DR-VIDAL - Doubly Robust Variational Information-theoretic Disentangled Adversarial Learning for Estimation of Counterfactuals and Treatment Effects

ABSTRACT

Randomized controlled trials (RCTs) have been the gold-standard, in biomedical and social sciences among others, to assess causal effects of interventions on outcomes, e.g. medical treatments or lifestyle changes, and to reduce the risk of biased estimations. However, conducting RCTs is not always feasible due to operational or ethical constraints. Alternatively, causal effects can be drawn from observational, real-world data, but the data generation and collection process might not be randomized and contain underlying bias. Several techniques for addressing bias in treatment assignments, and for predicting individualized treatment effects (ITEs) –including counterfactuals, i.e., outcomes for alternative treatment scenarios– have been proposed, from propensity score matching, to ensemble tree-based learning, to recent breakthrough in deep learning, e.g., the Causal Effect Variational Autoencoder (CEVAE) or the Generative Adversarial Nets for inference of Individualized Treatment Effects (GANITE). In this work, we propose a novel deep learning approach, the Doubly Robust Variational Information-theoretic Disentangled Adversarial Learning (DR-VIDAL) that incorporates the following key characteristics: (1) disentangled variational autoencoder with latent variables under a causal structure substantially different from the CEVAE; (2) feature attention on high-dimensional datasets and noise-reduction; (3) information-theoretic optimization for instance generation and prediction of counterfactuals; and (4) doubly-robust ITE estimation. Tests performed on real-world datasets showed that the DR-VIDAL outperforms several other state-of-the-art techniques. The utility of DR-VIDAL is not only with respect to prediction, for which the doubly robustness is assured, but also for more general inference tasks. The code is available under the MIT license on Github at: https://bitbucket.org/goingdeep2406/dr-vidal/src/master/

CCS CONCEPTS

• **Deep Learning** → **Representation Learning**; *Variational Autoencoder*; Generative Adversarial Network; • **Causal Inferences** → Doubly Robust Estimation .

KEYWORDS

causal AI, biomedical informatics, generative adversarial networks, variational inference, information theory, doubly robust

ACM Reference Format:
. 2018. DR-VIDAL - Doubly Robust Variational Information-theoretic Disentangled Adversarial Learning for Estimation of Counterfactuals and Treatment Effects. In *Woodstock '18: ACM Symposium on Neural Gaze Detection, June 03–05, 2018, Woodstock, NY*. ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

Evaluating causal effects of interventions on outcomes is key to knowledge and progress in many fields, e.g., medicine, psychology, public health, and policy making. A typical scenario in medicine is to determine whether a treatment (e.g., lipid-lowering medication) is effective to reduce risk or cure an illness (e.g., cardiovascular disease). Randomized controlled trials (RCTs) are often considered to be the best practice for evaluating causal effects [33]. In an RCT, a treatment is assigned randomly to individuals, with some getting it and some being assigned a placebo, making it independent of an individual's characteristics, thus avoiding selection bias and influence on the outcome from other sources rather then the treatment (albeit other possible bias, e.g., people understanding if they have been given a placebo). However, RCTs are not always feasible due to ethical and legal constraints. For instance, if one wanted to evaluate whether college education is the cause of good salary, it would not be ethical to randomly pick teenagers and randomize their attendance to college. So, in many cases, observational data, i.e., real-world data collected retrospectively and not randomized, are the only usable source. Unfortunately, observational data are often plagued with various biases –because the data generation processes are largely unknown– such as confounding (i.e., spurious causal effects on outcomes by features that are correlated with a true unmeasured cause) and colliders (i.e., mistakenly including effects of an outcome as predictors), making it difficult to infer causal claims [16]. In other words, without randomization, it is quite difficult to distinguish conditional association from causation. Another problem with causal effect estimation is that, in both RCTs and observational datasets, only factual outcomes are available, since clearly an individual cannot be treated and not-treated at the same time. Counterfactuals are alternative predictions that respond to the question "what outcome would have been observed if a person had been given the placebo instead of the treatment?" If models are biased, counterfactual predictions can be wrong, and interventions can be ineffective or harmful [28].

Traditional statistical approach for estimating treatment effects, taking into account possible bias from pre-treatment characteristics, are propensity score matching (PSM) and inverse probability weighting (IPW) [5]. The propensity score is a scalar estimate representing the conditional probability of receiving a certain treatment, given a set of measured pre-treatment covariates. By matching (or

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

Woodstock '18, June 03–05, 2018, Woodstock, NY
© 2018 Association for Computing Machinery.
ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/10.1145/1122445.1122456

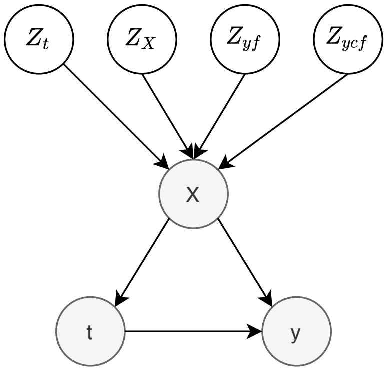

Figure 1: Directed acyclic graph modelling the causal relationships among a treatment t, outcome y and pre-treatment covariates X, under a latent space Z.

weighting) treated and control subjects according to their propensity score, a balance in pre-treatment covariates is induced, mimicking a randomization of the treatment group assignment. However, traditional PSM approach accounts only for measured covariates, and latent bias may remain after matching [14]. PSM has been implemented historically with logistic (linear) regression, coupled with different feature selection methods in presence of high-dimensional datasets [34]. A problem with PSM is that it often decreases the sample size due to matching, while IPW can be affected by skewed, heavy-tailed weight distributions. Machine learning approaches have been introduced more recently, e.g., Bayesian additive regression trees [17] and counterfactual random forests [35]. Big data also led to the flourishing of deep learning tailored to causal inference [19]. Notable examples include Treatment-Agnostic Representation Network (TARNet) [31], Dragonnet [32], Deep Counterfactual Network with Propensity-Dropout (DCN-PD) [2], Generative Adversarial Nets for inference of Individualized Treatment Effects (GANITE) [36], and Causal Effect Variational Autoencoder (CEVAE) [23].

1.1 Contribution

This work introduces a novel deep learning approach for treatment effect estimation and counterfactual prediction, named the *Doubly Robust Variational Information-theoretic Disentangled Adversarial Learning* (DR-VIDAL). The main features of DR-VIDAL are:

- Incorporation of an underlying causal structure where the observed pre-treatment covariate set X is decomposed into four independent latent variables $Z_t, Z_X, Z_{yf}, Z_{ycf}$, inducing confounding on both the treatment and the outcome (Figure 1).
- Latent variables are inferred using a variational autoencoder (VAE) [21] with disentanglement.
- A generative adversarial network (GAN) [15] with variational information maximization [8] generates (synthetic) complete tuples of covariates, treatment, factual and counterfactual outcomes.
- Individual treatment effects (ITE) are estimated on complete datasets with a downstream deep learning module which is doubly robust [12, 13].

To our knowledge, this is the first time in which VAE, GAN, information theory and doubly robustness are amalgamated together into a counterfactual prediction method. By performing test runs on three real world, popular benchmark datasets, we show that DR-VIDAL can outperform a number of state-of-art tools for estimating ITE.

2 PROBLEM FORMULATION

We utilize the *potential outcomes* framework [29, 30]. Let us consider a treatment T (binary for ease of reading, but the theory can be extended to multiple treatments) which can be prescribed to a population sample of size N. The individuals are characterized by a set of pre-treatment background covariates $\boldsymbol{X}$, and a health outcome Y is measured after treatment. We define each subject i with the tuple $\{\boldsymbol{X}, T, Y\}_{i=1}^{N}$, where Y_i^0 and Y_i^1 are the potential outcomes when applying treatments $T_i = 0$ and $T_i = 1$, respectively. The ITE $\tau(\boldsymbol{x})$ for subject i with pre-treatment covariates $\boldsymbol{X}_i = \boldsymbol{x}$, is defined as the difference in the average potential outcomes under both treatment interventions (i.e., treated vs. not treated), conditional on $\boldsymbol{x}$, i.e.,

$$\tau(\boldsymbol{x}) = \mathbb{E}[Y_i^1 - Y_i^0 \mid \boldsymbol{X}_i = \boldsymbol{x}] \tag{1}$$

The ITE cannot be calculated directly give the inaccessibility to both potential outcomes, as only factual outcomes can be observed, while the others (counterfactual) can be considered as missing values. However, when the potential outcomes are made independent of the treatment assignment, conditionally on the pre-treatment covariates, i.e. $\{Y^1, Y^0\} \perp T \mid \boldsymbol{X}$, the ITE can then be estimated as $\tau(\boldsymbol{x}) = \mathbb{E}[Y^1 \mid T = 1, \boldsymbol{X} = \boldsymbol{x}] - \mathbb{E}[Y^0 \mid T = 0, \boldsymbol{X} = \boldsymbol{x}] = \mathbb{E}[Y \mid T = 1, \boldsymbol{X} = \boldsymbol{x}] - \mathbb{E}[Y \mid T = 0, \boldsymbol{X} = \boldsymbol{x}]$. Such assumption is called strongly ignorable treatment assignment (SITA) [18, 26]. By further averaging over the distribution of $\boldsymbol{X}$, the average treatment effect (ATE) τ_{01} can be calculated as

$$\tau_{01} = \mathbb{E}[\tau(\boldsymbol{X})] = \mathbb{E}[Y \mid T = 1] - \mathbb{E}[Y \mid T = 0] \tag{2}$$

ITE and ATE can be calculated straightforward with stratification matching of $\boldsymbol{x}$ in treatment and control groups, but the calculation becomes unfeasible as the covariate space increases in dimensions.

