estimated $\hat{\beta}_{wy} = 1.43$ from the literature [13] (i.e., IV.Tetrad's estimated value) as the reference causal effect.

4) *Results:* From the results in Table I, we see the estimated causal effects of DIV.VAE for VitD and Schooling Returns are in their empirical intervals. On Sachs, the estimated causal effects by DIV.VAE are close to IV.Tetrad's estimated value. These results confirm that DIV.VAE is capable of recovering a latent IV representation from data. The causal effects estimated by IV.Tetrad are in the empirical intervals of VitD and Schooling Returns since both datasets are low dimensional and satisfy the assumptions of IV.Tetrad. The other baselines either work well on VitD, or work well on Schooling Returns, but not on both. The Sachs dataset is not applicable to TSLS, FIVR, and DeepIV since the dataset does not have a nominated IV. The two VAE-based methods do not work well on the Sachs dataset.

In sum, DIV.VAE, without needing a nominated IV, performs better or competitively with the state-of-the-art IV-based or VAE-based estimators on the three real-world datasets, further confirming the effectiveness of the proposed DIV.VAE method and suggesting the potential of DIV.VAE in real-world applications.

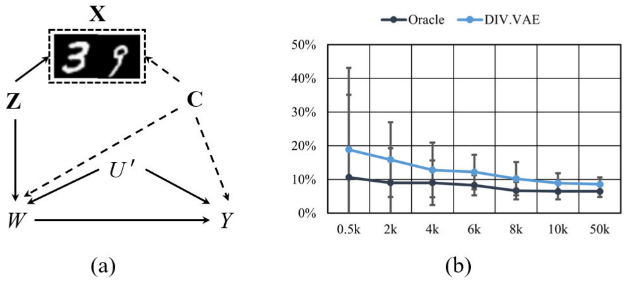
### D. Evaluation in Higher Dimension With Tabular Data

To evaluate the performance of our DIV.VAE with higher dimensional datasets, we generate synthetic datasets with a range of sample sizes: 0.5k, 2k, 4k, 6k, 8k, and 10k, and varying the number of measured variables as 8, 16, 32, and 64 using the same process described in Section IV-B.

Note that 64 variables are not considered as high dimensional in general machine learning settings. In causal effect estimation, however, the variables are not many since pretreatment variables are handpicked by domain experts [8]. We do not run DIV.VAE in a high-dimensional setting due to the faithfulness assumption it requires. Higher dimensionality is not a problem for DIV.VAE, but will pose a problem for simulation data generation. A dataset generated needs to be faithful to the underlying DAG for data generation, and to preserve the conditional independencies among a large number of variables, the size of the dataset (i.e., the number of samples) needs to be very large. A dataset with a large number of samples takes a long time for representation

TABLE II
ESTIMATION BIAS OF DIV.VAE IN EACH SETTING OVER 30 SYNTHETIC DATASETS (MEAN ± STD)

| Dimensions of variables | Sample sizes | | | | | |
|---|---|---|---|---|---|---|
| | 0.5k | 2k | 4k | 6k | 8k | 10k |
| 8 | 18.15 ± 39.36 | 17.25 ± 21.25 | 14.12 ± 19.11 | 16.83 ± 11.66 | 15.7 ± 24.66 | 13.88 ± 7.65 |
| 16 | 23.74 ± 10.07 | 27.57 ± 4.2 | 25.03 ± 5.52 | 22.54 ± 5.71 | 20.48 ± 4.22 | 16.55 ± 3.42 |
| 32 | 31.65 ± 18.72 | 30.39 ± 6.65 | 27.88 ± 5.21 | 25.04 ± 9.56 | 21.44 ± 5.38 | 18.16 ± 8.22 |
| 64 | 36.64 ± 12.7 | 30.63 ± 14.14 | 31.14 ± 12.19 | 26.56 ± 13.12 | 23.29 ± 13.73 | 18.56 ± 11.64 |



(a) (b)

Fig. 9. (a) True causal DAG for image data. **X** is replaced by image data, i.e., an image of two handwritten digits. (b) Estimation biases over 30 times for image data, where the horizontal axis represents the sample size and the vertical axis represents the estimation bias (%).

learning with VAE. This is why we just vary the number of variables up to 64.

We use the synthetic datasets with $Y_{\text{linear}}$ to examine the performance of DIV.VAE in this section. For each setting, we generate 30 datasets repeatedly to reduce the impact of random noise, as described in Section IV-B.

*Results:* We report the experimental results in Table II. We have the following observations.

1) As the number of variables increases, DIV.VAE has a large estimation bias. This is because a higher dimensional dataset needs a significantly larger dataset to ensure that the dataset and the underlying causal DAG are faithful to each other. When the faithfulness assumption is not satisfied, both bias and variance of estimates are large.
2) An increase in the number of samples results in a decrease in the bias. The reason is the same as before. Therefore, in the case of handling high-dimensional data, DIV.VAE requires a large sample size to reduce estimation bias.

### E. Capability of DIV.VAE in Image Data

In Sections IV-B–IV-D, we demonstrated the performance of DIV.VAE on both simulated and real-world relational datasets. However, in many applications, we do not have variables in table format and pixels do not have distinct semantic meaning as traditional variables. Instead, the learned representations of images can be mapped to traditional variables with semantic meaning. We design this experiment to demonstrate the capability of DIV.VAE with image data.

