2022 IEEE 25th International Conference on Computer Supported Cooperative Work in Design (CSCWD) | 978-1-6654-0527-0/22/$31.00 ©2022 IEEE | DOI: 10.1109/CSCWD54268.2022.9776065

Learning Disentangled Latent Factors for Individual Treatment Effect Estimation Using Variational Generative Adversarial Nets

Qingsen Bao
School of Computer Science
Nanjing University of Posts and Telecommunications
Nanjing, China
sdfcbqs@126.com

Zeyong Mao
School of Computer Science
Nanjing University of Posts and Telecommunications
Nanjing, China
zeyongmao@gmail.com

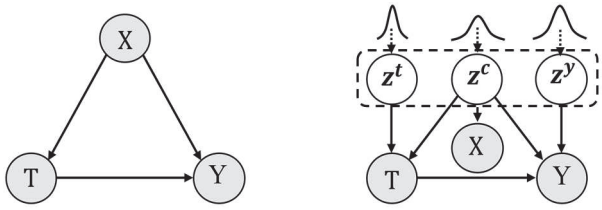
Lei Chen
Jiangsu Key Laboratory of Big Data Security and Intelligent Processing
Nanjing University of Posts and Telecommunications
Nanjing, China
chenlei@njupt.cn

***Abstract*—Estimating individual treatment effect (ITE) is a challenging task due to the need for individual potential outcomes to be learned from biased data and counterfactuals are inherently unobservable. Some researchers propose to use generative adversarial approaches to infer the counterfactual outcomes based on the distribution of factual outcomes. However, these methods assume that complete confounding factors are observed, and simply treat all observed variables as confounding factors, ignoring identification of possible instrumental factors and adjustment factors, which will bring large deviation to ITE estimation when facing biased data. To address these issues, we propose a novel Variational Generative Adversarial Nets for ITE estimation by designing the collaborative learning strategy with Variational AutoEncoder (VAE) and Generative Adversarial Nets (GAN). Specifically, we employ VAE to infer the latent representations of observed variables to access complete latent factors while using GAN to infer unseen counterfactual outcomes and guide VAE for disentangling these latent factors into three sets corresponding to the instrumental, confounding, and adjustment factors. Then the disentangled latent confounding factors can be used to further control data bias using an adaptive weighting scheme. Extensive experiments on real and synthetic data demonstrate learning disentangled latent factors for ITE estimation is effective, and our method has excellent performance even with high data bias.**

Keywords—individual treatment effect, biased data, counterfactual outcomes, latent confounding factors, Variational AutoEncoder, Generative Adversarial Nets

I. INTRODUCTION

The collaboration of causal inference and machine learning is currently attracting the attention of many researchers, some using ideas from causal inference to assist machine learning tasks like [1], [2], and others applying machine learning methods to solve causal inference problems such as [3], [4]. We focus on the latter of estimating individual treatment effect (ITE) in causal inference. For example, in medicine, accurately estimating the treatment effect for each patient helps doctors decide which treatment (e.g., taking the drug or not) is appropriate for the patient [5]. The gold standard for evaluating treatment effect is randomized controlled trials (RCTs), yet RCTs are expensive and sometimes unethical. Therefore researchers have shifted their focus on treatment effect especially ITE estimation from observed data. The ITE estimation compares the differences between potential outcomes for an individual under different treatments. But in observed data, only one outcome can be observed which is called factual outcome, and unobserved outcomes are called counterfactual outcomes. So learning the ITE estimation model requires answering counterfactual questions (e.g., would the patient have recovered more quickly if he had received different treatment at that time?). ITE estimation is different from the standard supervised learning problem. Firstly the counterfactual outcomes are not observed, so the labels of data are incomplete. Second, unlike RCTs, observational studies have data bias due to the presence of confounding factors that affect both treatment and outcome. Fig. 1 (a) illustrates the data bias termed selection bias $P(\mathrm{T}|\mathrm{X}) \neq P(\mathrm{T})$, i.e., the assignment of treatment is not random, depends on the observed variables.



(a) No-disentangled latent factors (b) Disentangled latent factors

Fig. 1. We denote the graphical model among the observed variables X, the treatment T, and the outcome Y. $\boldsymbol{z}^t$, $\boldsymbol{z}^c$, and $\boldsymbol{z}^y$ are latent instrumental, confounding, and adjustment factors sampled from three Gaussian distributions respectively.

There are classical works to solve the problem. [6] uses a matching method to pair treated (e.g., taking the drug) and control (e.g., no-taking the drug) patients with similar features, and [7] uses propensity scoring weighting to account for the selection bias. More recent works focus on using representation learning to balance confounders and solve this question like [8-11]. Estimation of ITE using Generative Adversarial Nets (GANITE) [10] is an unusual work, which treats the ITE estimation as an incomplete labels questions. Missing counterfactual labels can be generated using generative adversarial nets (GAN). However, GANITE ignores the identification of confounding and non-confounding factors, treats all observed variables as confounding factors to reduce the influence caused by selection bias. [12] and [13] has been demonstrated that controlling non-confounders (e.g., instrumental factors) will increase the error of estimating ITE. Moreover, GANITE assumes that the observed variables already contain complete confounders, but the assumption is difficult to

satisfy in real applications, therefore greatly limiting the scenarios in which GANITE can be used.