The propensity score $\pi(x)$ represents the probability of receiving the treatment $T = 1$ conditioned on the pre-treatment covariates $X = x$ [29], denoted as

$$\pi(\boldsymbol{x}) = P(T = 1|\boldsymbol{X} = \boldsymbol{x}). \tag{3}$$

The propensity score can be identified using a regression function such as logistic regression, and then ITE/ATE can be calculated by matching (PSM) or weighting (IPW) instances through $\pi(\boldsymbol{x})$, in a doubly robust way [27], or with a glut of other approaches [4, 9–11, 24, 25, 27, 35]. In the next section, we describe those based on deep learning.

3 RELATED WORK

A comprehensive work on characterizing the conditions and the limits of heterogeneous treatment effect estimation using deep learning has been provided [1]. The sample size plays an important role, e.g., estimations on small sample sizes are affected by selection bias, whilst on large sample sizes by algorithmic design. Our work is motivated mostly from the advancement in ITE estimation

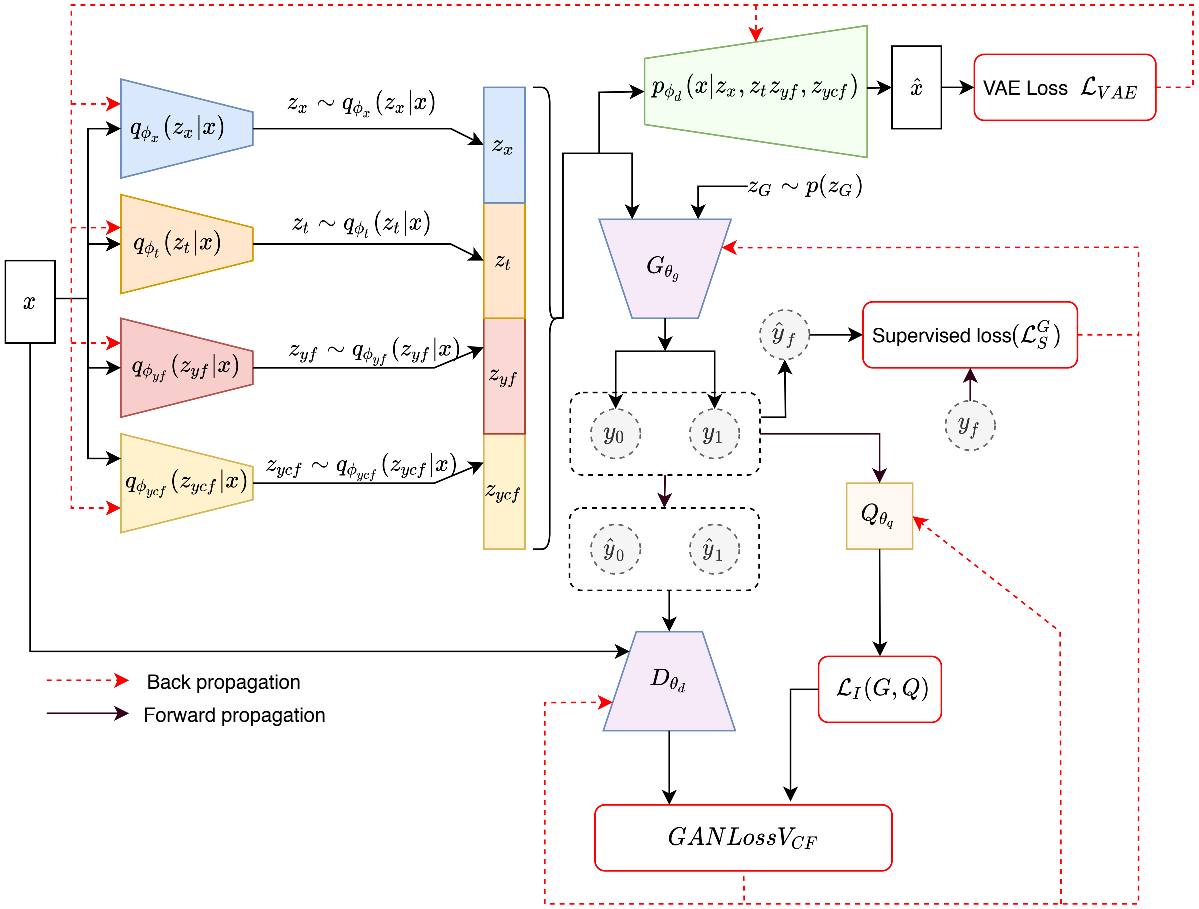
x
$q_{\phi_x}(z_x|x)$
$q_{\phi_t}(z_t|x)$
$q_{\phi_{yf}}(z_{yf}|x)$
$q_{\phi_{ycf}}(z_{ycf}|x)$
$z_x \sim q_{\phi_x}(z_x|x)$
$z_t \sim q_{\phi_t}(z_t|x)$
$z_{yf} \sim q_{\phi_{yf}}(z_{yf}|x)$
$z_{ycf} \sim q_{\phi_{ycf}}(z_{ycf}|x)$
z_x
z_t
z_{yf}
z_{ycf}
$p_{\phi_d}(x|z_x, z_t z_{yf}, z_{ycf})$
$\hat{x}$
VAE Loss $\mathcal{L}_{VAE}$
$z_G \sim p(z_G)$
G_{θ_g}
$\hat{y}_f$
Supervised loss($\mathcal{L}_S^G$)
y_f
y_0
y_1
$\hat{y}_0$
$\hat{y}_1$
Q_{θ_q}
D_{θ_d}
$\mathcal{L}_I(G,Q)$
$GANLossV_{CF}$
Back propagation
Forward propagation

Figure 2: Architecture of the counterfactual network to estimate the counterfactual outcome.

brought by DCN-PD [2], CEVAE [23], GANITE [36], TARNet [31], and Dragonnet [32]. DCN-PD is a doubly robust, multitask network for counterfactual prediction, where propensity scores are used to determine a dropout probability of samples to regularize training, carried out in alternating phases, using treated and control batches. CEVAE uses VAE and GAN to identify latent variables from an observed pre-treatment vector and to generate counterfactuals. GANITE generates proxies of counterfactual outcomes using a GAN, and feeding these proxies to an ITE generator. In TARNet, each sample from the treated and control group is associated with a weight indemnifying group imbalance. Dragonnet is a modified architecture of TARNet that introduces targeted regularization based on propensity scores.

In the following sections, we discuss in detail the novelty and the differences in architectural design and training mechanisms of DR-VIDAL with respect to the aforementioned approaches.

4 PROPOSED METHODOLOGY

DR-VIDAL architecture can be decomposed into three main parts: (1) a VAE to disentanglement of the latent variables, (2) a GAN to generate the counterfactual outcomes (3) a doubly robust module to estimate ITE. The architectural layout and the training algorithm of the first two components (VAE and GAN) are illustrated in Figure 2 and Algorithm 1, respectively. For the third component (doubly robust estimator), architecture and training algorithm are shown in Figure 3 and Algorithm 2, respectively.

4.1 Latent variable disentanglement with VAE

We assume that the observed covariates x are generated from an independent latent space z, composed by $z_x \sim p(z_x)$, $z_t \sim p(z_t)$, $z_{yf} \sim p(z_{yf}), and z_{ycf} \sim p(z_{ycf})$, which denote the latent variables for the covariates x, treatment indicator t, and factual outcomes y_f and y_{cf}, respectively. This decomposition follows the causal structure shown in Figure 1. The goal is to infer the posterior distribution $p(z_x, z_t, z_{yf}, z_{ycf}|x)$, which is however harder to optimize. So, we use the theory of variational inference [6] to learn the variational posteriors $q_{\phi_x}(z_x|x)$, $q_{\phi_t}(z_t|x)$, $q_{\phi_{yf}}(z_{yf}|x), q_{\phi_{ycf}}(z_{ycf}|x)$, using 4 different neural network encoders with parameters $\phi_x, \phi_t, \phi_{yf}, and \phi_{ycf}$, respectively. Using the latent factors sampled from the learned variational posteriors, we reconstruct x by estimating the likelihood $p_{\phi_d}(x|z_x, z_t, z_{yf}, z_{ycf})$

via a single decoder parameterized by ϕ_d. The latent factors, assumed Gaussian, are defined as follows:

$$p(z_x) = \prod_{i=1}^{D_{z_x}} \mathcal{N}(z_{x_i}|0,1); \qquad p(z_t) = \prod_{i=1}^{D_{z_t}} \mathcal{N}(z_{t_i}|0,1) \tag{4}$$

$$p(z_{yf}) = \prod_{i=1}^{D_{z_{yf}}} \mathcal{N}(z_{yf_i}|0,1); \quad p(z_{ycf}) = \prod_{i=1}^{D_{z_{ycf}}} \mathcal{N}(z_{ycf_i}|0,1) \tag{5}$$