To simulate this, we replaced the covariates **X** with the pixels of two handwritten digits from the modified national institute of standards and technology (MNIST) dataset [43] as done in the literature [44], [45]. The dimension of **X** is $2 * 28 * 28 = 1568$. The datasets are generated based on the DAG in Fig. 9(a), and the specifications are as follows: $U' \sim N(0, 0.05)$, where $N(,)$ denotes the normal distribution. The treatment assignment $W$ is generated from $n$ (where $n$ is the sample size) Bernoulli trials using the assignment probability based on variables $\{\mathbf{Z}, \mathbf{C}\}$, where **Z** represents the ten digits of a two-digit handwritten number (i.e., IV), and **C** represents the one digit of a two-digit handwritten number and latent variables $\{U'\}$ as $P(W = 1|\mathbf{Z}, \mathbf{C}, U') = [1 + \exp\{2 - U' - \mathbf{Z} - \mathbf{C}\}]^{-1}$. The outcome is formulated by $Y = 2 * \mathbf{C} + 10 * W + U' + \epsilon_y$. Note that the causal effect of $W$ on $Y$ is fixed at 10.

To demonstrate the representation learning ability of DIV.VAE, we consider an Oracle setting as the baseline. The Oracle set can access the outcome label from image data. **Z** and **C** are read from a two-digit handwritten number. Note that in this case, bias is not zero because of $\{U'\}$ and $\epsilon_y$. Instead, DIV.VAE learns the representation of **Z** and **C** from **X** (i.e., a two-digit handwritten number, by concatenating or combining two images from the MNIST dataset).

*Results:* The results are visualized in Fig. 9(b) with the mean and standard deviation (std) of 30 runs. The Oracle generates some estimation bias due to inconsistency between the generated datasets and the true causal DAG. This bias decreases as the sample size increases. DIV.VAE performs worse than Oracle with small sample sizes but improves significantly as the sample size grows. At sample sizes of 10k and 50k, DIV.VAE performs similarly to the Oracle, demonstrating its ability to extract valid IV from high-dimensional image data.

### F. Parameter Analysis

With the DIV.VAE algorithm, two tuning parameters, namely, $\alpha_W$ and $\alpha_Y$, are used to balance $\mathcal{L}_{\text{ELBO}}$ and the two classifiers. We examine the parameter settings $\alpha_W$ and $\alpha_Y$ across a range of values, specifically $\{0.01, 0.1, 1, 10, 100, 1000, 10\,000\}$, to analyze the sensitivity of DIV.VAE on synthetic datasets with a sample size of 10k. Note that the package *Pyro* requires $\alpha_W$ and $\alpha_Y$ to be the same. These datasets are generated using the same data generation process presented in Section IV-B. The estimation biases of DIV.VAE are reported in Table III. From Table III, we observe that DIV.VAE achieves the smallest estimation bias when both parameters, $\alpha_W$ and $\alpha_Y$, are set to 100. There is a need to tune parameters $\alpha_W$ and $\alpha_Y$ in an application.

### G. Evaluation of the Dimension of Latent IV Representation

In our implementation, we set $|\mathbf{Z}|$ to 1 in the disentangling process. We use this experiment to demonstrate the effectiveness of the setting. To do so, we use the same data generation

TABLE III

Estimation Bias in Different Settings and Different Values of Tuning Parameters $\alpha_W$ and $\alpha_Y$

| $\{\alpha_W, \alpha_Y\}$ | Dataset | |
|---|---|---|
| | Linear | Nonlinear |
| 0.01 | 80.5 ± 51.17 | 86.65 ± 96.38 |
| 0.1 | 67.25 ± 21.59 | 65.15 ± 40.72 |
| 1 | 52.42 ± 25.97 | 57.2 ± 33.12 |
| 10 | 32.57 ± 22.72 | 25.5 ± 7.85 |
| 100 | 13.88 ± 7.65 | 23.13 ± 9.18 |
| 1,000 | 26.26 ± 13.15 | 25.81 ± 15.06 |
| 10,000 | 35.06 ± 9.99 | 31.82 ± 16.62 |

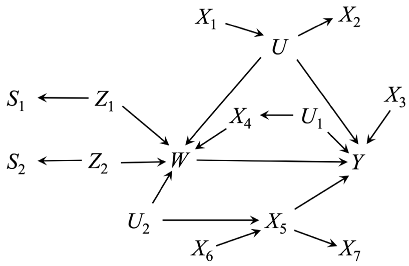


Fig. 10. True causal DAG with a latent common cause $U$ between $W$ and $Y$ is used to generate the synthetic datasets. $\{Z_1, Z_2\}$ and $\{S_1, S_2\}$ are latent IVs and SIVs, respectively. $\{U_1, U_2\}$ are two latent variables, and other measured variables are pretreatment variables of $(W, Y)$.

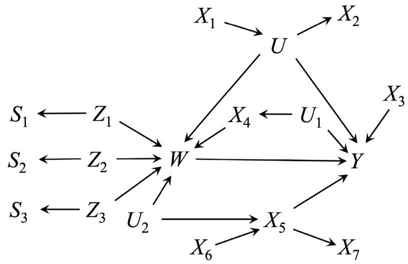


Fig. 11. True causal DAG with a latent common cause $U$ between $W$ and $Y$ is used to generate the synthetic datasets. $\{Z_1, Z_2, Z_3\}$ and $\{S_1, S_2, S_3\}$ are latent IVs and SIVs, respectively. $\{U_1, U_2\}$ are two latent variables, and other measured variables are pretreatment variables of $(W, Y)$.

process as used in Section IV-B with three causal DAGs in Figs. 5, 10, and 11 to generate three groups of synthetic datasets such that the three groups of datasets contain one SIV, two SIVs, and three SIVs, respectively, with other variables and causal relationships remaining unchanged. We repeatedly generate 30 datasets for each group to avoid bias in data generation. The estimation biases of DIV.VAE on the three groups of synthetic data are reported in Fig. 12.