To address the limitations of GANITE for estimating ITE, a collaborative learning framework is termed Variational Generative Adversarial Nets for ITE estimation (VGANITE). As shown in Fig. 1(b), the observed variables are considered as proxies for latent factors which can be decomposed into instrumental factors that only affect treatment assignment, confounding factors that affect both treatment assignment and outcome, and adjustment factors that only affect the outcome. VGANITE incorporates the framework of variational autoencoder (VAE) and generative adversarial nets (GAN) to learn disentangled latent factors then generate the counterfactual outcomes. Our contributions can be summarized as follows:

- Explicit identification of the latent factors $(\boldsymbol{z}^t, \boldsymbol{z}^c, \boldsymbol{z}^y)$ in observational data can better handle selection bias and achieve better performance in estimating ITE;
- The collaborative learning strategy with VAE and GAN can infer the latent representations of observed variables to access the unobserved confounding factors;
- We validate the effectiveness of the proposed VGANITE on a wide range of real and synthetic datasets.

II. Related Work

The estimation of ITE from observed data has gradually attracted the attention of machine learning researchers. We believe previous studies can be divided into two main categories. First, to estimate ITE from observed data, most of the traditional methods either employ propensity scores including matching, stratification, weighting, and doubly robust or directly optimize sample weight in [14], [15]. However, these traditional methods are difficult to adapt to high-dimensional data scenarios. Secondly, representation learning is used in [16-19]. Some methods define ITE estimation from observed data as a domain adaptive scenario where the model is trained on the source domain (factual) data but need to perform well on the target domain (counterfactual) data. However, we cannot control that the distribution of source domain and target domain are consistent. The reason is precisely the selection bias ($P(\mathrm{T}|\mathrm{X}) \neq P(\mathrm{T})$). These methods learn a representation space in which the distribution of observed variables is consistent across different treatment groups and then learn a model to estimate ITE based on this consistent representation space. However, since confounding factors can affect both treatment assignment and outcome, these approaches need to balance between preserving predictive confounding and reducing biased confounding factors, resulting in the learned representation space having remaining selection bias. To solve this problem, [17] proposes a counterfactual regression method with importance sampling weights. But in [10], GANITE directly uses all observed variables to infer the counterfactual distribution using a GAN framework, then the ITE can be computed using a supervised approach.

However, existing most methods treat all observed variables as confounders to explain selection bias, ignoring the importance of identifying unconfounding factors. Moreover, most methods assume that there are no latent confounding factors but this assumption is difficult to satisfy. Recently, variable decomposition based on representation learning has been used for ITE estimation in [20-22]. But [20] only considers decomposing confounding and adjustment factors, does not consider the identification of instrumental factors.

And they also assume complete confounding factors are observed. Another work that is closely related to ours is the Disentangled Variational Autoencoder (TEDVAE) [22] improves on Causal Effect Variational Autoencoder (CEVAE) [23], which considers the presence of latent factors, but did not attempt to explain the distribution of counterfactual outcomes and address selection bias.

III. Method

A. Notations, Problems, and Assumptions

Let X denote the s-dimensional feature space, Y the set of potential outcomes, and T the set of possible treatments. For each sample indexed by i, we have its context features vector $\boldsymbol{x}_i \in \mathrm{X}$, its treatment $t_i \in \mathrm{T}$, and the potential outcome $y_i(t_i) \in \mathrm{Y}$ as an outcome of choosing the treatment t_i. In our context, we are interested in the case of a binary treatment (i.e., $t_i \in \{0, 1\}$). In binary treatment, for sample i the y_i^f is denoted as the observed factual outcome, and y_i^{cf} is denoted as the counterfactual outcome, which $y_i^f = t_i y_i(1) + (1 - t_i) y_i(0)$, $y_i^{cf} = (1 - t_i) y_i(1) + t_i y_i(0)$. Based on the potential outcomes of sample i, the $ITE_i = y_i(1) - y_i(0)$.

There are some problems with estimating ITE from observed data. First, for training sample i the counterfactual outcome y_i^{cf} cannot be observed. Second, the assignment of treatment t_i depends on the observed features vector $\boldsymbol{x}_i$, which leads to the domain adaptive problem. In particular, the set of observed samples is $D^f = \{(\boldsymbol{x}_i, t_i), y_i^f\}_{i=1}^n$. However, computing ITE requires inferring the counterfactual outcome on the set $D^{cf} = \{(\boldsymbol{x}_i, 1 - t_i), y_i^{cf}\}_{i=1}^n$. As shown in Fig. 1(a), previous methods treat observed variables as confounders and define $P^f(\mathrm{X}, \mathrm{T})$ as the factual distribution and $P^{cf}(\mathrm{X}, \mathrm{T})$ as the counterfactual distribution. Specially, $P^f(\mathrm{X}, \mathrm{T}) = P(\mathrm{X}) \cdot P^f(\mathrm{T}|\mathrm{X})$, $P^{cf}(\mathrm{X}, \mathrm{T}) = P(\mathrm{X}) \cdot P^{cf}(\mathrm{T}|\mathrm{X})$. The difference between the factual and the counterfactual distribution on lies precisely in the treatment assignment $P(\mathrm{T}|\mathrm{X})$. If T and X are independent, the distribution of factual and counterfactual are equal. In our methods, as shown in Fig. 1(b), the latent confounding factors are learned, so our method can control the impact of selection bias more precisely.

In the paper, we assume the following three assumptions for ITE estimation in [14]:

- *(SUTVA): The Stable Unit Treatment Value Assumption requires that the potential outcomes of a sample are not affected by the treatment of others.*
- *(Overlap): Every sample has a non-zero probability to receive either treatment or control when given the observed variables:* $0 < P(t = 1|\boldsymbol{x}) < 1$.
- *(Latent-confounder): Given latent confounding factors, treatment assignment does not affect potential outcomes:* $t \perp\!\!\!\perp (y(0), y(1))|\boldsymbol{z}^c$.