where $D_{z_x}, D_{z_t}, D_{z_{yf}}, D_{z_{ycf}}$ are the dimensions of the latent factors $z_x, z_t, z_{yf}, z_{ycf}$, respectively. The variational posteriors of the inference of models are defined as:

$$q_{\phi_x}(z_x|x) = \prod_{i=1}^{D_{z_x}} \mathcal{N}(\mu = \hat{\mu}_x, \sigma^2 = \hat{\sigma}_x^2) \tag{6}$$

$$q_{\phi_t}(z_t|x) = \prod_{i=1}^{D_{z_t}} \mathcal{N}(\mu = \hat{\mu}_t, \sigma^2 = \hat{\sigma}_t^2) \tag{7}$$

$$q_{\phi_{yf}}(z_{yf}|x) = \prod_{i=1}^{D_{z_{yf}}} \mathcal{N}(\mu = \hat{\mu}_{yf}, \sigma^2 = \hat{\sigma}_{yf}^2) \tag{8}$$

$$q_{\phi_{ycf}}(z_{ycf}|x) = \prod_{i=1}^{D_{z_{ycf}}} \mathcal{N}(\mu = \hat{\mu}_{ycf}, \sigma^2 = \hat{\sigma}_{ycf}^2) \tag{9}$$

where $\hat{\mu}_x, \hat{\mu}_t, \hat{\mu}_{yf}, \hat{\mu}_{ycf}$ and $\hat{\sigma}_x^2, \hat{\sigma}_t^2, \hat{\sigma}_{yf}^2, \hat{\sigma}_{ycf}^2$ are the means and variances of the Gaussian distributions parameterized by encoders $E_{\phi_x}, E_{\phi_t}, E_{\phi_{yf}}, E_{\phi_{ycf}}$ with parameters $\phi_x, \phi_t, \phi_{yf}, \phi_{ycf}$ respectively.

The overall evidence lower bound (ELBO) loss of the VAE is expressed as $\mathcal{L}_{ELBO}$ in the following equation,

$$\begin{aligned}
&\mathcal{L}_{ELBO}(\phi_x, \phi_t, \phi_{yf}, \phi_{ycf}; x, z_x, z_t, z_{yf}, z_{ycf}) \\
&= \mathbb{E}_{q_{\phi_x}, q_{\phi_t}, q_{\phi_{yf}}, q_{\phi_{ycf}}}[\log p_{\phi_d}(x|z_x, z_t, z_{yf}, z_{ycf})] \\
&\quad - KL\big(q_{\phi_x}(z_x|x)||p_{\phi_d}(z_x))\big) \\
&\quad - KL\big(q_{\phi_t}(z_t|x)||p_{\phi_d}(z_t))\big) \\
&\quad - KL\big(q_{\phi_{yf}}(z_{yf}|x)||p_{\phi_d}(z_{yf}))\big) \\
&\quad - KL\big(q_{\phi_{ycf}}(z_{ycf}|x)||p_{\phi_d}(z_{ycf}))\big)
\end{aligned}$$

We minimize the optimization function of the VAE as $\mathcal{L}_{VAE}$ to obtain the optimal parameter of the encoders $\phi_x, \phi_t, \phi_{yf}, \phi_{ycf}$, and of the decoder ϕ_d.

$$\begin{aligned}
&\mathcal{L}_{VAE}(\phi_x, \phi_t, \phi_{yf}, \phi_{ycf}; x, z_x, z_t, z_{yf}, z_{ycf}) = \\
&- \mathcal{L}_{ELBO}(\phi_x, \phi_t, \phi_{yf}, \phi_{ycf}; x, z_x, z_t, z_{yf}, z_{ycf})
\end{aligned} \tag{10}$$

A detailed derivation of the ELBO loss for the VAE is given in the Appendix A.1.

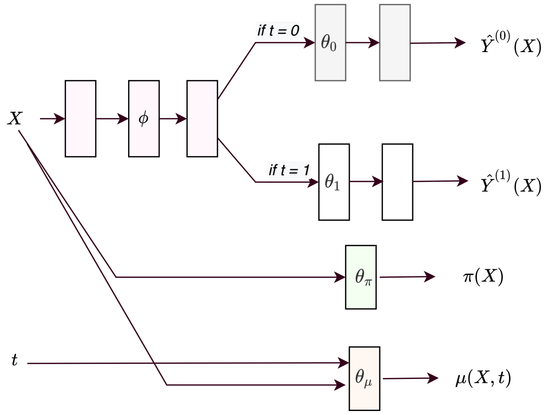

Figure 3: Architecture of the doubly robust network to calculate the average treatment effect.

4.2 Generation of counterfactuals via GAN

After learning the hidden latent codes $z_x, z_t, z_{yf}, z_{ycf}$ from the VAE, we concatenate the latent codes to form z_c, passed to the generator of the GAN block G_{θ_g}, along with a random noise $z_G \sim \mathcal{N}(0, Id)$. G_{θ_g} is parameterized by θ_g, and it outputs the vector $\overline{y}$ of the potential (both factual and counterfactual) outcomes. We then replace the true factual outcome y_f in the generated outcome vector $\overline{y}$ to form $\hat{y}_0$ and $\hat{y}_1$, which are passed to the counterfactual discriminator D_{θ_d}, along with the true covariate vector x. D_{θ_d} is parameterized by θ_d, and is responsible to predict the treatment variable, similarly to GANITE. The loss of the GAN block is defined as:

$$\begin{aligned}
V_{GAN}(G, D) = \mathbb{E}_{x, z_G, z_c}\big[&t^T \log D(x, G(z_G, z_c)) \\
&+ (1-t)^T \log(1 - D(x, G(z_G, z_c))\big]
\end{aligned}$$

where $x \sim p(x), z_G \sim p(z_G)$ and z_c denote the concatenated latent codes $z_x \sim q_{\phi_x}(z_x|x)$, $z_t \sim q_{\phi_t}(z_t|x)$, $z_{yf} \sim q_{\phi_{yf}}(z_{yf}|x)$ and $z_{ycf} \sim q_{\phi_{ycf}}(z_{ycf}|x)$.

From $\overline{y}$, we also calculate the predicted factual outcome $\hat{y}_f$. As also done in GANITE, we make sure to include the supervised loss $\mathcal{L}_S^G(y_f, \hat{y}_f)$, which enforces the predicted factual outcome $\hat{y}_f$ to be as close as to the true factual outcome y_f.

$$\mathcal{L}_S^G(y_f, \hat{y}_f) = \frac{1}{n}\sum_{i=1}^{n}\big(y_f(i) - \hat{y}_f(i)\big)^2 \tag{11}$$

The complete loss function of counterfactual GAN is given by

$$V_{CF}(G, D) = V_{GAN}(G, D) + \gamma \mathcal{L}_S^G(y_f, \hat{y}_f)$$

We also employ an additional regularization $\lambda I(z_c; G(z_G, z_c))$ to maximize the mutual information between the learnt concatenated latent code z_c and the generated output by the generator $G(z_G, z_c)$, as mentioned in [8]. So, we propose to solve the following minimax game:

$$\min_G \max_D V_{CF\_I}(G, D) = V_{CF}(G, D) + \lambda I(z_c; G(z_G, z_c)) \tag{12}$$

$I(z_c; G(z_G, z_c))$ is harder to solve because of the presence of the posterior $p(z_c|x)$ [8], so we obtain the lower bound of it using an auxiliary distribution $Q(z_c|x)$ to approximate $p(z_c|x)$.

Finally, the optimization function of the counterfactual information-theoretic GAN –*InfoGAN*– incorporating the variational regularization of mutual information and a hyperparameter λ is given by:

$$\min_{G,Q} \max_D V_{CF\_infoGAN}(G, D, Q) = V_{CF}(G, D) - \lambda \mathcal{L}_I(G, Q) \tag{13}$$

The counterfactual InfoGAN is used to generate the missing counterfactual outcome y_{cf} to form the quadruple $\{X, t, y_f, y_{cf}\}_{i=1}^{N}$ and sent to the doubly robust block to estimate the ITE.