The second group of datasets are generated from the DAG in Fig. 10. The generation processes different from Section IV-B are discussed in the following: $Z_1, Z_2 \sim N(0, 1)$; $\epsilon_{S_1,S_2} \sim N(0, 0.5)$; $S_1 \sim N(0, 1) + Z_1 + \epsilon_{S_1}$; and $S_2 \sim N(0, 1) + Z_2 + \epsilon_{S_2}$.

The treatment assignment $W$ is generated from $n$ Bernoulli trials by using the assignment probability $P(W = 1|U, Z_1, Z_2, X_4, X_5, U_2) = [1 + \exp\{2 - 2 * U - 2 * Z_1 - 2 * Z_2 3 * X_4 - X_5 - 3 * U_2\}]^{-1}$. The generation processes of other variables are the same as done in Section IV-B.

The third group of datasets is generated from the DAG in Fig. 11. The generation processes different from Section IV-B are discussed in the following: $Z_1, Z_2, Z_3 \sim N(0, 1)$; $\epsilon_{S_1,S_2,S_3} \sim N(0, 0.5)$; $S_1 \sim N(0, 1) + Z_1 + \epsilon_{S_1}$; $S_2 \sim N(0, 1) + Z_2 + \epsilon_{S_2}$; and$S_3 \sim N(0, 1) + Z_3 + \epsilon_{S_3}$.

The treatment assignment $W$ is generated from $n$ Bernoulli trials by using the assignment probability: $P(W = 1|U, Z_1, Z_2, Z_3, X_4, X_5, U_2) = [1+\exp\{2-2*U-2*Z_1-2*Z_2-2*Z_3+3*X_4-X_5-3*U_2\}]^{-1}$. The generation processes of other variables are the same as discussed in Section IV-B.

From Fig. 12, we see that the estimation biases of DIV.VAE on all three groups of synthetic datasets are consistently small. That means it is safe to set $|\mathbf{Z}|$ to 1 for DIV.VAE when we have no enough knowledge about the number of SIVs in real-world datasets.

### H. Ablation Study

Here, we develop a variant DIV.VAE to explore the component of the OPR for the average causal effect estimation from data with latent confounders. The variant of DIV.VAE is the loss function in (7) in the main text without the OPR term. The variant DIV.VAE is referred to as DIV.VAE$_{\text{w/o.OPR}}$.

We conduct the ablation study experiment on the synthetic datasets that are used in Section IV-B of the main text to show the performance of the proposed DIV.VAE with DIV.VAE$_{\text{w/o.OPR}}$. The experimental results are visualized in Fig. 13.

*Results:* From the results in Fig. 13, DIV.VAE$_{\text{w/o.OPR}}$ has a larger estimation bias than DIV.VAE on all synthetic datasets. Moreover, the variance of DIV.VAE$_{\text{w/o.OPR}}$ is also larger than DIV.VAE. The observations show the importance of the OPR in learning and disentangling the latent IV representation $\mathbf{Z}$ from the latent representation $\mathbf{\Phi} = (\mathbf{Z}, \mathbf{C})$ for causal effect estimation from data with latent confounders.

### I. Empirical Evaluation on the Independence Relation of Z and C

In this section, we conduct an empirical evaluation of the independence relation of $\mathbf{Z}$ and $\mathbf{C}$ by using the synthetic datasets generated in Section IV-B. In the empirical evaluation, we use the Pearson product-moment correlation coefficients (PCCs) as the evaluation index. In general, two variables are unrelated when PCC is less than 0.3. We report the mean with the std over 30 replications. In all experiments, $|\mathbf{Z}| = 1$ and $|\mathbf{C}| = 10$. Hence, we calculate the PCC of each pair $(Z, C_i)$ where $C_i \in \mathbf{C}$. The PCCs of each pair $(Z, C_i)$ in $(\mathbf{Z}, \mathbf{C})$ by using DIV.VAE$_{\text{w/o.OPR}}$ and DIV.VAE are reported in Tables IV and V, respectively.

*Results:* From Tables IV and V, we have three observations: 1) the PCC between $Z$ and each $C_i$ in $\mathbf{C}$ is less than 0.3, i.e., $Z$ and each $C_i$ in $\mathbf{C}$ are uncorrelated; 2) as the sample increases, the mean of the PCCs between $Z$ and each $C_i$ in $\mathbf{C}$ dropped significantly; and 3) as the sample increases, the std of the PCCs between $Z$ and each $C_i$ in $\mathbf{C}$ is also decreased.

Therefore, these results show that $Z$ and each $C_i$ in $\mathbf{C}$ are well-disentangled.

By comparing Tables IV and V, we know that the PCCs of DIV.VAE is significantly smaller than the PCCs of DIV.VAE$_{\text{w/o.OPR}}$'s on all synthetic datasets. It further confirms that the OPR term plays an important role in