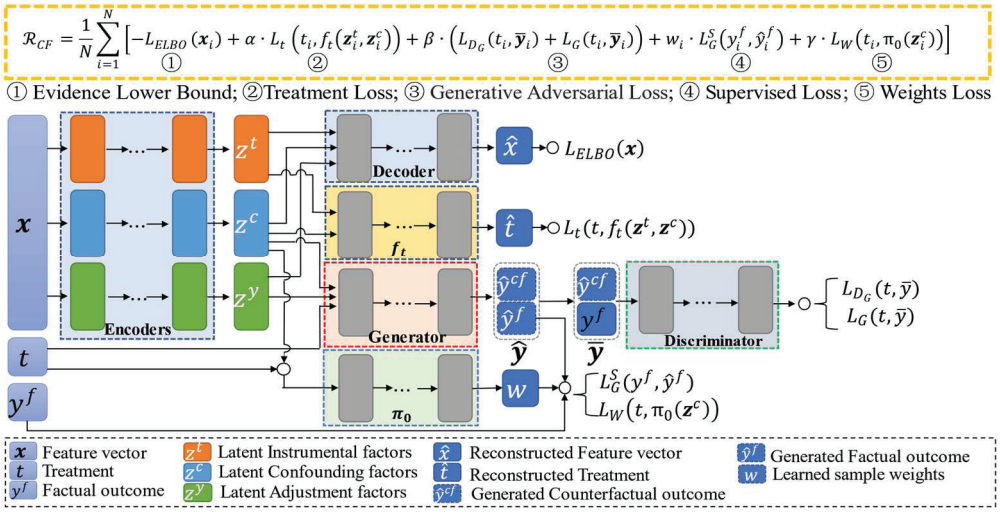


Fig 2. The overall architecture of the disentangling latent factors and generating counterfactual outcomes

B. *Disentangling latent factors*

In this work, our goal is based on observed features vector $\boldsymbol{x}$ to infer the posterior distribution of latent representations $\boldsymbol{z} = \{\boldsymbol{z}^t, \boldsymbol{z}^c, \boldsymbol{z}^y\}$, and disentangling $\boldsymbol{z}$ corresponding to latent instrumental $\boldsymbol{z}^t$, confounding $\boldsymbol{z}^c$, and adjustment $\boldsymbol{z}^y$ factors. As shown in Fig. 2, the features vector $\boldsymbol{x}$ is disentangled into $\boldsymbol{z}^t, \boldsymbol{z}^c$, and $\boldsymbol{z}^y$ using three independent encoders including $q_{\phi_t}(\boldsymbol{z}^t|\boldsymbol{x}), q_{\phi_c}(\boldsymbol{z}^c|\boldsymbol{x}), q_{\phi_y}(\boldsymbol{z}^y|\boldsymbol{x})$. Then three latent factors are used to reconstruct $\boldsymbol{x}$ by the decoder $p_\theta(\boldsymbol{x}|\boldsymbol{z}^t, \boldsymbol{z}^c, \boldsymbol{z}^y)$ following the VAE framework. The prior distributions $p(\boldsymbol{z}^t)$, $p(\boldsymbol{z}^c)$, $p(\boldsymbol{z}^y)$ are chosen to be Gaussian distributions.

$$p(\boldsymbol{z}^t) = \prod_{j=1}^{D_t} \mathcal{N}(z^{tj}|0,1); \quad p(\boldsymbol{z}^c) = \prod_{j=1}^{D_c} \mathcal{N}(z^{cj}|0,1); \quad p(\boldsymbol{z}^y) = \prod_{j=1}^{D_y} \mathcal{N}(z^{yj}|0,1) \tag{1}$$

where D_t, D_c, and D_y denote the dimensionality of the latent instrument, confounding, and adjustment factors.

In encoders, the variational approximations of the posteriors are defined as:

$$q_{\phi_t}(\boldsymbol{z}^t|\boldsymbol{x}) = \prod_{j=1}^{D_t} \mathcal{N}\left(\mu = \hat{\mu}_{tj}, \sigma^2 = \hat{\sigma}_{tj}^2\right);$$
$$q_{\phi_c}(\boldsymbol{z}^c|\boldsymbol{x}) = \prod_{j=1}^{D_c} \mathcal{N}\left(\mu = \hat{\mu}_{cj}, \sigma^2 = \hat{\sigma}_{cj}^2\right);$$
$$q_{\phi_y}(\boldsymbol{z}^y|\boldsymbol{x}) = \prod_{j=1}^{D_y} \mathcal{N}\left(\mu = \hat{\mu}_{yj}, \sigma^2 = \hat{\sigma}_{yj}^2\right) \tag{2}$$

where $\hat{\mu}_{tj}, \hat{\mu}_{cj}, \hat{\mu}_{yj}$ and $\hat{\sigma}_{tj}^2, \hat{\sigma}_{cj}^2, \hat{\sigma}_{yj}^2$ denote the mean and variance of the parametric Gaussian distribution.