Algorithm 1 Training of the generative adversarial network for counterfactual outcome calculation

```
Input: Training set $X = \{(x^{(1)}, t^{(1)}, y_f^{(1)}),..., (x^{(n)}, t^{(n)}, y_f^{(n)})\}$; hyper-parameters $\gamma > 0$; $\lambda > 0$; Encoders: $E_{\phi_x}, E_{\phi_t}, E_{\phi_{yf}}, E_{\phi_{ycf}}$ with parameters $\phi_x, \phi_t, \phi_{yf}, \phi_{ycf}$ respectively; Decoder $D_{\phi_d}$ with parameter $D_{\phi_d}$; Generator $G_{\theta_g}$, Discriminator $D_{\theta_d}$, Q network $D_{\theta_q}$ with parameters $\theta_g, \theta_d, \theta_q$ respectively
Initialize parameters: $\phi_x, \phi_t, \phi_{yf}, \phi_{ycf}, \phi_d, \theta_g, \theta_d, \theta_q$
while training do
    $x \leftarrow$ batch of samples from the dataset
    $z_{\mu_x}, z_{\sigma_x} \leftarrow E_{\phi_x}(x)$
    $z_{\mu_t}, z_{\sigma_t} \leftarrow E_{\phi_t}(x)$
    $z_{\mu_{yf}}, z_{\sigma_{yf}} \leftarrow E_{\phi_{yf}}(x)$
    $z_{\mu_{ycf}}, z_{\sigma_{ycf}} \leftarrow E_{\phi_{ycf}}(x)$
    $z_x \leftarrow z_{\mu_x} + \epsilon z_{\sigma_x}$ , where $\epsilon \sim \mathcal{N}(0, Id)$
    $z_t \leftarrow z_{\mu_t} + \epsilon z_{\sigma_t}$ , where $\epsilon \sim \mathcal{N}(0, Id)$
    $z_{yf} \leftarrow z_{\mu_{yf}} + \epsilon z_{\sigma_{yf}}$ , where $\epsilon \sim \mathcal{N}(0, Id)$
    $z_{ycf} \leftarrow z_{\mu_{ycf}} + \epsilon z_{\sigma_{ycf}}$ , where $\epsilon \sim \mathcal{N}(0, Id)$
    Concatenate $z_x, z_t, z_{yf}, z_{ycf}$ to form $z_c$
    $\hat{x} \leftarrow D_{\phi_d}(z_c)$
    Calculate $\mathcal{L}_{VAE}(\phi_x, \phi_t, \phi_{yf}, \phi_{ycf}; x, z_x, z_t, z_{yf}, z_{ycf})$
    $\phi_x \overset{-}{\longleftarrow} \nabla_{\phi_x}\mathcal{L}_{VAE}$; $\phi_t \overset{-}{\longleftarrow} \nabla_{\phi_t}\mathcal{L}_{VAE}$; $\phi_{yf} \overset{-}{\longleftarrow} \nabla_{\phi_{yf}}\mathcal{L}_{VAE}$; $\phi_{ycf} \overset{-}{\longleftarrow} \nabla_{\phi_{ycf}}\mathcal{L}_{VAE}$; $\phi_d \overset{-}{\longleftarrow} \nabla_{\phi_d}\mathcal{L}_{VAE}$
    $z_G \sim \mathcal{N}(0, Id)$
    $y_0, y_1 \leftarrow G_{\theta_g}(z_G, z_c)$
    $\hat{y}_0 = ((1 - T) * y_f + T * y_0)$
    $\hat{y}_1 = (T * y_f + (1 - T) * y_1)$
    $d_{logit} \leftarrow D_{\theta_d}(x, \hat{y}_0, \hat{y}_1)$
    Calculate $\mathcal{L}^D(\theta_d)$
    $\theta_d \overset{-}{\longleftarrow} \nabla_{\theta_d}\mathcal{L}^D(\theta_d)$
    $\hat{y}_f \leftarrow T * y_1 + (1 - T) * y_0$
    Compute$\mathcal{L}_S^G(y_f, \hat{y}_f)$
    Concatenate $y_0, y_1$ to form $q_{input}$
    $q_\mu, q_\sigma \leftarrow Q_{\theta_q}(q_{input})$
    Compute $\mathcal{L}_I(G, Q)$ by treating $Q(c|x)$ as factored Gaussian using $q_\mu, q_\sigma$ and $z_c$
    Compute $\mathcal{L}^G(\theta_g)$
    $\theta_g \overset{-}{\longleftarrow} \nabla_{\theta_g}\mathcal{L}^G(\theta_g)$
end while
```

Information-theoretic GAN optimization

The GAN generator G_{θ_g} works to fool the discriminator D_{θ_d}. To get the optimal Discriminator $D^*_{\theta_d}$, we maximize $V_{CF\_infoGAN}$ as

$$\max_{D} \mathcal{L}^D(\theta_d) = V_{CF\_infoGAN}(G, D, Q) \tag{14}$$

To get the optimal Generator $G^*_{\theta_g}$, we maximize $V_{CF\_infoGAN}$ as

$$\min_{G,Q} \mathcal{L}^G(\theta_g) = V_{CF\_infoGAN}(G, D, Q) \tag{15}$$

A detailed derivation the information-theoretic loss of the GAN is given in the Appendix A.2.

4.3 Doubly robust ITE estimation

Algorithm 2 Training of the doubly robust multitask network for ITE estimation

```
Input: Complete dataset $\tilde{X} = \{(x^{(1)}, t^{(1)}, y_f^{(1)}, y_{cf}^{(1)}),..., (x^{(n)}, t^{(n)}, y_f^{(n)}, y_{cf}^{(n)})\}$ after training the GAN module for counterfactual prediction; hyper-parameters $\alpha > 0$; $\beta > 0$; outcome heads with shared parameters $\phi$ and outcome specific parameters $\theta_0, \theta_1$; propensity head with parameters $\theta_\pi$; regressor head with parameters $\theta_\mu$
Initialize parameters: $\theta_0, \theta_1, \theta_\pi, \theta_\mu$
while training do
    $x \leftarrow$ batch of samples from the dataset
    Calculate $\hat{y}_i^{(0)}, \hat{y}_i^{(1)}, \hat{y}_f^{(i)}, \hat{y}_{cf}^{(i)}$
    Calculate the predicted loss $\mathcal{L}_i^p(\theta_1, \theta_0, \phi)$
    Calculate $\hat{y}_{fDR}^{(i)}, \hat{y}_{cfDR}^{(i)}$
    Calculate the Doubly Robust loss $\mathcal{L}_i^{DR}(\theta_1, \theta_0, \theta_\pi, \theta_\mu, \phi)$
    Calculate the final loss $\mathcal{L}_{ITE}(\theta_1, \theta_0, \theta_\pi, \theta_\mu, \phi)$
    Calculate gradients of the loss $\mathcal{L}_{ITE}(\theta_1, \theta_0, \theta_\pi, \theta_\mu, \phi)$
    Update the parameters $\theta_1, \theta_0, \theta_\pi, \theta_\mu, \phi$
end while
```

As above introduced, the propensity score $\pi(x)$ represents the probability of receiving a treatment $T = 1$ (over the alternative $T = 0$) conditioned on the pre-treatment covariates X. By combining IPW through $\pi(x)$ with outcome regression by both treatment variable and the covariates, Jonsson defined the doubly robust estimation of causal effect [13] as follows:

$$\hat{\delta}_{DR} = \frac{1}{n}\sum_{i=1}^{n}\Bigg[\frac{y_i t_i - (t_i - \pi(x_i))\mu(x_i, t_i)}{\pi(x_i)} - \frac{y_i(1 - t_i) - (t_i - \pi(x_i))\mu(x_i, t_i)}{1 - \pi(x_i)}\Bigg] \tag{16}$$

where $\mu(x, t) = \hat{\alpha_0} + \hat{\alpha_1}x_1 + \hat{\alpha_2}x_2 + \cdots + \hat{\alpha_n}x_n + \hat{\delta}t$, and $(t_i - \pi(x_i))\mu(x_i, t_i)$ is used for the IPW estimator.

After getting the counterfactual outcome y_{cf} from the counterfactual GAN to form the quadruple $\{X, t, y_f, y_{cf}\}_{i=1}^{N}$, we pass this as the input to the doubly robust multitask network to estimate the ITE, using the architecture shown in Figure 3. To predict the outcomes $y^{(0)}$ and $y^{(1)}$, we use a configuration similar to TARNet, which contains a number of shared layers, denoted by f_ϕ, parameterized by ϕ, and two outcome specific heads f_{θ_0} and f_{θ_1}, parameterized by θ_0 and θ_1.

To ensure doubly robustness, we introduce two more heads that predict the propensity score $\pi(X) = \mathbb{P}(t = 1|X)$ and the regressor $\mu(X, t)$. These two are calculated using two neural networks, parameterized by θ_π and θ_μ respectively. The factual and counterfactual outcome $y_i^{(0)}$ and $y_i^{(1)}$ of the i^{th} sample are then calculated as:

$$\hat{y}_i^{(0)} = f_{\theta_0}(f_\phi(x_i)) \qquad \text{if } t_i = 0 \tag{17}$$

$$\hat{y}_i^{(1)} = f_{\theta_1}(f_\phi(x_i)) \qquad \text{if } t_i = 1 \tag{18}$$

$$\hat{y}_f^{(i)} = t_i\hat{y}_i^{(1)} + (1 - t_i)\hat{y}_i^{(0)} \tag{19}$$

$$\hat{y}_{cf}^{(i)} = (1 - t_i)\hat{y}_i^{(1)} + t_i\hat{y}_i^{(0)} \tag{20}$$

Next, the predicted loss $\mathcal{L}_i^p(\theta_1, \theta_0, \phi)$ is calculated as:

$$\mathcal{L}_i^p(\theta_1, \theta_0, \phi) = (\hat{y}_f^{(i)} - y_f^{(i)})^2 + (\hat{y}_{cf}^{(i)} - y_{cf}^{(i)})^2 + \alpha \text{BinaryCrossEntropy}(\pi(x_i), t_i) \tag{21}$$

where α is a hyperparameter. With the help of the propensity score $\pi(X)$ and the regressor $\mu(X, t)$, the doubly robust outcomes are calculated as

$$\hat{y}_{f_{DR}}^{(i)} = t_i\left[\frac{t_i\hat{y}_i^{(1)} - (t_i - \pi(x_i)\mu(x_i, t_i))}{\pi(x_i)}\right] + (1 - t_i)\left[\frac{(1 - t_i)\hat{y}_i^{(0)} - (t_i - \pi(X_i)\mu(x_i, t_i))}{1 - \pi(x_i)}\right] \tag{22}$$

$$\hat{y}_{cf_{DR}}^{(i)} = (1 - t_i)\left[\frac{(1 - t_i)\hat{y}_i^{(1)} - (t_i - \pi(x_i)\mu(x_i, t_i))}{\pi(x_i)}\right] + t_i\left[\frac{t_i\hat{y}_i^{(0)} - (t_i - \pi(x_i)\mu(x_i, t_i))}{1 - \pi(x_i)}\right] \tag{23}$$