TABLE IV

PCCs of Each Pair of $(Z, C_i)$ for DIV.VAE$_{\text{w/o.OPR}}$ (Mean ± Std)

| N | 0.5k | 2k | 4k | 6k | 8k | 10k |
|---|---|---|---|---|---|---|
| PCC($Z$,$C_1$) | 0.246 ± 0.15 | 0.232 ± 0.15 | 0.141 ± 0.09 | 0.174 ± 0.11 | 0.107 ± 0.09 | 0.107 ± 0.07 |
| PCC($Z$,$C_2$) | 0.263 ± 0.15 | 0.209 ± 0.14 | 0.189 ± 0.12 | 0.150 ± 0.10 | 0.159 ± 0.09 | 0.092 ± 0.08 |
| PCC($Z$,$C_3$) | 0.256 ± 0.16 | 0.201 ± 0.13 | 0.175 ± 0.10 | 0.143 ± 0.12 | 0.133 ± 0.10 | 0.101 ± 0.07 |
| PCC($Z$,$C_4$) | 0.307 ± 0.15 | 0.208 ± 0.14 | 0.193 ± 0.14 | 0.185 ± 0.10 | 0.143 ± 0.10 | 0.123 ± 0.08 |
| PCC($Z$,$C_5$) | 0.277 ± 0.17 | 0.196 ± 0.12 | 0.166 ± 0.12 | 0.166 ± 0.12 | 0.141 ± 0.11 | 0.091 ± 0.07 |
| PCC($Z$,$C_6$) | 0.237 ± 0.16 | 0.218 ± 0.14 | 0.189 ± 0.13 | 0.187 ± 0.10 | 0.144 ± 0.12 | 0.097 ± 0.08 |
| PCC($Z$,$C_7$) | 0.243 ± 0.15 | 0.181 ± 0.14 | 0.149 ± 0.11 | 0.147 ± 0.10 | 0.110 ± 0.09 | 0.086 ± 0.06 |
| PCC($Z$,$C_8$) | 0.254 ± 0.17 | 0.185 ± 0.12 | 0.173 ± 0.11 | 0.136 ± 0.09 | 0.125 ± 0.08 | 0.093 ± 0.07 |
| PCC($Z$,$C_9$) | 0.267 ± 0.17 | 0.230 ± 0.14 | 0.160 ± 0.10 | 0.147 ± 0.11 | 0.139 ± 0.10 | 0.103 ± 0.10 |
| PCC($Z$,$C_{10}$) | 0.293 ± 0.12 | 0.207 ± 0.15 | 0.174 ± 0.13 | 0.164 ± 0.13 | 0.142 ± 0.11 | 0.101 ± 0.09 |
| Average | 0.264 ± 0.16 | 0.207 ± 0.14 | 0.171 ± 0.11 | 0.160 ± 0.11 | 0.134 ± 0.10 | 0.101 ± 0.08 |

TABLE V

PCCs of Each Pair of $(Z, C_i)$ for DIV.VAE (Mean ± Std)

| N | 0.5k | 2k | 4k | 6k | 8k | 10k |
|---|---|---|---|---|---|---|
| PCC($Z$,$C_1$) | 0.206 ± 0.13 | 0.177 ± 0.11 | 0.142 ± 0.11 | 0.092 ± 0.08 | 0.082 ± 0.08 | 0.055 ± 0.04 |
| PCC($Z$,$C_2$) | 0.203 ± 0.14 | 0.172 ± 0.11 | 0.100 ± 0.08 | 0.094 ± 0.06 | 0.080 ± 0.06 | 0.062 ± 0.04 |
| PCC($Z$,$C_3$) | 0.180 ± 0.14 | 0.205 ± 0.10 | 0.152 ± 0.09 | 0.104 ± 0.10 | 0.080 ± 0.05 | 0.052 ± 0.04 |
| PCC($Z$,$C_4$) | 0.225 ± 0.14 | 0.177 ± 0.09 | 0.083 ± 0.09 | 0.081 ± 0.06 | 0.064 ± 0.06 | 0.089 ± 0.06 |
| PCC($Z$,$C_5$) | 0.194 ± 0.13 | 0.152 ± 0.11 | 0.113 ± 0.08 | 0.088 ± 0.06 | 0.080 ± 0.07 | 0.070 ± 0.07 |
| PCC($Z$,$C_6$) | 0.199 ± 0.11 | 0.164 ± 0.11 | 0.141 ± 0.11 | 0.101 ± 0.08 | 0.071 ± 0.05 | 0.089 ± 0.07 |
| PCC($Z$,$C_7$) | 0.165 ± 0.11 | 0.189 ± 0.11 | 0.168 ± 0.10 | 0.090 ± 0.08 | 0.097 ± 0.07 | 0.071 ± 0.06 |
| PCC($Z$,$C_8$) | 0.207 ± 0.11 | 0.153 ± 0.13 | 0.167 ± 0.11 | 0.071 ± 0.06 | 0.097 ± 0.08 | 0.081 ± 0.07 |
| PCC($Z$,$C_9$) | 0.210 ± 0.13 | 0.199 ± 0.13 | 0.118 ± 0.09 | 0.069 ± 0.05 | 0.071 ± 0.05 | 0.075 ± 0.06 |
| PCC($Z$,$C_{10}$) | 0.192 ± 0.12 | 0.205 ± 0.14 | 0.137 ± 0.11 | 0.106 ± 0.08 | 0.079 ± 0.08 | 0.070 ± 0.06 |
| Average | 0.198 ± 0.13 | 0.179 ± 0.11 | 0.132 ± 0.10 | 0.089 ± 0.07 | 0.080 ± 0.06 | 0.071 ± 0.06 |