Consistent with the standard VAE [24], the parameters of encoders and decoder can be optimized by maximizing the evidence lower bound (ELBO):

$$L_{ELBO} = E_{q_{\phi_t}, q_{\phi_c}, q_{\phi_y}}[log p_\theta(\boldsymbol{x}|\boldsymbol{z}^t, \boldsymbol{z}^c, \boldsymbol{z}^y)] - KL(q_{\phi_t}(\boldsymbol{z}^t|\boldsymbol{x})||p_{\theta_t}(\boldsymbol{z}^t)) - KL(q_{\phi_c}(\boldsymbol{z}^c|\boldsymbol{x})||p_{\theta_c}(\boldsymbol{z}^c)) - KL(q_{\phi_y}(\boldsymbol{z}^y|\boldsymbol{x})||p_{\theta_y}(\boldsymbol{z}^y)) \tag{3}$$

As shown in Fig. 1(b), the latent instrumental factors $\boldsymbol{z}^t$and confounding factors $\boldsymbol{z}^c$ are associated with treatment T, and the latent confounding factors $\boldsymbol{z}^c$ and adjustment factors $\boldsymbol{z}^y$ are associated with outcome Y. To have a better separation of latent factors, we assume that the separated $\boldsymbol{z}^t$, $\boldsymbol{z}^c$ can predict t well, and $\boldsymbol{z}^c$, $\boldsymbol{z}^y$ can predict y well. We use a network $f_t(\boldsymbol{z}^t, \boldsymbol{z}^c)$ to inference the posterior distribution $p(\hat{t}|\boldsymbol{z}^t, \boldsymbol{z}^c) = Bern(f_t(\boldsymbol{z}^t, \boldsymbol{z}^c))$, where f_t is a neural network with logistic activation function, and $Bern(p)$ denotes the standard Bernoulli distribution with parameter p. The loss function of f_t is named as treatment loss:

$$L_t(t_i, f_t(\boldsymbol{z}_i^t, \boldsymbol{z}_i^c)) = -\left[t_i \, log\left(f_t(\boldsymbol{z}_i^t, \boldsymbol{z}_i^c)\right) + (1 - t_i) \, log\left(1 - f_t(\boldsymbol{z}_i^t, \boldsymbol{z}_i^c)\right)\right] \tag{4}$$

Similarly, a framework of GAN is used to infer the factual outcome distribution $p(y^f|t, \boldsymbol{z}^c, \boldsymbol{z}^y)$. We will introduce the block detailly in the next section.

C. *Counterfactual inference*

In this block, we use GANITE as the baseline model. we replace the input $\boldsymbol{x}$ of GANITE's counterfactual generator with the learning disentangled latent confounding factors $\boldsymbol{z}^c$ and the latent adjustment factors $\boldsymbol{z}^y$. As well as a sample-weighted approach is used to balance the confounding factors thus reducing the selection bias.

Counterfactual generator ($\boldsymbol{G}$): The counterfactual generator $\boldsymbol{G}$ uses $\boldsymbol{z}^c, \boldsymbol{z}^y$, and t to generator the potential outcome vector $\hat{\boldsymbol{y}}$ that contains both predicted factual ($\hat{y}^f$) and counterfactual ($\hat{y}^{cf}$) outcomes, where $\boldsymbol{z}^c, \boldsymbol{z}^y$ are not random noise compared with standard GAN. We use η to record the subscript of the factual outcome in $\hat{\boldsymbol{y}}$ (e.g., if $t_i = 1$, the factual outcome $y_i^f = y_i(1)$ and $\eta = 1$, on the contrary, the $\eta = 0$). Then we replace $\hat{y}(\eta)$ with factual outcome y^f, which is defined as $\bar{\boldsymbol{y}}$ containing the counterfactual outcome $\hat{y}^{cf}$ generated by $\boldsymbol{G}$ and the factual outcome y^f. As shown in Fig. 2 (Generator). And $\bar{\boldsymbol{y}}$ is used as input to the counterfactual discriminator.

Counterfactual discriminator ($\boldsymbol{D_G}$): The input of the $\boldsymbol{D_G}$ is $\bar{\boldsymbol{y}}$, exporting the probability that the i-th component corresponding of $\bar{\boldsymbol{y}}$ is the factual outcome, equivalently the probability that $\eta = i$. This is different from the framework of

the standard GAN, where the discriminator is given a sample from two distributions and tries to determine which distribution it comes from. Here, the discriminator is given a sample that is composed of components from two different distributions, i.e., the factual outcome and the counterfactual outcome distribution, then the discriminator tries to determine which component comes from which distribution. It should be noted that "factual" and "counterfactual" do not correspond to specific treatments, for any given sample, any treatment may be the factual one. Therefore, it makes sense to try and push counterfactual outcomes from one sample toward the factual outcomes from other (similar) samples. As shown in Fig. 2 (Discriminator).

Based on the above analysis, we define the optimization function of the $\boldsymbol{D_G}$ and $\boldsymbol{G}$ as:

$$V_{CF}(t_i, \bar{\boldsymbol{y}}_i) = t_i \log(D_G(\bar{\boldsymbol{y}}_i)) + (1 - t_i)\, log(1 - D_G(\bar{\boldsymbol{y}}_i)] \quad (5)$$

where $\bar{\boldsymbol{y}}_i = [\hat{y}_i^{cf}, y_i^f]$.

We also introduce supervised loss to strengthen the fit of the factual outcome:

$$L_G^S(y_i^f, \hat{y}_i^f) = (y_i^f - \hat{y}_i^f)^2 \quad (6)$$

With the above two functions, $\boldsymbol{D_G}$ and $\boldsymbol{G}$ are iteratively optimized as follow:

$$\min_{D_G} -\sum_{i=1}^{k_G} V_{CF}(t_i, \bar{\boldsymbol{y}}_i) \quad (7)$$

$$\min_{G} \sum_{i=1}^{k_G} [L_G^S(y_i^f, \hat{y}_i^f) + \gamma \cdot V_{CF}(t_i, \bar{\boldsymbol{y}}_i)] \quad (8)$$

where k_G is the mini-batches and γ is a hyper-parameter. To optimize the training process, we let the $\boldsymbol{D_G}$ iteratively train for two rounds first while normalizing the inputs.

