PITE: Multi-Prototype Alignment for Individual Treatment Effect Estimation

Fuyuan Cao[1,2], **Jiaxuan Zhang**[1,*], **Xiaoli Li**[3]

[1]School of Computer and Information Technology, Key Laboratory of Computational Intelligence and Chinese Information Processing of Ministry of Education, Shanxi University, Taiyuan, China
[2]Shanxi Taihang Laboratory, Taiyuan, China
[3]Singapore University of Technology and Design, Singapore
cfy@sxu.edu.cn, jiaxuan7531@163.com, xiaoli_li@sutd.edu.sg

Abstract

Estimating Individual Treatment Effects (ITE) from observational data is challenging due to confounding bias. Most studies tackle this bias by balancing distributions globally, but ignore individual heterogeneity and fail to capture the local structure that represents the natural clustering among individuals, which ultimately compromises ITE estimation. While instance-level alignment methods consider heterogeneity, they similarly overlook the local structure information. To address these issues, we propose an end-to-end Multi-**P**rototype alignment method for **ITE** estimation (**PITE**). PITE effectively captures local structure within groups and enforces cross-group alignment, thereby achieving robust ITE estimation. Specifically, we first define prototypes as cluster centroids based on similar individuals under the same treatment. To identify local similarity and the distribution consistency, we perform instance-to-prototype matching to assign individuals to the nearest prototype within groups, and design a multi-prototype alignment strategy to encourage the matched prototypes to be close across treatment arms in the latent space. PITE not only reduces distribution shift through fine-grained, prototype-level alignment, but also preserves the local structures of treated and control groups, which provides meaningful constraints for ITE estimation. Extensive evaluations on benchmark datasets demonstrate that PITE outperforms 13 state-of-the-art methods, achieving more accurate and robust ITE estimation.

Extended version — http://arxiv.org/abs/2511.10320

Introduction

Estimating the Individual Treatment Effects (ITE) from observational data is critical for personalized decision-making in fields such as healthcare and E-commerce (Li et al. 2024; Liu, Wei, and Zhang 2021; Chu et al. 2022; Bica and Van der Schaar 2022), where understanding the causal impact of interventions guides critical decisions at an individual level. Unlike randomized controlled trials (RCTs), observational studies suffer from *confounding bias* due to confounders (Kong et al. 2023; Cheng, Hardt, and Mendler-Dünner 2024), variables that influence both treatments and outcomes, which make treated and control groups follow different covariate distributions.

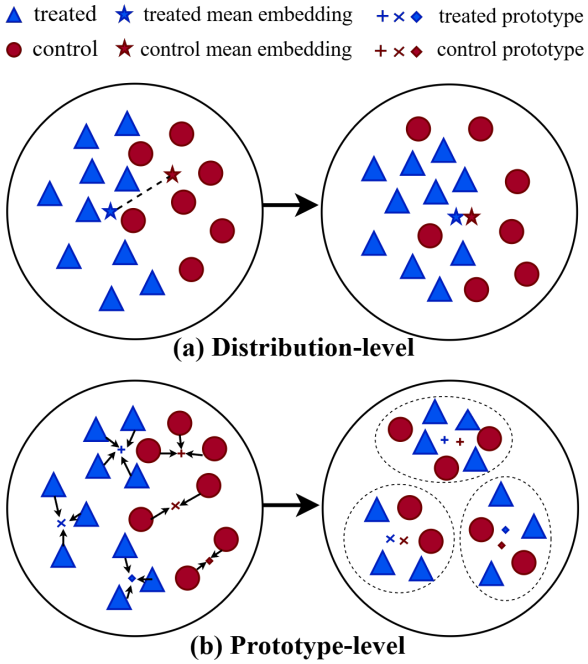

Figure 1: (a) Existing distribution-level alignment methods align the overall distributional statistics (e.g., mean) of covariates between treated and control groups but fail to preserve semantic information. (b) Our method achieves both fine-grained prototype-level alignment to reduce distribution shift and preserves local structure between treated and control groups.

One popular approach for handling confounding bias in treatment effect estimation is distribution-level covariate balance, which aligns the overall distributional statistics of treated and control groups, as shown in Figure 1 (a). For example, Maximum Mean Discrepancy (MMD) minimizes the distribution discrepancy between treated and control groups by aligning their mean representations. Similarly, adversarial training (Zhou et al. 2022; Du et al. 2021) makes factual and counterfactual distributions indistinguishable, naturally mitigating distribution shift. And optimal transport

*Corresponding author

methods (Yan et al. 2024; Li et al. 2021) achieve distribution alignment by moving masses from one distribution to another with minimal transport cost, effectively reducing distribution bias. However, these methods only achieve global distribution balance, neglecting individual-level heterogeneity and the underlying local structures that reflect natural clustering among individuals, which ultimately compromises ITE estimation.

Recent works adapt instance-level alignment methods to address individual heterogeneity, but still neglect local structure constraints during the representation learning process. They generally employ contrastive learning to learn an embedding space where 'positive samples' are pulled closer together and 'negative samples' are pushed apart (Li and Yao 2022; Zhang et al. 2025). However, the instance-to-instance matching overemphasizes the characteristics of each instance, leading to poor generalization. Meanwhile, these methods also disrupt the natural clustering structure of data during the representation learning process, which degrades the estimation performance.