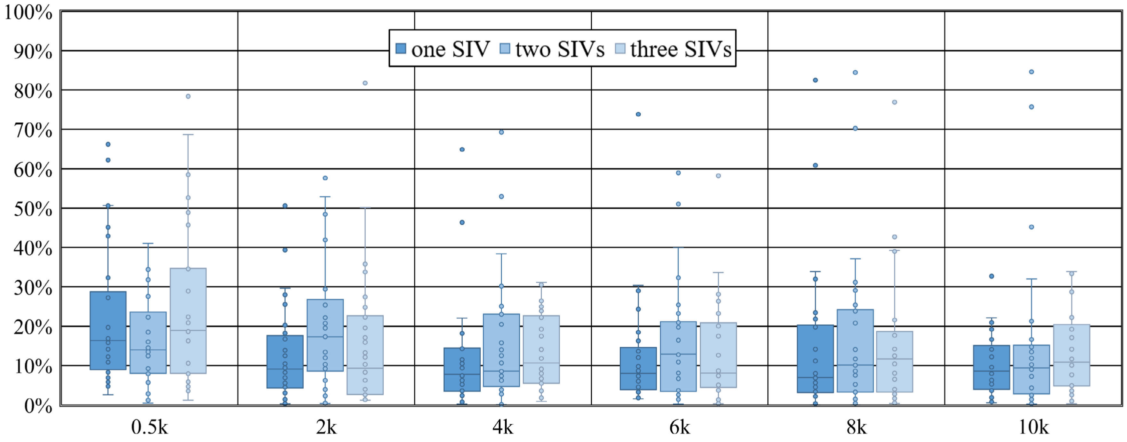


Fig. 12. Experimental results of DIV.VAE on all three groups of synthetic datasets with $|\mathbf{Z}| = 1$, where the horizontal axis represents the sample size and the vertical axis represents the estimation bias (%). Mismatching between the number of latent IVs and the number of SIVs does not cause performance deterioration.

encouraging $\mathbf{Z} \perp\!\!\!\perp \mathbf{C}$ in learning and disentangling the latent IV representation $\mathbf{Z}$ from the latent representation $\mathbf{\Phi} = (\mathbf{Z}, \mathbf{C})$ of $\mathbf{X}$.

## V. Related Work

In this section, we review the research closely related to this work, including IV-based methods with a given IV, data-driven IV-based methods without a given IV, and deep learning (including VAE)-based causal effect estimation.

### A. IV-Based Methods With a Given IV

In practice, before we use an IV method, one needs to nominate a valid IV based on domain knowledge. Under the assumption that a valid IV has been given, several IV-based counterfactual prediction methods have been developed for heterogeneous causal effect estimation, such as instrumental random forest regression [33], generalized method of moments (GMMs) [46], DeepIV [34], and kernel IV (KIV) regression [47]. Yuan et al. [45] proposed a novel AutoIV algorithm to automatically generate IV representation for the downstream IV-based counterfactual prediction under the assumption that the latent confounder between $W$ and $Y$ is independent of the set of measured covariates, but this assumption may be violated in many real applications. Furthermore, AutoIV requires $S \perp\!\!\!\perp Y | W$, which is more strict than DIV.VAE. Hence, AutoIV does not solve the same

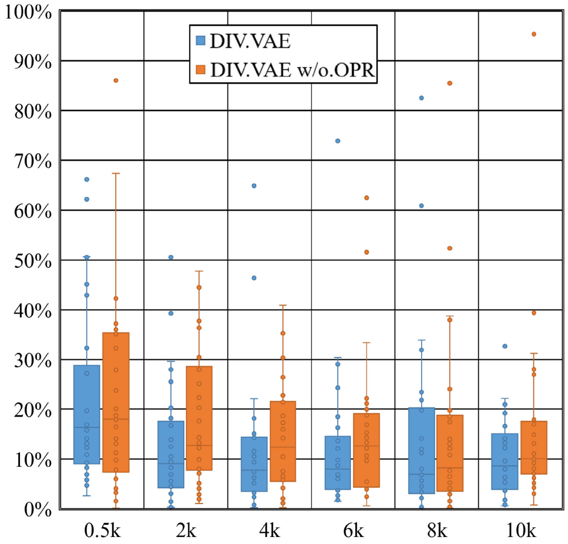


Fig. 13. Ablation study of DIV.VAE on synthetic datasets, where the horizontal axis represents the sample size and the vertical axis represents the estimation bias (%).

problem as DIV.VAE does and is not compared in the experiments. Different from the abovementioned IV methods, we focus on learning a valid IV representation from data without nominating a valid IV from domain knowledge.

### *B. Data-Driven IV-Based Method Without a Given IV*

In the absence of a given IV, a few data-driven methods have been proposed for finding valid IVs [13] or synthesizing IVs [48], [49] or eliminating the effect of invalid IVs by using statistical analysis [14], [50], [51]. For example, IV.Tetrad [13] uses the tetrad constraint to perform statistical tests for discovering pairs of valid IVs. Kuang et al. [49] proposed the Ivy method to combine IV candidates as a summary IV for identifying all invalid IVs or dependencies. Kang et al. [14] proposed the sisVIVE method to estimate causal effects when at least half of the covariates are valid IVs (i.e., majority assumption). Hartford et al. [51] developed a ModeIV algorithm by employing a deep learning-based IV estimator [34] under the majority assumption. Both sisVIVE and ModeIV rely on the majority assumption, but it is difficult to verify the majority assumption and this limits their applications. Unlike this type of data-driven method, DIV.VAE only needs an SIV and makes use of VAE to recover IV information.

### *C. Deep Learning (Including VAE)-Based Causal Effect Estimation*

The existing VAE-based causal effect estimators [19], [20], [27] simply assume no latent confounders, and thus, they are not for dealing with latent confounders. Moreover, this type of VAE-based estimator relies on an impractical assumption [44], i.e., they require all covariates to be measured as the proxy variables of the latent confounders or latent representations [27]. Under the unconfoundedness assumption, researchers focus on designing deep learning models for estimating causal effects from observational data, e.g., balance learning representation (BLR) neutral network [52], counterfactual regression (CFR) [53], and generative adversarial nets for estimating individualized treatment effects (GANITE) [54]. However, none of these deep learning-based estimators can obtain an unbiased estimation of the causal effect of $W$ on $Y$ in the presence of a latent confounder between $(W, Y)$. Hence, this work is the first one to use the VAE model in learning and disentangling the latent IV representation from the latent representation of the measured covariates for causal effect estimation from data without the unconfoundedness assumption.