Moreover, we use importance sample weighting as [17] to reduce selection bias. A logistic regression model π_0 is learned using neural networks to learn sample weights, as shown in Fig. 2 (π_0 network), and the loss function names as weights loss:

$$L_w(t_i, \pi_0(\boldsymbol{z}_i^c)) = -[t_i\, log(\pi_0(\boldsymbol{z}_i^c)) + (1 - t_i)\, log(1 - \pi_0(\boldsymbol{z}_i^c))] \quad (9)$$

Unlike [17], we use latent confounders and treatment to infer sample weights:

$$w(\boldsymbol{z}_i^c, t_i) = 1 + \frac{P(t_i)}{1 - P(t_i)} \cdot \frac{1 - \pi_0(\boldsymbol{z}_i^c)}{\pi_0(\boldsymbol{z}_i^c)} \quad (10)$$

where $P(t_i)$ is the probability of $t_i = 1$ or $t_i = 0$ in a dataset.

Using the learned sample weights, the supervised loss (6) of the $\boldsymbol{G}$ is weighted as $w_i L_G^S(y_i^f, \hat{y}_i^f)$ to further control the selection bias. As shown in Fig. 2 the total loss function is :

$$\mathcal{L}_{CF} = -L_{ELBO} + \alpha \cdot L_t + \beta \cdot (L_{D_G} + L_G) + w \cdot L_G^S + \gamma \cdot L_W \quad (11)$$

where $L_{D_G} = -V_{CF}$, $L_G = V_{CF}$, and α, β, γ are parameters.

D. ITE estimation

After training the counterfactual generator, we use $\boldsymbol{G}$ to generate counterfactual outcomes and acquire the complete data $\hat{D} = (\boldsymbol{x_i}, y_i^f, \hat{y}_i^{cf}) = (\boldsymbol{x}_i, \bar{\boldsymbol{y}}_i)$, then use $\hat{D}$ for the ITE estimation.

ITE generator ($\boldsymbol{I}$): The ITE generator generates the potential outcomes $\tilde{\boldsymbol{y}}$, which contains the predicted potential

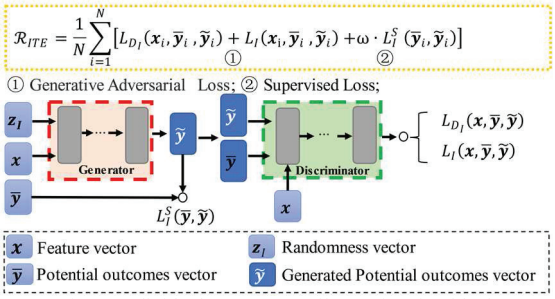


Fig. 3. Individual Treatment Effect estimation block

factual and counterfactual outcomes based on the variable $\boldsymbol{x}$ and randomness vector $\boldsymbol{z}_I \sim U(-1,1)$. As shown in Fig. 3.

ITE discriminator ($\boldsymbol{D_I}$): Like the standard conditional GAN [25] discriminator. The discriminator determines which of the inputs is the true potential outcomes ($\bar{\boldsymbol{y}}$). Similarly, we define the ITE estimation module to the optimization function:

$$V_{ITE}(\boldsymbol{x}_i, \bar{\boldsymbol{y}}_i, \tilde{\boldsymbol{y}}_i) = log(D_I(\boldsymbol{x}_i, \bar{\boldsymbol{y}}_i)) + log(\boldsymbol{1} - D_I(\boldsymbol{x}_i, \tilde{\boldsymbol{y}}_i)) \quad (12)$$

To better optimize the objective function, we introduce supervised loss $L_I^S(\bar{\boldsymbol{y}}_i, \tilde{\boldsymbol{y}}_i) = (\bar{\boldsymbol{y}}_i - \tilde{\boldsymbol{y}}_i)^2$. $\boldsymbol{D_I}$ and $\boldsymbol{I}$ are iteratively optimized as follow:

$$\min_{D_I} -\sum_{i=1}^{k_I} V_{ITE}(\boldsymbol{x}_i, \bar{\boldsymbol{y}}_i, \tilde{\boldsymbol{y}}_i) \quad (13)$$

$$\min_{I} \sum_{i=1}^{k_I} [L_S^I(\bar{\boldsymbol{y}}_i, \tilde{\boldsymbol{y}}_i) + \xi \cdot V_{ITE}(\boldsymbol{x}_i, \bar{\boldsymbol{y}}_i, \tilde{\boldsymbol{y}}_i)] \quad (14)$$

where k_I is the mini-batches and ξ is a hyper-parameter. As shown in Fig. 3, the loss of ITE estimation is:

$$\mathcal{L}_{ITE} = (L_{D_I}(\boldsymbol{x}_i, \bar{\boldsymbol{y}}_i, \tilde{\boldsymbol{y}}_i) + L_I(\boldsymbol{x}_i, \bar{\boldsymbol{y}}_i, \tilde{\boldsymbol{y}}_i)) + \omega \cdot L_I^S(\bar{\boldsymbol{y}}_i, \tilde{\boldsymbol{y}}_i) \quad (15)$$

where $L_{D_I} = -V_{ITE}$, $L_I = V_{ITE}$, and ω is a hyperparameter.