The doubly robust loss $\mathcal{L}_i^{DR}(\theta_1, \theta_0, \theta_\phi, \theta_\mu, \phi)$ is calculated as:

$$\mathcal{L}_i^{DR}(\theta_1, \theta_0, \theta_\pi, \theta_\mu, \phi) = (\hat{y}_{f_{DR}}^{(i)} - y_f^{(i)})^2 + (\hat{y}_{cf_{DR}}^{(i)} - y_{cf}^{(i)})^2 \tag{24}$$

Finally, the loss function of the ITE is:

$$\mathcal{L}^{ITE}(\theta_1, \theta_0, \theta_\pi, \theta_\mu, \phi) = \frac{1}{n}\sum_{i=1}^{n}\left(\mathcal{L}_i^p + \beta\mathcal{L}_i^{DR}\right) \tag{25}$$

where β is a hyperparameter and the whole network is trained using end-to-end strategy .

4.4 Differences with CEVAE and GANITE

The counterfactual outcome predictor block of DR-VIDAL uses together VAE and GAN, used in CEVAE and GANITE, respectively. CEVAE also incorporates a causal graph, but it is more simplistic than DR-VIDAL, as it infers only the observed proxy X from Z. We instead consider multiple latent variables causally related to the treatment and the outcome in addition to the direct link to the pre-treatment covariates. Upon that, we use GAN to generate counterfactual examples, but, unlike GANITE, we first disentangle the multiple latent factors using a VAE, then we optimize the GAN with the mutual information, finally generating the entire potential outcome vector.

4.5 Differences with TARNet and Dragonnet

The design doubly robust model of DR-VIDAL is closely related to that of TARNet and Dragonnet. The doubly robust network without the propensity score and the regressor heads is the TARNET and only the he propensity score head is essentially the Dragonnet.

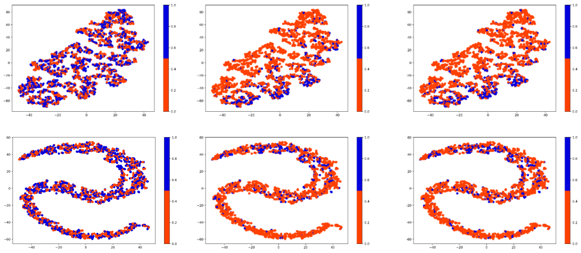

Figure 4: t-SNE visualization of the Latent representation of the Twins dataset, learned by the encoder of the counterfactual network. From the left to right, the plots show the representations of treatment, factual and counterfactual outcomes respectively. The top and bottom panels show the representations before and after training the network respectively.

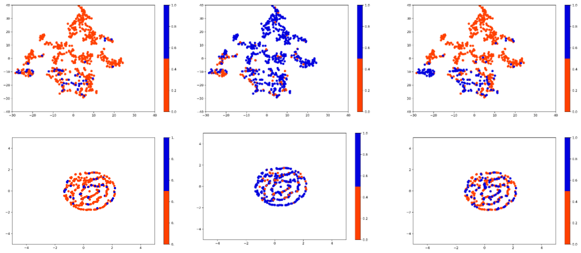

Figure 5: t-SNE visualization of the Latent representation of the Jobs dataset, learned by the encoder of the counterfactual network. From the left to right, the plots show the representations of treatment, factual and counterfactual outcomes respectively. The top and bottom panels show the representations before and after training the network respectively.

However the in TARNet, the weights corresponding to each of samples is calculated as the probability of the treatment assignment whereas in DR-VIDAL, the weights are calculated based on the propensity score which is the probability of treatment assignment given the covariates. For the Dragonnet, the targeted regularization was implemented without taking the regressed out into account which is estimated using the treatment assignment and the covariates together as in DR-VIDAL. Another major difference between TARNet and Dragonnet with DR-VIDAL is the training strategy employed. For both TARNet and Dragonnet, the counterfactual outcome does not exist, so for each sample the overall loss function has been estimated with the factual outcome only and the parameters of the outcome head of the factual outcome have been updated during training. In DR-VIDAL, we have the entire potential outcome vector comprising of both the factual and the counterfactual outcomes, so for each training sample, the loss function is calculated for both the outcomes and the corresponding parameters of both the outcome heads are updated as shown in Algorithm 2.

5 EXPERIMENTAL SETUP

5.1 Datasets

Estimating ITE from a observational dataset is dataset is a difficult task due to the unavailability of the counterfactuals. However, the importance of the dataset holds immense importance in estimating the ITE. In this work, we used two semi-synthetic datasets Infant Health and Development Program (IHDP)[17], Twins [3] and a real world dataset - Jobs [22]. These datasets are well described in [23, 31, 36]. In all the experiments for all the datasets, we used the same settings described in GANITE [36], where all the datasets are divided into 56/24/20 % train-validation-test splits. We ran 1000, 10 and 100 realizations of IHDP, Jobs and Twins datasets respectively and report the performance metrics on both the within and out-of-samples. We developed DR-VIDAL using Pytorch framework.

5.2 DR-VIDAL settings

Adversarial module To reduce the model complexity and parameters for the encoder of the VAE, we have a shared neural network connected to 4 other neural networks for estimating the 4 posterior distributions $q_{\phi_x}(z_x|x)$, $q_{\phi_t}(z_t|x)$, $q_{\phi_{yf}}(z_{yf}|x)$, $q_{\phi_{ycf}}(z_{ycf}|x)$. The shared neural network has 3 layers, each having 15 nodes. The networks with outputs as $q_{\phi_x}(z_x|x)$, $q_{\phi_t}(z_t|x)$, $q_{\phi_{yf}}(z_{yf}|x)$, $q_{\phi_{ycf}}(z_{ycf}|x)$ have a single layer with 5, 1, 1, 1 nodes respectively. The decoder is 4-layered neural network, each with 15 nodes to calculate the data likelihood $p_{\phi_d}(x|z_x, z_t, z_{yf}, z_{ycf})$. For the GAN, the generator network has 2 shared layers and 2 outcome specific layers for each outcomes, each with 100 nodes. The discriminator and Qnetwork is a 3-layered neural network, each with 30 nodes and 8 nodes respectively. All the layers of the VAE and GAN used RELU activation functions and the parameters were updated using Adam optimizer [20]. The random noise z_G is sampled from a 92-dimensional $\mathcal{N}(0, 1)$.

The hyperparameters γ is set as 1 for the 3 datasets and λ is set as 0.2, 0.01 and 10 for IHDP, Jobs and Twins datasets respectively. The batch sizes of IHDP, Jobs and Twins are set as 64, 64, 256 for the for IHDP, Jobs and Twins datasets respectively. The learning rate of the VAE, generator and discriminator is set as 1e-3, 1e-4, 5e-4 respectively.

Doubly robust module For the Doubly robust module, the shared network f_ϕ and outcome specific networks f_{θ_0} and f_{θ_1} are both 3-layered neural network each with 200 and 100 nodes respectively. The propensity network π has 2 layers each with 200 nodes. The regressor network μ has 6 layers with 200 nodes and 100 nodes in the first and last 3 layers respectively. All the layers of the VAE and GAN used ELU activation functions and the parameters were updated using Adam optimizer [20]. The batch sizes of IHDP, Jobs and Twins are set as 64, 64, 256 for the for IHDP, Jobs and Twins datasets respectively. We set the learning rate of all the networks as 1e-4 and the hyperparameters α and β are both set as 1 for all 3 datasets.