When there is a latent confounder between $(W, Y)$, the causal effect of $W$ on $Y$ is nonidentifiable with covariate adjustment [1], [55] and these deep learning methods do not work as they are based on covariate adjustment. Our method takes the IV approach, a practical way to address this challenging problem very well.

## VI. Conclusion

Causal effect estimation from data with latent variables is crucial for many real-world applications, but there is a lack of effective data-driven methods for dealing with latent confounders. In this work, we make a connection between SIVs studied in the causal inference and statistics communities and the VAE model widely used in the machine learning community for latent representation learning. This connection has provided the theoretical guarantees for us to develop the DIV.VAE method to learn the latent IV representation through VAE-based disentangled representation learning. This, in turn, enables us to leverage the IV approach to obtain unbiased causal effect estimation from data in the presence of latent confounders. To the best of authors' knowledge, this is the first work to establish a link between generative modeling and the IV approach. Extensive experiments on synthetic and real-world datasets demonstrate that DIV.VAE is very effective in estimating the average causal effect from data with latent variables. We believe that the findings presented in this work have the potential to significantly improve the real-world applications of the IV approach for inferring unbiased causal effects from data with latent variables.

In our future work, we plan to explore connections between balanced representation learning [8], [53], proximal causal learning [18], [56], and latent IV representation learning using the disentanglement technique presented in this work. Furthermore, we will extend our DIV.VAE to applications in recommendation systems [57], natural language processing [58], and other areas [59].

## Acknowledgment

The authors thank the action editors and the reviewers for their invaluable comments and suggestions.

## References

[1] J. Pearl, *Causality*. Cambridge, U.K.: Cambridge Univ. Press, 2009.
[2] M. A. Hernán and J. M. Robins, *Causal Inference*. Boca Raton, FL, USA: CRC Press, 2010.
[3] T. C. Chalmers et al., "A method for assessing the quality of a randomized control trial," *Controlled Clin. Trials*, vol. 2, no. 1, pp. 31–49, May 1981.

[4] P. Spirtes et al., *Causation, Prediction, and Search*. Cambridge, MA, USA: MIT Press, 2000.

[5] W. Chen, R. Cai, K. Zhang, and Z. Hao, "Causal discovery in linear non-Gaussian acyclic model with multiple latent confounders," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 7, pp. 2816–2827, Jul. 2022.

[6] D. Cheng, J. Li, L. Liu, K. Yu, T. D. Le, and J. Liu, "Toward unique and unbiased causal effect estimation from data with hidden variables," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 34, no. 9, pp. 6108–6120, Sep. 2023.

[7] M. A. Hernán and J. M. Robins, "Instruments for causal inference: An epidemiologist's dream?" *Epidemiology*, vol. 17, no. 4, pp. 360–372, 2006.

[8] G. W. Imbens and D. B. Rubin, *Causal Inference in Statistics, Social, and Biomedical Sciences*. Cambridge, U.K.: Cambridge Univ. Press, 2015.

[9] J. D. Angrist and G. W. Imbens, "Two-stage least squares estimation of average causal effects in models with variable treatment intensity," *J. Amer. Stat. Assoc.*, vol. 90, no. 430, pp. 431–442, Jun. 1995.

[10] C. Brito and J. Pearl, "Generalized instrumental variables," in *Proc. 18th Conf. Uncertainty Artif. Intell.*, 2002, pp. 85–93.

[11] A. Sjolander and T. Martinussen, "Instrumental variable estimation with the R package ivtools," *Epidemiol. Methods*, vol. 8, no. 1, pp. 1–20, Dec. 2019.

[12] J. Pearl, "On the testability of causal models with latent and instrumental variables," in *Proc. 11th Conf. Uncertainty Artif. Intell.*, 1995, pp. 435–443.

[13] R. Silva and S. Shimizu, "Learning instrumental variables with structural and non-Gaussianity assumptions," *J. Mach. Learn. Res.*, vol. 18, no. 1, pp. 120:1–120:49, Jan. 2017.

[14] H. Kang, A. Zhang, T. T. Cai, and D. S. Small, "Instrumental variables estimation with some invalid instruments and its application to Mendelian randomization," *J. Amer. Stat. Assoc.*, vol. 111, no. 513, pp. 132–144, Jan. 2016.

[15] M. Kuroki and Z. Cai, "Instrumental variable tests for directed acyclic graph models," in *Proc. Int. Conf. Artif. Intell. Statist.*, 2005, pp. 190–197.

[16] F. Xie, Y. He, Z. Geng, Z. Chen, R. Hou, and K. Zhang, "Testability of instrumental variables in linear non-Gaussian acyclic causal models," *Entropy*, vol. 24, no. 4, p. 512, Apr. 2022.

[17] M. R. Montgomery, M. Gragnolati, K. A. Burke, and E. Paredes, "Measuring living standards with proxy variables," *Demography*, vol. 37, no. 2, pp. 155–174, May 2000.

[18] W. Miao, Z. Geng, and E. J. Tchetgen Tchetgen, "Identifying causal effects with proxy variables of an unmeasured confounder," *Biometrika*, vol. 105, no. 4, pp. 987–993, Dec. 2018.