There above is the training phase, and in the testing phase, just the generator and discriminator of the ITE estimation block can be used. ITE estimation generator ($\boldsymbol{I}$) can infer the potential outcomes $\tilde{y}_i(0)$, $\tilde{y}_i(1)$ based on $\boldsymbol{x}_i$, knowing the potential outcomes, the $\widehat{ITE}_i$ can be directly obtained. And the discriminator ($\boldsymbol{D_I}$) can output the probability of how probable the estimation is equal to the true outcome, this probability can be used as a confidence measure, which is useful in many areas.

IV. Experiment

A. Experimental settings

We validate VGANITE in a variety of settings. On real datasets, we compare VGANITE with other methods to verify whether learning disentangled latent factors can achieve better performance in estimating ITE and Average Treatment Effect (ATE). These methods include **BLR**: balancing linear regression [8], **K-NN**: k-nearest neighbor [26], **BART**: Bayesian additive regression trees [27], **R-Forest**: random forests [28], **C-Forest**: causal forests [29], **TARNET**: treatment-agnostic representation network [16], **CFR_WASS** and **CFR_MMD**: counterfactual regression with Wasserstein and MMD distance [16], **CEVAE**: causal effect variational autoencoder [23], **DR-CFR**: disentangled representations for counterfactual regression [21], **TEDVAE**: treatment effect by disentangled variational autoencoder [22], and **GANITE** [10]. Second, we generate synthetic datasets to verify VGANITE whether has robust performance in estimating ITE under

TABLE I. ITE AND ATE ESTIMATION PERFORMANCE OF VGANITE ON TWO REAL DATASETS

Methods	Datasets							
	IHDP (*PEHE*)		IHDP (ϵ_{ATE})		Twins (*PEHE*)		Twins (ϵ_{ATE})	
	In-sample	Out-sample	In-sample	Out-sample	In-sample	Out-sample	In-sample	Out-sample
BLR	5.82±0.30	5.82±0.30	0.72±0.04	0.93±0.05	0.312±0.001	0.323±0.018	0.0057±0.0025	0.0334±0.0027
CEVAE	2.70±0.12	2.60±0.14	0.34±0.01	0.46±0.01	0.335±0.001	0.344±0.001	0.0411±0.0056	0.0563±0.0070
k-NN	2.11±0.10	4.11±0.20	0.14±0.01	0.9±0.05	0.333±0.009	0.345±0.005	0.0038±0.0039	**0.0051±0.0040**
BART	2.14±0.10	2.32±0.10	0.23±0.01	0.34±0.02	0.347±0.002	0.338±0.003	0.1206±0.0021	0.1265±0.0025
R-Forest	4.25±0.20	6.63±0.30	0.73±0.05	0.96±0.06	**0.306±0.002**	0.321±0.002	0.0049±0.0091	0.0080±0.0056
C-Forest	3.82±0.20	3.84±0.20	0.18±0.01	0.40±0.03	0.366±0.002	0.316±0.018	0.0286±0.0012	0.0335±0.0016
TARNET	0.88±0.02	0.92±0.02	0.26±0.01	0.28±0.01	0.319±0.005	0.315±0.002	0.0108±0.0015	0.0151±0.0025
CFR_WASS	0.71±0.02	0.76±0.02	0.25±0.01	0.27±0.01	0.315±0.007	**0.313±0.004**	0.0112±0.0025	0.0284±0.0046
CFR_MMD	0.71±0.02	0.79±0.02	0.28±0.01	0.30±0.01	0.318±0.005	0.316±0.008	0.0111±0.0010	0.0285±0.0012
DR-CFR	0.65±0.02	0.78±0.03	0.24±0.03	0.26±0.03	0.319±0.002	0.322±0.018	0.0060±0.0010	0.0090±0.0013
TEDVAE	0.65±0.11	0.70±0.14	0.12±0.02	0.16±0.02	0.320±0.002	0.321±0.025	0.0060±0.0012	0.0060±0.0016
GANITE	1.91±0.40	2.41±0.40	0.43±0.05	0.49±0.05	0.330±0.005	0.332±0.016	0.0058±0.0017	0.0089±0.0075
VGANITE	**0.64±0.03**	**0.67±0.03**	**0.11±0.02**	**0.15±0.02**	0.315±0.007	0.318±0.005	**0.0035±0.0013**	0.0062±0.0038

[a] The best results of all methods are highlighted with the bold font and the second with an underscore for each dataset

different selection biases. And we also validate the contribution of disentangled instrumental, confounding, and adjustment factors for ITE estimation, respectively.

B. Experiments on Real Datasets

IHDP: The original RCT data of the Infant Health and Development Program (IHDP) aims to evaluate the effect of specialist home visits on the future cognitive test scores of premature infants. The dataset included 747 samples (139 treated, 608 control) with 25 observed variables. We report the results from 100 experiments (with 63/27/10 proportion of train/validation/test splits) like [16], in which the true potential outcomes are simulated used to compute the true ITE.

Twins: The dataset comes from all twins birthed in the USA during 1989-1991. We define the $t = 1$ for heavier twins (and $t = 0$ for lighter twins). The outcome is defined as the 1-year mortality rate. For each twin pair, we obtained 30 observed variables. The final dataset is 11,400 pairs of twins. All features can be considered the same except for weight, so the ground truth of ITE is known in this dataset. To simulate an observational study, we selectively observe one of the two twins to create the selection bias as follows: $t|\mathbf{x}\sim Bern(\text{Sigmoid}(\mathbf{w}^T\mathbf{x}+n))$, $\mathbf{w}^T\sim U((-0.1,0.1)^{30\times 1})$ and $n\sim\mathcal{N}(0,0.1)$ like [10].