5.3 Validation and performance metrics

We report the expected Precision in Estimation of Heterogeneous Effect (PEHE) (ϵ_{PEHE}), average treatment effect (ATE) (ϵ_{ATE}) as discussed in [17, 31, 36] for datasets IHDP and Twins due to the availability of the factual and the counterfactual outcomes both. For Jobs, the counterfactual outcome does not exist, so we report the policy risk ($R_{pol}(\pi)$) and the error in average treatment effect on the treated (ATT) (ϵ_{ATT}) as mentioned by [31, 36].

$$\epsilon_{PEHE} = \frac{1}{N}\sum_{n=0}^{N}\Big(\mathbb{E}_{y_j(n)\sim\mu_j(n)}\big[y_1(n) - y_0(n)\big] - \big[\hat{y_1}(n) - \hat{y_0}(n)\big]\Big)^2 \tag{26}$$

$$\epsilon_{ATE} = ||\frac{1}{N}\sum_{n=0}^{N}\mathbb{E}_{y(n)\sim\mu(n)}\left[y(n)\right] - \frac{1}{N}\sum_{n=0}^{N}\hat{y}(n)||_2^2 \tag{27}$$

$$R_{pol}(\pi) = \frac{1}{N}\sum_{n=0}^{N}\Big[1 - \Big(\sum_{i=1}^{k}[\frac{1}{|\Pi_i \cap T_i \cap E|} \sum_{x(n)\in\Pi_i\cap\ T_i\cap E} y_i(n) \times \frac{|\Pi_n \cap E|}{|E|}]\Big)\Big] \tag{28}$$

where $\pi_i = \{\boldsymbol{x(n)} : i = \arg\max \hat{\boldsymbol{y}}$, $T_i = \boldsymbol{x(n)} : t_i(n) = 1\}$, and E is the randomized sample. For datasets where only factual outcomes are available with treatment being binary, such as Jobs and a randomized controlled trail(RCT) generates the testing set, the true average treatment effect on the treated (ATT) and the error ϵ_{ATT} are define by [31] as follows:

$$ATT = \frac{1}{|T_1 \cap E|}\sum_{x_i\in T_1\cap E} Y_1(x_i) - \frac{1}{|T_0 \cap E|}\sum_{x_i\in C\cap E} Y_0(x_i) \tag{29}$$

$$\epsilon_{ATT} = \left|ATT - \frac{1}{|T_1 \cap E|}\sum_{x_i\in T_1\cap E}\hat{Y}_1(x_i) - \hat{Y}_0(x_i)\right| \tag{30}$$

where T_1, T_0 and E are the subsets corresponding to treated, controlled samples and randomized controlled trials respectively.

6 RESULTS

We evaluated DR-VIDAL for 3 datasets and compare the performance with least squares regression using treatment as a feature (OLS/LR1), separate least squares regressions for each treatment (OLS/LR2), balancing linear regression (BLR) [19], k-nearest neighbor (k-NN) [10], Bayesian additive regression trees (BART) [9], random forests (RForest) [7], causal forests (C Forest) [35], balancing neural network (BNN) [19], treatment-agnostic representation network (TARNET) [31], counterfactual regression with Wasserstein distance (CFRW ASS) [31], CEVAE [23], and GANITE [36]. We report the performance for both the in-sample and out-of-sample in Table 2, 3 and 4 for IHDP, Jobs and Twins respectively. We can see that DR-VIDAL outperformed all the other models for Jobs dataset. For IHDP, it also surpassed all the models by a significant margin except TARNET and CFR$_{WASS}$. Due to the disentanglement of the hidden factors and the adversarial learning, the performance gain for the IHDP dataset was consequential compared to the other 2 generative models CEVAE and GANITE even with the large number of parameters of DR-VIDAL for a small dataset like IHDP. For Twins, the performance of DR-VIDAL is fairly competitive with most of the state of the art methods. The representations learned by the VAE of the adversarial model of DR-VIDAL for Twins and Jobs dataset is shown in Figures 4 and 5 respectively for both before and

	IHDP		Jobs		Twins	
	$\sqrt{\epsilon_{PEHE}^{out-of-s}}$	$\epsilon_{ATE}^{out-of-s}$	$R_{Pol}^{out-of-s}$	$\epsilon_{ATT}^{out-of-s}$	$\sqrt{\epsilon_{PEHE}^{out-of-s}}$	$\epsilon_{ATE}^{out-of-s}$
DR-VIDAL	**1.453 ± 0.11**	**0.45 ± 0.12**	**0.102 ± 0.01**	**0.056 ± 0.02**	**0.318 ± 0.008**	**0.0111 ± 0.0137**
DR-VIDAL (w/o DR loss)	1.476 ± 0.17	0.46 ± 0.18	0.110 ± 0.01	0.091 ± 0.04	0.324 ± 0.007	0.0131 ± 0.0152
DR-VIDAL (w/o Info loss)	1.461 ± 0.11	0.45 ± 0.12	0.109 ± 0.01	0.056 ± 0.02	0.318 ± 0.012	0.0115 ± 0.0171
DR-VIDAL (w/o DR loss & Info loss)	1.476 ± 0.13	0.46 ± 0.15	0.113 ± 0.01	0.094 ± 0.04	0.326 ± 0.008	0.0124 ± 0.0172

Table 1: Performance on the out-of-sample test sets (mean ± st.dev) of the all the variants of DR-VIDAL algorithm on 100, 10, 100 realizations of the IHDP, Jobs and Twins datasets respectively.

Methods	$\sqrt{\epsilon_{PEHE}^{within-s}}$	$\epsilon_{ATE}^{within-s}$	$\sqrt{\epsilon_{PEHE}^{out-of-s}}$	$\epsilon_{ATE}^{out-of-s}$
OLS/LR1	5.8 ± 0.3	0.73 ± 0.04	5.8 ± 0.3	0.94 ± 0.06
OLS/LR2	2.4 ± 0.1	0.14 ± 0.01	2.5 ± 0.1	0.31 ± 0.02
BLR	5.8 ± 0.3	0.72 ± 0.04	5.8 ± 0.3	0.93 ± 0.05
k-NN	2.1 ± 0.1	0.14 ± 0.01	4.1 ± 0.2	0.79 ± 0.05
BART	2.1 ± 0.2	0.23 ± 0.01	2.3 ± 0.1	0.34 ± 0.02
R Forest	4.2 ± 0.2	0.73 ± 0.05	6.6 ± 0.3	0.96 ± 0.06
C Forest	3.8 ± 0.2	0.18 ± 0.01	3.8 ± 0.2	0.40 ± 0.03
BNN	2.2 ± 0.1	0.37 ± 0.03	2.1 ± 0.1	0.42 ± 0.03
TARNET	0.88 ± 0.02	0.26 ± 0.01	0.95 ± 0.02	0.28 ± 0.01
CFR$_{WASS}$	0.71 ± 0.0	0.25 ± 0.01	0.76 ± 0.0	0.27 ± 0.01
GANITE	1.9 ± 0.4	0.43 ± 0.05	2.4 ± 0.4	0.49 ± 0.05
CEVAE	2.7 ± 0.1	0.34 ± 0.01	2.6 ± 0.1	0.46 ± 0.02
DR-VIDAL	**1.44 ± 0.03**	**0.44 ± 0.11**	**1.46 ± 0.12**	**0.44 ± 0.19**

Table 2: Performance on the within-sample and out-of-sample test sets (mean ± st.dev) of various models on the IHDP dataset.

Methods	$R_{Pol}^{within-s}$	$\epsilon_{ATT}^{within-s}$	$R_{Pol}^{out-of-s}$	$\epsilon_{ATT}^{out-of-s}$
OLS/LR1	0.22 ± 0.0	0.01 ± 0.00	0.23 ± 0.0	0.08 ± 0.04
OLS/LR2	0.21 ± 0.0	0.01 ± 0.01	0.24 ± 0.0	0.08 ± 0.03
BLR	0.22 ± 0.0	0.01 ± 0.01	0.25 ± 0.0	0.08 ± 0.03
k-NN	0.02 ± 0.0	0.21 ± 0.01	0.26 ± 0.0	0.13 ± 0.05
BART	0.23 ± 0.0	0.02 ± 0.00	0.25 ± 0.0	0.08 ± 0.03
R Forest	0.23 ± 0.0	0.03 ± 0.01	0.28 ± 0.0	0.09 ± 0.04
C Forest	0.19 ± 0.0	0.03 ± 0.01	0.20 ± 0.0	0.07 ± 0.03
BNN	0.20 ± 0.0	0.03 ± 0.01	0.24 ± 0.0	0.09 ± 0.04
TARNET	0.17 ± 0.0	0.05 ± 0.02	0.21 ± 0.0	0.11 ± 0.04
CFR$_{WASS}$	0.17 ± 0.0	0.04 ± 0.01	0.21 ± 0.0	0.09 ± 0.03
GANITE	0.13 ± 0.01	0.01 ± 0.01	0.14 ± 0.01	0.06 ± 0.03
CEVAE	0.15 ± 0.0	0.02 ± 0.01	0.26 ± 0.0	0.03 ± 0.01
DR-VIDAL	**0.09 ± 0.005**	**0.04 ± 0.03**	**0.10 ± 0.01**	**0.05 ± 0.02**

Table 3: Performance on the within-sample and out-of-sample test sets (mean ± st.dev) of various models on the Jobs dataset.