[19] N. Hassanpour and R. Greiner, "Learning disentangled representations for counterfactual regression," in *Proc. 8th Int. Conf. Learn. Represent.*, 2020, pp. 1–11.

[20] W. Zhang, L. Liu, and J. Li, "Treatment effect estimation with disentangled latent factors," in *Proc. AAAI Conf. Artif. Intell.*, May 2021, pp. 10923–10930.

[21] Y. Cui and E. T. Tchetgen, "A semiparametric instrumental variable approach to optimal treatment regimes under endogeneity," *J. Amer. Stat. Assoc.*, vol. 116, no. 533, pp. 162–173, Jan. 2021.

[22] V. Chernozhukov et al., "Double/debiased machine learning for treatment and structural parameters," *Econometrics J.*, vol. 21, no. 1, pp. C1–C68, Feb. 2018.

[23] V. Syrgkanis et al., "Machine learning estimation of heterogeneous treatment effects with instruments," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 15167–15176.

[24] S. Greenland, "An introduction to instrumental variables for epidemiologists," *Int. J. Epidemiol.*, vol. 29, no. 4, pp. 722–729, Aug. 2000.

[25] E. P. Martens et al., "Instrumental variables: Application and limitations," *Epidemiology*, vol. 17, no. 3, pp. 260–267, 2006.

[26] T. Richardson and P. Spirtes, "Ancestral graph Markov models," *Ann. Statist.*, vol. 30, no. 4, pp. 962–1030, 2002.

[27] C. Louizos et al., "Causal effect inference with deep latent-variable models," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 6446–6456.

[28] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," in *Proc. 2nd Int. Conf. Learn. Represent.*, 2014, pp. 1–14.

[29] D. P. Kingma and M. Welling, "An introduction to variational autoencoders," *Found. Trends Mach. Learn.*, vol. 12, no. 4, pp. 307–392, 2019.

[30] P. Xie, W. Wu, Y. Zhu, and E. P. Xing, "Orthogonality-promoting distance metric learning: Convex relaxation and theoretical analysis," in *Proc. 35th Int. Conf. Mach. Learn.*, 2018, pp. 5399–5408.

[31] F. Locatello et al., "Challenging common assumptions in the unsupervised learning of disentangled representations," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2019, pp. 4114–4124.

[32] I. Khemakhem, D. P. Kingma, R. Monti, and A. Hyvarinen, "Variational autoencoders and nonlinear ICA: A unifying framework," in *Proc. 23rd Int. Conf. Artif. Intell. Statist. (AISTATS)*, 2020, pp. 2207–2217.

[33] S. Athey, J. Tibshirani, and S. Wager, "Generalized random forests," *Ann. Statist.*, vol. 47, no. 2, pp. 1148–1178, Apr. 2019.

[34] J. Hartford, G. Lewis, K. Leyton-Brown, and M. Taddy, "Deep IV: A flexible approach for counterfactual prediction," in *Proc. 34th Int. Conf. Mach. Learn.*, 2017, pp. 1414–1423.

[35] A. Paszke et al., "PyTorch: An imperative style, high-performance deep learning library," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, Dec. 2019, pp. 8024–8035.

[36] E. Bingham et al., "Pyro: Deep universal probabilistic programming," *J. Mach. Learn. Res.*, vol. 20, no. 28, pp. 28:1–28:6, 2019.

[37] K. Battocchi et al. (2019). *EconML: A Python Package for ML-Based Heterogeneous Treatment Effects Estimation*. [Online]. Available: https://github.com/microsoft/EconML

[38] D. Cheng, J. Li, L. Liu, J. Zhang, T. D. Le, and J. Liu, "Ancestral instrument method for causal inference without complete knowledge," in *Proc. 31st Int. Joint Conf. Artif. Intell.*, Jul. 2022, pp. 4843–4849.

[39] T. Martinussen, D. N. Sørensen, and S. Vansteelandt, "Instrumental variables estimation under a structural cox model," *Biostatistics*, vol. 20, no. 1, pp. 65–79, Jan. 2019.

[40] D. Card, "Using geographic variation in college proximity to estimate the return to schooling," Nat. Bur. Econ. Res., Cambridge, MA, USA, NBER Working Papers 4483, 1993.

[41] K. Sachs, O. Perez, D. Pe'er, D. A. Lauffenburger, and G. P. Nolan, "Causal protein-signaling networks derived from multiparameter single-cell data," *Science*, vol. 308, no. 5721, pp. 523–529, Apr. 2005.

[42] M. Verbeek, *A Guide To Modern Econometrics*. Hoboken, NJ, USA: Wiley, 2008.

[43] Y. LeCun, C. Cortes, and C. Burges. (2010). *MNIST Handwritten Digit Database*. ATT Labs. [Online]. Available: http://yann.lecun.com/exdb/mnist

[44] S. Rissanen and P. Marttinen, "A critical look at the consistency of causal estimation with deep latent variable models," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 4207–4217.

[45] J. Yuan et al., "Auto IV: Counterfactual prediction via automatic instrumental variable decomposition," *ACM Trans. Knowl. Discovery Data*, vol. 16, no. 4, pp. 74:1–74:20, Aug. 2022.

[46] A. Bennett, N. Kallus, and T. Schnabel, "Deep generalized method of moments for instrumental variable analysis," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 3559–3569.

[47] R. Singh, M. Sahani, and A. Gretton, "Kernel instrumental variable regression," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 4595–4607.

[48] S. Burgess and S. G. Thompson, "Use of allele scores as instrumental variables for Mendelian randomization," *Int. J. Epidemiol.*, vol. 42, no. 4, pp. 1134–1144, Aug. 2013.