Similar to [10], [16], [17], we adopt the Precision in Estimation of Heterogeneous Effect which $PEHE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\left(\tilde{y}_i(1)-\tilde{y}_i(0)\right)-\left(y_i(1)-y_i(0)\right)\right)^2}$ as the ITE performance metric and we adopt the bias of the ATE which $\epsilon_{ATE} = \left|ATE-\widehat{ATE}\right|$ to evaluate ATE performance, where $ATE = \frac{1}{N}\sum_{i=1}^{N}(y_i(1)-y_i(0))$. We evaluate both in-sample and out-of-sample performance and report the results, the mean, and standard deviation in Table I.

In Table I, VGANITE achieved significant performance on the IHDP dataset. But on the Twins dataset, the performance is hardly improved compared with advanced methods. We believe this is due to that in the Twins dataset, each variable is discrete. Thus, the lack of information leads to insignificant performance improvement of VGANITE.

C. Experiments on Synthetic datasets

Referring to the [21], we generate our synthetic datasets according to the following procedure where we consider all latent factors $\boldsymbol{z}\in\{\boldsymbol{z}^t,\boldsymbol{z}^c,\boldsymbol{z}^y\}$ of dimension $m_t,m_c,m_y\in\{0,4,8\}$, where $\boldsymbol{z}\sim\mathcal{N}(0,1)$. We connect $\boldsymbol{z}^t,\boldsymbol{z}^c,\boldsymbol{z}^y$ to generate $\mathbf{x}$, $\boldsymbol{z}_t,\boldsymbol{z}_c$ to produce Ψ and $\boldsymbol{z}^c,\boldsymbol{z}^y$ to produce Φ. We then generate the selection bias by $t = Bern(1,1/(1+\exp(-\zeta r)))$, where $r = \Psi\cdot\theta,\theta\sim\mathcal{N}(0,1)^{m_t+m_c}$ and ζ determines the slope of the logistic curve. For potential outcomes, we define $y(0) = (\Phi\circ\Phi\circ\Phi+0.5)\cdot v^0/(m_c+m_y)+\epsilon$ and $y(1) = (\Phi\circ\Phi)\cdot v^1/(m_c+m_y)+\epsilon$, where v^0,v^1 are obtained from $\mathcal{N}(0,1)^{m_c+m_y}$ and ϵ is the noise variable sampled from $\mathcal{N}(0,0.1)$.$\circ$ is the symbol for element-wise (Hadamard) product. We consider generating all feasible datasets through $m_t,m_c,m_y\in\{0,4,8\}$. And we remove the set of (0,0,0), (4,0,0), (8,0,0), giving a total of 24 groups.

We control the size of selection bias by setting the size of ζ and use the latent factors dimensionality setting of (8,8,8) to compare with GANITE and TEDVAE. As shown in Fig. 4, we can see that as ζ increases (i.e., the selection bias increases), VGANITE outperforms the other methods.

Finally, we investigate the contribution of the disentangled latent instrumental, confounding, and adjustment factors for ITE estimation respectively. We compare the performance when setting one latent dimension parameter of m_t,m_c,m_y to zero with the performance when setting any dimension

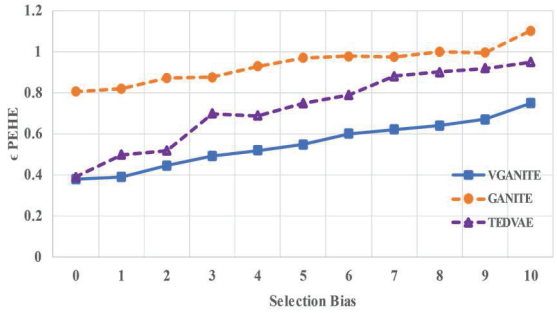


Fig. 4. As the selection bias increases, VGANITE performs better than other models.

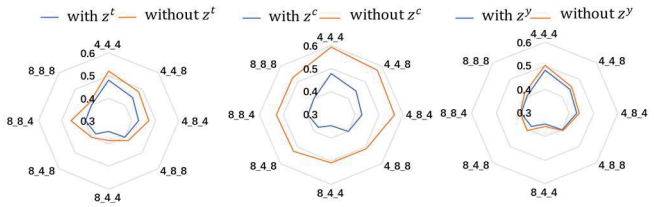


Fig. 5. In the radar charts, each vertex on the polygon is identified using a sequence of latent factor dimensions from the synthetic datasets, with each polygon representing the model's PEHE metric (the small polygon is better).

parameters to non-zero (e.g., setting $m_t = 0$ means without $\boldsymbol{z}^t$ to force VGANITE to ignore the disentangle of instrumental factors). If VGANITE performs better when considering disentangle latent factors than ignoring one of the latent factors, we can know the contribution of disentangling the latent factors for ITE estimation. As shown in Fig. 5. VGANITE can perform better when considering learn disentangled latent instrumental, confounding, and adjustment factors, especially disentangling latent confounding factors.

V. Conclusion

In this paper, we focus on the ITE estimation from observational data. We argue that most previous methods ignore the importance of latent confounding and non-confounding factors identification. Hence, we propose the VGANITE method combining the frameworks of VAE and GAN to learn disentangled latent instrumental, confounding, and adjustment factors and generate the counterfactual outcomes for ITE estimation. Experiments have shown that learning disentangled latent factors is effective and our approach is more robust compared with baseline methods.

However, the disentanglement of VGANITE relies on the correlations between three variables and treatment or outcome, which may not be pure enough for a three-part decomposition. And VGANITE is limited to dealing with binary discrete treatment variables. For future work, we will consider better disentangling latent factors and extending the approach to continuous treatment variables.