Methods	$\sqrt{\epsilon_{PEHE}^{within-s}}$	$\epsilon_{ATE}^{within-s}$	$\sqrt{\epsilon_{PEHE}^{out-of-s}}$	$\epsilon_{ATE}^{out-of-s}$
OLS/LR1	0.319 ± 0.005	0.0038 ± 0.0025	0.297 ± 0.016	0.0069 ± 0.0056
OLS/LR2	0.320 ± 0.001	0.0039 ± 0.0025	0.318 ± 0.007	0.0070 ± 0.0059
BLR	0.312 ± 0.002	0.0057 ± 0.0036	0.320 ± 0.003	0.0334 ± 0.0092
k-NN	0.333 ± 0.003	0.0028 ± 0.0021	0.323 ± 0.018	0.0051 ± 0.0039
BART	0.347 ± 0.009	0.1206 ± 0.0236	0.338 ± 0.016	0.1265 ± 0.0234
R Forest	0.306 ± 0.002	0.0049 ± 0.0034	0.321 ± 0.005	0.0080 ± 0.0051
C Forest	0.366 ± 0.003	0.0286 ± 0.0035	0.316 ± 0.011	0.0335 ± 0.0083
BNN	0.325 ± 0.003	0.0056 ± 0.0032	0.321 ± 0.018	0.0203 ± 0.0071
TARNET	0.317 ± 0.005	0.0108 ± 0.0017	0.315 ± 0.003	0.0151 ± 0.0018
CFR$_{WASS}$	0.315 ± 0.007	0.0112 ± 0.0016	0.313 ± 0.008	0.0284 ± 0.0032
GANITE	0.289 ± 0.12	0.0058 ± 0.0017	0.297 ± 0.05	0.0089 ± 0.0075
CEVAE	n.r	n.r	n.r	n.r
DR-VIDAL	**0.317 ± 0.002**	**0.0102 ± 0.0128**	**0.318 ± 0.008**	**0.0111 ± 0.0137**

Table 4: Performance on the within-sample and out-of-sample test sets (mean ± st.dev) of various models on the Twins dataset.

after training. From these figures, we can observe how the learned representations are clustered within a small space after we train the model. Also Table 1 shows the different variations of DR-VIDAL and how DR-VIDAL with the doubly robust loss and information loss performs the best for all the three datasets.

7 DISCUSSION

In this work, DR-VIDAL - a novel deep learning based model showed the power of adversarial representation learning with doubly robust regression. However, there are some limitation of this work. One limitation is the causal graph that we assumed to work on is fairly simple. Also it will be quite, interesting to see how TARNET and Dragonnet will perform as a downstream model after the counterfactuals, generated from the adversarial network of Dr-VIDAL. Another possible extension will be the usage of attention in the encoded representations in VAE and while calculating the propensity score which will give us the more important covariates in the covariate space.

In conclusion, DR-VIDAL framework is a promising approach to estimate the counterfactuals and the ITE. Our experiments proved that this method outperforms the state-of-the-art models empirically making the estimation more robust.

REFERENCES

[1] Ahmed Alaa and Mihaela van der Schaar. 2018. Limits of Estimating Heterogeneous Treatment Effects: Guidelines for Practical Algorithm Design. In *Proceedings of the 35th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 80)*, Jennifer Dy and Andreas Krause (Eds.). PMLR, Stockholmsmässan, Stockholm Sweden, 129–138. http://proceedings.mlr.press/v80/alaa18a.html

[2] Ahmed M Alaa, Michael Weisz, and Mihaela Van Der Schaar. 2017. Deep counterfactual networks with propensity-dropout. *arXiv preprint arXiv:1706.05966* (2017).

[3] Douglas Almond, Kenneth Y Chay, and David S Lee. 2005. The costs of low birth weight. *The Quarterly Journal of Economics* 120, 3 (2005), 1031–1083.

[4] Susan Athey and Guido Imbens. 2016. Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences* 113, 27 (2016), 7353–7360.

[5] Peter C Austin. 2011. An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate behavioral research* 46, 3 (2011), 399–424.

[6] David M Blei, Alp Kucukelbir, and Jon D McAuliffe. 2017. Variational inference: A review for statisticians. *Journal of the American statistical Association* 112, 518 (2017), 859–877.

[7] Leo Breiman. 2001. Random forests. *Machine learning* 45, 1 (2001), 5–32.

[8] Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. 2016. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. *Advances in neural information processing systems* 29 (2016), 2172–2180.

[9] Hugh A Chipman, Edward I George, Robert E McCulloch, et al. 2010. BART: Bayesian additive regression trees. *The Annals of Applied Statistics* 4, 1 (2010), 266–298.

[10] Richard K Crump, V Joseph Hotz, Guido W Imbens, and Oscar A Mitnik. 2008. Nonparametric tests for treatment effect heterogeneity. *The Review of Economics and Statistics* 90, 3 (2008), 389–405.

[11] Rajeev H Dehejia and Sadek Wahba. 2002. Propensity score-matching methods for nonexperimental causal studies. *Review of Economics and statistics* 84, 1 (2002), 151–161.

[12] Miroslav Dudík, Dumitru Erhan, John Langford, Lihong Li, et al. 2014. Doubly robust policy evaluation and optimization. *Statist. Sci.* 29, 4 (2014), 485–511.

[13] Michele Jonsson Funk, Daniel Westreich, Chris Wiesen, Til Stürmer, M Alan Brookhart, and Marie Davidian. 2011. Doubly robust estimation of causal effects. *American journal of epidemiology* 173, 7 (2011), 761–767.

[14] MM Garrido et al. 2014. Methods for Constructing and Assessing Propensity Scores. *Health Services Research* 49, 5 (2014), 1701––20.

[15] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. *Advances in neural information processing systems* 27 (2014), 2672–2680.

[16] M.A. Hernan and J.M. Robins. 2019. *Causal Inference.* Taylor & Francis. https://books.google.com/books?id=_KnHIAAACAAJ

[17] Jennifer L Hill. 2011. Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics* 20, 1 (2011), 217–240.

[18] Guido W Imbens. 2000. The role of the propensity score in estimating dose-response functions. *Biometrika* 87, 3 (2000), 706–710.

[19] Fredrik Johansson, Uri Shalit, and David Sontag. 2016. Learning representations for counterfactual inference. In *International conference on machine learning.* 3020–3029.

[20] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[21] Diederik P Kingma and Max Welling. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114* (2013).

[22] Robert J LaLonde. 1986. Evaluating the econometric evaluations of training programs with experimental data. *The American economic review* (1986), 604–620.

[23] Christos Louizos, Uri Shalit, Joris M Mooij, David Sontag, Richard Zemel, and Max Welling. 2017. Causal effect inference with deep latent-variable models. In *Advances in neural information processing systems.* 6446–6456.

[24] Min Lu, Saad Sadiq, Daniel J Feaster, and Hemant Ishwaran. 2018. Estimating individual treatment effect in observational data using random forest methods. *Journal of Computational and Graphical Statistics* 27, 1 (2018), 209–219.

[25] Jared K Lunceford and Marie Davidian. 2004. Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Statistics in medicine* 23, 19 (2004), 2937–2960.

[26] J. Pearl, M. Glymour, and N.P. Jewell. 2016. *Causal Inference in Statistics: A Primer.* Wiley. https://books.google.com/books?id=L3G-CgAAQBAJ

[27] Kristin E Porter, Susan Gruber, Mark J Van Der Laan, and Jasjeet S Sekhon. 2011. The relative performance of targeted maximum likelihood estimators. *The International Journal of Biostatistics* 7, 1 (2011).

[28] Mattia Prosperi, Yi Guo, Matt Sperrin, James S. Koopman, Jae S. Min, Xing He, Shannan Rich, Mo Wang, Iain E. Buchan, and Jiang Bian. 2020. Causal inference and counterfactual prediction in machine learning for actionable healthcare. *Nature Machine Intelligence* 2 (2020), 369–375. https://doi.org/10.1038/s42256-020-0197-y

[29] Paul R Rosenbaum and Donald B Rubin. 1983. The central role of the propensity score in observational studies for causal effects. *Biometrika* 70, 1 (04 1983), 41–55. https://doi.org/10.1093/biomet/70.1.41

[30] Donald B Rubin. 1974. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology* 66, 5 (1974), 688–701.

[31] Uri Shalit, Fredrik D. Johansson, and David Sontag. 2017. Estimating individual treatment effect: generalization bounds and algorithms. In *Proceedings of the 34th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 70)*, Doina Precup and Yee Whye Teh (Eds.). PMLR, International Convention Centre, Sydney, Australia, 3076–3085. http://proceedings.mlr.press/v70/shalit17a.html

[32] Claudia Shi, David Blei, and Victor Veitch. 2019. Adapting neural networks for the estimation of treatment effects. In *Advances in neural information processing systems.* 2507–2517.

[33] Bonnie Sibbald and Martin Roland. 1998. Understanding controlled trials: Why are randomised controlled trials important? *BMJ* 316, 7126 (1998), 201. https://doi.org/10.1136/bmj.316.7126.201

[34] Yuxi Tian, Martijn J Schuemie, and Marc A Suchard. 2018. Evaluating large-scale propensity score performance through real-world and synthetic data experiments. *International journal of epidemiology* 47, 6 (2018), 2005–2014.

[35] Stefan Wager and Susan Athey. 2018. Estimation and inference of heterogeneous treatment effects using random forests. *J. Amer. Statist. Assoc.* 113, 523 (2018), 1228–1242.