[49] Z. Kuang et al., "Ivy: Instrumental variable synthesis for causal inference," in *Proc. 23rd Int. Conf. Artif. Intell. Statist.*, 2020, pp. 398–410.

[50] Z. Guo, H. Kang, T. Tony Cai, and D. S. Small, "Confidence intervals for causal effects with invalid instruments by using two-stage hard thresholding with voting," *J. Roy. Stat. Soc. B, Stat. Methodol.*, vol. 80, no. 4, pp. 793–815, Sep. 2018.

[51] J. S. Hartford et al., "Valid causal inference with (some) invalid instruments," in *Proc. 38th Int. Conf. Mach. Learn.*, 2021, pp. 4096–4106.

[52] F. Johansson, U. Shalit, and D. Sontag, "Learning representations for counterfactual inference," in *Proc. Int. Conf. Mach. Learn.*, 2016, pp. 3020–3029.

[53] U. Shalit, F. D. Johansson, and D. Sontag, "Estimating individual treatment effect: Generalization bounds and algorithms," in *Proc. 34th Int. Conf. Mach. Learn.*, 2017, pp. 3076–3085.

[54] J. Yoon, J. Jordon, and M. Van Der Schaar, "GANITE: Estimation of individualized treatment effects using generative adversarial nets," in *Proc. 6th Int. Conf. Learn. Represent.*, 2018, pp. 1–22.

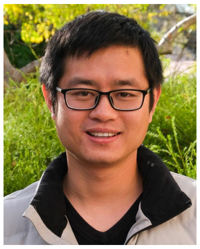
[55] D. Cheng, J. Li, L. Liu, J. Liu, and T. D. Le, "Data-driven causal effect estimation based on graphical causal modelling: A survey," *ACM Comput. Surv.*, vol. 56, no. 5, pp. 1–37, May 2024.

[56] E. J. T. Tchetgen, A. Ying, Y. Cui, X. Shi, and W. Miao, "An introduction to proximal causal learning," 2020, *arXiv:2009.10982*.

[57] Y. Wang, D. Liang, L. Charlin, and D. M. Blei, "Causal inference for recommender systems," in *Proc. 14th ACM Conf. Recommender Syst.*, Sep. 2020, pp. 426–431.

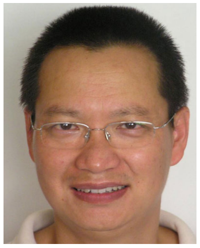
[58] A. Feder et al., "Causal inference in natural language processing: Estimation, prediction, interpretation and beyond," *Trans. Assoc. Comput. Linguistics*, vol. 10, pp. 1138–1158, Oct. 2022.

[59] B. Schölkopf et al., "Toward causal representation learning," *Proc. IEEE*, vol. 109, no. 5, pp. 612–634, May 2021.



**Debo Cheng** received the Ph.D. degree in computer and information science from the University of South Australia (UniSA), Adelaide, SA, Australia, in 2021.

He currently serves as a Research Fellow with UniSA STEM, UniSA. His main interests include data mining, machine learning, causal inference, and causal machine learning.

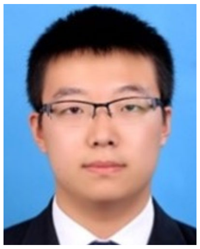


**Jiuyong Li** (Member, IEEE) received the Ph.D. degree in computer science from Griffith University, Brisbane, QLD, Australia, in 2002.

He is currently a Professor with the University of South Australia (UniSA), Adelaide, SA, Australia. His research work has been supported by eight Australian Research Council Discovery projects and he has led several industrial and applied projects. His main research interests include data mining, causal discovery, privacy and fairness, and bioinformatics.

**Lin Liu** received the Ph.D. degree in computer systems engineering from the University of South Australia (UniSA), Adelaide, SA, Australia, in 2006.

She is currently a Professor with UniSA. Her research interests include data mining, causal inference, and bioinformatics.



**Ziqi Xu** received the M.S. degree in computer science and the Ph.D. degree in STEM from the University of Adelaide (UniSA), Adelaide, SA, Australia, in 2020 and 2024, respectively.

He is currently a Lecturer with the School of Computing Technologies, RMIT University, Melbourne, VIC, Australia. His research interests include data mining, causal inference, generative model, and fairness.

**Weijia Zhang** received the B.Sc. degree in mathematics and the M.Sc. degree in computer science from Nanjing University, Nanjing, China, in 2011 and 2014, respectively, and the Ph.D. degree in computer science from the University of South Australia, Adelaide, SA, Australia, in 2018.

He is currently a Lecturer with the School of Information and Physical Sciences, University of Newcastle, Callaghan, NSW, Australia. His main research interests include causal inference and machine learning.

**Jixue Liu** received the Ph.D. degree in computer science from the University of South Australia (UniSA), Adelaide, SA, Australia, in 2001.

He is currently an Associate Professor with UniSA. He has authored widely in databases and artificial intelligence. His work covers the topics of integrity constraint discovery, data analytics in texts and time series, entity linking, fairness computing, privacy in data, extensible markup language (XML) functional dependencies, and data integration and transformation.

**Thuc Duy Le** is currently an Associate Professor with the University of South Australia (UniSA), Adelaide, SA, Australia. His research focuses on the development of causal discovery methods and their applications in bioinformatics.

Dr. Le is currently a DECRA Fellow and was also a National Health and Medical Research Council (NHMRC) ECR Fellow from 2017 to 2019. He has served as an academic editor and a reviewer for conferences in data mining and journals in bioinformatics.