References

[1] X. Zhang, P. Cui, R. Xu, L, Zhou, Y. He, and Z. Shen, "Deep Stable Learning for Out-Of-Distribution Generalization," *in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5372-5382, 2021.

[2] Y. Niu, K. Tang, H. Zhang, Z. Lu, X. S. Hua, and J. R. Wen, "Counterfactual VQA: A cause-effect look at language bias," *in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12700-12710, 2021.

[3] R. Guo, L. Cheng, J. Li, P. R. Hahn, and H. Liu, "A survey of learning causality with data: Problems and methods," *ACM Computing Surveys*, vol. 53, no. 4, pp. 1-37, 2020.

[4] L.Yao, Z. Chu, S. Li, Y. Li, J. Gao, and A. Zhang, "A survey on causal inference," *arXiv preprint arXiv*:2002.02770, 2020.

[5] S. Athey, and G. Imbens, "Recursive partitioning for heterogeneous causal effects," *National Academy of Sciences,* vol. 113, no. 27, pp. 7353-7360, 2016.

[6] R. H. Dehejia, and S. Wahba. "Propensity score-matching methods for nonexperimental causal studies," *Review of Economics and Statistics* vol. 84, no. 1, pp. 151-161,2002.

[7] J. K. Lunceford, and M. Davidian, "Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study," *Statistics in medicine*, vol. 23, no. 19, pp. 2937-2960, 2004.

[8] F. Johansson, U. Shalit, and D. Sontag, "Learning representations for counterfactual inference," *in Proceedings of the International conference on machine learning*, pp. 3020-3029, 2016.

[9] A. M. Alaa, M. Weisz, and M. Van Der Schaar, "Deep counterfactual networks with propensity-dropout," *arXiv preprint arXiv*:1706.05966, 2017.

[10] J. Yoon, J. Jordon, and M. Van Der Schaar, "GANITE: Estimation of individualized treatment effects using generative adversarial nets, " *in Proceedings of the International Conference on Learning Representations*, 2018.

[11] A. Wu, K. Kuang, J. Yuan, B. Li, P. Zhou, J. Tao, Q. Zhu, Y. Zhuang, F. Wu, "Learning decomposed representation for counterfactual inference, " *arXiv preprint arXiv*:2006.07040, 2020.

[12] A. Abadie, and G. W. Imbens, "Large sample properties of matching estimators for average treatment effects, " *Econometrica*, vol. 74, no. 1, pp. 235-267, 2006.

[13] J. Häggström, "Data-driven confounder selection via Markov and Bayesian networks,". *Biometrics*, vol.74, no. 2, pp. 389-398, 2018.

[14] P. R. Rosenbaum, and D. B. Rubin, "The central role of the propensity score in observational studies for causal effects, " *Biometrika*, vol. 70, no. 1, pp. 41-55, 1983.

[15] D. A. Freedman, and R. A. Berk, "Weighting regressions by propensity scores," *Evaluation Review*, vol. 32, no. 4, pp. 392-409, 2008.

[16] U. Shalit, F. D. Johansson, and D. Sontag, "Estimating individual treatment effect: generalization bounds and algorithms," *in Proceedings of the International Conference on Machine Learning*, pp. 3076-3085, 2017.

[17] N. Hassanpour, and R. Greiner, "CounterFactual Regression with Importance Sampling Weights," *in Proceeding of the International Joint Conference on Artificial Intelligence,* pp. 5880-5887, 2019.

[18] J. R. Zubizarreta, "Stable weights that balance covariates for estimation with incomplete outcome data," *Journal of the American Statistical Association*, vol. 110, no. 511, pp. 910-922, 2015.

[19] C. Cortes, Y. Mansour, and M. Mohri, "Learning Bounds for Importance Weighting," *in Proceeding of the Neural Information Processing Systems*, pp. 442-450, 2010.

[20] K. Kuang, P. Cui, B. Li, M. Jiang, S. Yang, et al. "Treatment effect estimation with data-driven variable decomposition," *in Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 31, no. 1, 2017.

[21] N. Hassanpour, and R. Greiner," Learning disentangled representations for counterfactual regression," *in Proceedings of the International Conference on Learning Representations*, 2019.

[22] W. Zhang, L. Liu, and J. Li, "Treatment Effect Estimation with Disentangled Latent Factors, " *in Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 12, pp. 10923-10930, 2020.

[23] C. Louizos, U. Shalit, J. M. Mooij, D. Sontag, R. S. Zemel, and M. Welling, "Causal Effect Inference with Deep Latent-Variable Models," *in Proceeding of the Neural Information Processing Systems*, pp. 6449-6459, 2017.

[24] D. P. Kingma, and M. Welling, "Auto-encoding variational Bayes, " *in Proceedings of the International Conference on Learning Representations*, 2014.

[25] M. Mirza, and S. Osindero, " Conditional generative adversarial nets," *arXiv preprint arXiv*:1411.1784, 2014.

[26] R. K. Crump, V. J. Hotz, G. W. Imbens, and O. A. Mitnik, "Nonparametric tests for treatment effect heterogeneity," *The Review of Economics and Statistics,* vol. 90, no. 3, pp. 389-405, 2008.

[27] H. A. Chipman, E. I. George, and R. E. McCulloch, "BART: Bayesian additive regression trees," *The Annals of Applied Statistics*, vol. 4, no. 1, pp. 266-298, 2010.

[28] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5-32, 2001.

[29] S. Wager, and S. Athey, "Estimation and inference of heterogeneous treatment effects using random forests," *Journal of the American Statistical Association,* vol. 113, no. 523, pp. 1228-1242, 2018.