[36] Jinsung Yoon, James Jordon, and Mihaela van der Schaar. 2018. GANITE: Estimation of Individualized Treatment Effects using Generative Adversarial Nets. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings.* OpenReview.net. https://openreview.net/forum?id=ByKWUeWA-

A APPENDIX

A.1 Derivation of the Loss function ELBO loss of VAE

From Figure 1, $p_{\phi_d}(x|z_x, z_t, z_{yf}, z_{y_cf})$ and $p_{\phi_d}(z_x, z_t, z_{yf}, z_{y_cf}|x)$ are the true likelihood and true posterior respectively. The posterior is hard to evaluate, so we have to approximate the true posterior to the product of the factorized known distributions $q_{\phi_x}(z_x|x)$, $q_{\phi_t}(z_t|x)$, $q_{\phi_{yf}}(z_{yf}|x)$ and $q_{\phi_{ycf}}(z_{ycf}|x)$ by minimising the KL divergence as follows,

$$
\begin{aligned}
&KL\big(q_{\phi_x}(z_x|x)q_{\phi_t}(z_t|x)q_{\phi_{yf}}(z_{yf}|x)q_{\phi_{ycf}}(z_{ycf}|x)||\\
&\quad p_{\phi_d}(z_x, z_t, z_{yf}, z_{ycf}|x)\big)\\
&= \int\int\int\int q_{\phi_x}(z_x|x)q_{\phi_t}(z_t|x)q_{\phi_{yf}}(z_{yf}|x)q_{\phi_{ycf}}(z_{ycf}|x)\\
&\quad \big[\log \frac{q_{\phi_x}(z_x|x)q_{\phi_t}(z_t|x)q_{\phi_{yf}}(z_{yf}|x)}{p_{\phi_d}(z_x, z_t, z_{yf}, z_{ycf}|x)}\big] dz_x dz_t dz_{yf} dz_{ycf}\\
&= \int\int\int\int q_{\phi_x}(z_x|x)q_{\phi_t}(z_t|x)q_{\phi_{yf}}(z_{yf}|x)q_{\phi_{ycf}}(z_{ycf}|x)\\
&\quad \big[\log q_{\phi_x}(z_x|x) + \log q_{\phi_t}(z_t|x)\\
&\quad + \log q_{\phi_{yf}}(z_{yf}|x) + \log q_{\phi_{ycf}}(z_{ycf}|x)\\
&\quad - \log p_{\phi_d}(z_x, z_t, z_{yf}, z_{ycf}|x)\big] dz_x dz_t dz_{yf} dz_{ycf}\\
&= \int\int\int\int q_{\phi_x}(z_x|x)q_{\phi_t}(z_t|x)q_{\phi_{yf}}(z_{yf}|x)q_{\phi_{ycf}}(z_{ycf}|x)\\
&\quad \big[\log q_{\phi_x}(z_x|x) + \log q_{\phi_t}(z_t|x)\\
&\quad + \log q_{\phi_{yf}}(z_{yf}|x) + \log q_{\phi_{ycf}}(z_{ycf}|x)\\
&\quad - \log p_{\phi_d}(x|z_x, z_t, z_{yf}, z_{ycf}) - \log p_{\phi_d}(z_x, z_t, z_{yf}, z_{ycf})\\
&\quad + \log p_{\phi_d}(x)\big] dz_x dz_t dz_{yf} dz_{ycf}\\
&= \int q_{\phi_x}(z_x|x) \log \frac{q_{\phi_x}(z_x|x)}{p_{\phi_d}(z_x)} dz_x\\
&\quad + \int q_{\phi_t}(z_t|x) \log \frac{q_{\phi_t}(z_t|x)}{p_{\phi_d}(z_t)} dz_t\\
&\quad + \int q_{\phi_{yf}}(z_{yf}|x) \log \frac{q_{\phi_{yf}}(z_{yf}|x)}{p_{\phi_d}(z_{yf})} dz_{yf}\\
&\quad + \int q_{\phi_{ycf}}(z_{ycf}|x) \log \frac{q_{\phi_x}(z_{ycf}|x)}{p_{\phi_d}(z_x)} dz_{ycf}\\
&\quad - \int\int\int\int \big[q_{\phi_x}(z_x|x)q_{\phi_t}(z_t|x)q_{\phi_{yf}}(z_{yf}|x)\\
&\quad q_{\phi_{ycf}}(z_{ycf}|x) \log p_{\phi_d}(x|z_x, z_t, z_{yf}, z_{ycf})\big] dz_x dz_t dz_{yf} dz_{ycf}\\
&\quad + \log p_{\phi_d}(x)\\
&= KL\big(q_{\phi_x}(z_x|x)||p_{\phi_d}(z_x))\big) + KL\big(q_{\phi_t}(z_t|x)||p_{\phi_d}(z_t))\big)\\
&\quad + KL\big(q_{\phi_{yf}}(z_{yf}|x)||p_{\phi_d}(z_{yf}))\big)\\
&\quad + KL\big(q_{\phi_{ycf}}(z_{ycf}|x)||p_{\phi_d}(z_{ycf}))\big)\\
&\quad - \mathbb{E}_{q_{\phi_x}, q_{\phi_t}, q_{\phi_{yf}}, q_{\phi_{ycf}}}[\log p(x|z_x, z_t, z_{yf}, z_{ycf})]\\
&\quad + \log p_{\phi_d}(x)
\end{aligned}
$$

where, the distributions $q_{\phi_x}(z_x|x)$, $q_{\phi_t}(z_t|x)$, $q_{\phi_{yf}}(z_{yf}|x)$, $q_{\phi_{ycf}}(z_{ycf}|x)$ and $p_{\phi_d}(x|z_x, z_t, z_{yf}, z_{ycf}|x)$ are parameterized by the parameters $\phi_x, \phi_t, \phi_{yf}, \phi_{ycf}, \phi_d$. The KL divergence of two distributions is always greater than or equal to zero. So,

$$
\begin{aligned}
&KL\big(q_{\phi_x}(z_x|x)q_{\phi_t}(z_t|x)q_{\phi_{yf}}(z_{yf}|x)q_{\phi_{ycf}}(z_{ycf}|x)||\\
&\quad p_{\phi_d}(z_x, z_t, z_{yf}, z_{ycf}|x)\big) \geq 0,\\
&\log p_{\phi_d}(x) \geq \mathcal{L}_{ELBO} \qquad \text{where,}\\
&\mathcal{L}_{ELBO}(\phi_x, \phi_t, \phi_{yf}, \phi_{ycf}; x, z_x, z_t, z_{yf}, z_{ycf})\\
&= \mathbb{E}_{q_{\phi_x}, q_{\phi_t}, q_{\phi_{yf}}, q_{\phi_{ycf}}}[\log p(x|z_x, z_t, z_{yf}, z_{ycf})]\\
&\quad - KL\big(q_{\phi_x}(z_x|x)||p_{\phi_d}(z_x))\big) - KL\big(q_{\phi_t}(z_t|x)||p_{\phi_d}(z_t))\big)\\
&\quad - KL\big(q_{\phi_{yf}}(z_{yf}|x)||p_{\phi_d}(z_{yf}))\big)\\
&\quad - KL\big(q_{\phi_{ycf}}(z_{ycf}|x)||p_{\phi_d}(z_{ycf}))\big)
\end{aligned}
$$

A.2 Variational information maximization

$$
\begin{aligned}
&I(z_c; G(z_G, z_c)) = H(z_c) - H(z_c|G(z_G, z_c))\\
&= H(z_c) + \int\int p(Z_c = z'_c, X = G(z_G, z_c))\\
&\qquad \log p(Z_c = z'_c|X = G(z_G, z_c))\, dz_c dx\\
&= H(z_c) + \mathbb{E}_{x\sim G(z_G,z_c)}\mathbb{E}_{z'_c\sim p(z_c|x)} \log(p(z'_c|x))\\
&= H(z_c) + \mathbb{E}_{x\sim G(z_G,z_c)}\mathbb{E}_{z'_c\sim p(z_c|x)} \log\left[\frac{p(z'_c|x)}{Q(z'_c|x)} Q(z'_c|x)\right]\\
&= H(z_c) + \mathbb{E}_{x\sim G(z_G,z_c)}\mathbb{E}_{z'_c\sim p(z_c|x)} \log\left[\frac{p(z'_c|x)}{Q(z'_c|x)}\right]\\
&\qquad + \mathbb{E}_{x\sim G(z,c)}\mathbb{E}_{z'_c\sim p(z_c|x)} \log\left[Q(z'_c|x)\right]\\
&= H(z_c) + \mathbb{E}_{x\sim G(z,c)} \int p(z'_c|x) \log \frac{p(z'_c|x)}{Q(z'_c|x)} dc'\\
&\qquad + \mathbb{E}_{x\sim G(z,c)}\mathbb{E}_{c'\sim p(z_c|x)} \log\left[Q(z'_c|x)\right]\\
&= H(z_c) + \mathbb{E}_{x\sim G(z,c)}\big[KL\big(p(z'_c|x||Q(z'_c|x))\big)\big]\\
&\qquad + \mathbb{E}_{x\sim G(z,c)}\mathbb{E}_{z'_c\sim p(z_c|x)} \log\left[Q(z'_c|x)\right]\\
&\geq H(z_c) + \mathbb{E}_{x\sim G(z,c)}\mathbb{E}_{z'_c\sim p(z_c|x)} \log\left[Q(z'_c|x)\right]\\
&\geq H(z_c) + \mathbb{E}_{z_c\sim p(z_c)}\mathbb{E}_{x\sim G(z,c)}\mathbb{E}_{z'_c\sim p(z_c|x)} \log\left[Q(z'_c|x)\right]\\
&\geq H(z_c) + \mathbb{E}_{z_c\sim p(z_c)}\mathbb{E}_{x\sim G(z,c)} \log\left[Q(z_c|x)\right]\\
&\text{(by Lemma 5.1 of [8])}\\
&= L_I(G, Q)
\end{aligned}
$$