The local structural information overlooked by both distribution-level and sample-level alignment methods, provides meaningful constraint for ITE estimation. For example, in precision medicine, these methods ignore the inherent patient subgroups during representation learning. Patients typically fall into three response categories based on drug sensitivity: normal responders, hyper-responders, and low responders (Feuerriegel et al. 2024). This leads to mismatched pairs where, for instance, a hyper-responder from the treated group might be incorrectly matched with the subgroup of normal responders or low responders from the control group, rather than with the subgroup of hyper-responders. Such prototype-agnostic alignment frequently produces significant errors for hyper-responders, who risk adverse effects.

To overcome these limitations, we propose a Multi-**P**rototype alignment method for **ITE** estimation (**PITE**). PITE effectively captures local structure within groups and enforces cross-group alignment. Specifically, we first define prototypes as cluster centroids based on similar individuals under the same treatment, and then integrate two key techniques: (1) Within-group Prototype Matching, which performs instance-to-prototype matching to assign individuals to the nearest prototype within groups. Instead of global matching, matching a sample to a prototype is more robust to abnormal instances, especially in scenarios with significant individual heterogeneity. (2) Cross-group Prototype Alignment, which establishes correspondence between treated and control prototypes to encourage the matched prototypes to be close in the latent space. This dual strategy in PITE enables robust prototype-level alignment, effectively mitigating distribution shift while preserving local structure similarity, thereby making PITE more accurate and robust for instance-level treatment effect predictions. The proposed prototype-level alignment method introduces k prototypes, where the flexible choice of k unifies distribution-level and instance-level alignment. When $k = 1$, it performs distribution-level alignment; when $k = n$, it aligns instances. For $1 < k < n$, this group-level alignment effectively balances the trade-off between global and individual.

Our main contributions are summarized as follows:

- We define prototypes as cluster centroids of similar instances and perform instance-to-prototype matching, thereby capturing the local structure constraints within groups.
- We provide a novel algorithm, PITE, to capture local structure within groups and enforces cross-group alignment for individual treatment effects estimation.
- We conduct a comprehensive evaluation of PITE. Importantly, we find that PITE significantly outperforms distribution-level and instance-level methods, with up to 33.8% and 39.3% reduction in estimation error on IHDP, achieving more accurate ITE estimation.

Related Work

Recently, numerous deep learning studies have analyzed the relationship between treatment and outcome at the individual level through mitigating distribution shift, which can be broadly categorized into distribution-level alignment and instance-level alignment methods.

Distribution-Level Alignment

Current distribution-level alignment methods aim to balance the distributions globally by learning first-order moments, primarily employing distance metrics, adversarial training, and optimal transport techniques. For example, Shalit, Johansson, and Sontag (2017) developed TARNet / CFRNet to mitigate confounding bias by reducing the distribution divergence between treated and control groups in the representation space, adopting Maximum Mean Discrepancy (MMD) and Wasserstein distance. GANITE (Yoon, Jordon, and Van Der Schaar 2018) utilized adversarial training to make the discriminator unable to distinguish whether the input data come from the factual distribution or the generate counterfactual distribution. CBRE (Zhou et al. 2022) introduced an information loop to preserve predictive information that might otherwise be lost during the raw-to-latent space transformation in adversarial training. Alternatively, optimal transport-based methods have also been explored, where Yan et al. (2024) reduces the balancing error under the framework of optimal transport with learnable marginal distributions and the cost function. Similarly, Wang et al. (2023) proposed an estimator based on optimal transport to handle both mini-batch sampling effects and unobserved confounder effects issues.

While these methods focus on global distributional alignment, they often neglect the individual-level heterogeneity and intrinsic structure of data such as subgroup similarity or local clustering, which leads to less informative representations and compromises ITE estimation.

Instance-Level Alignment

Instance-level alignment methods work by matching similar units from different groups to construct locally balanced distributions. The propensity score matching (Rosenbaum and Rubin 1983) computes unit similarity based on propensity scores. Instead, representation learning-based methods

perform instance-level alignment in learned representation spaces. For example, SITE (Yao et al. 2018) employed representation learning to capture instance-level variation by selecting specific sample pairs for alignment in the learned embedding space. Similarly, Li and Yao (2022) designed a contrastive task for ITE estimation based on propensity score learning within a representation framework, regarding samples with propensity scores close to 0.5 as positive samples to learn balanced representations. FCCL (Zhang et al. 2025) further integrated diffeomorphic counterfactual generation and contrastive learning to address distribution shift through instance-level alignment in the representation space. However, these approaches only achieve partial balance and fail to effectively mitigate the distribution shift.

Compared with instance-level alignment methods, we not only account for individual heterogeneity by performing instance-to-prototype matching that preserves local structural information, but also achieve distributional balance across treatment groups through prototype-level alignment in the latent space, thereby enabling more robust and accurate ITE estimation.

Preliminary

Following the Neyman-Rubin potential outcome framework (Rubin 2005; Shalit, Johansson, and Sontag 2017), we formally define the problem setup. Let $\mathcal{X} \subset \mathbb{R}^d$ denote the d-dimensional covariate space, $\mathcal{T} = \{0, 1\}$ represent the binary treatment space, and $\mathcal{Y} \subset \mathbb{R}$ denote the potential outcome space. We assume the observed dataset contains n independent and identically distributed samples, represented as $\mathcal{D} = \{x_i, t_i, y_i\}_{i=1}^n$. For each sample, the covariates are denoted by $x_i \in \mathcal{X}$, and the treatment assignment is defined by the binary variable $t_i \in \mathcal{T}$, where $t_i = 0$ indicates that the i-th sample belongs to the control group, and $t_i = 1$ indicates that the i-th sample belongs to the treatment group. Each sample has two potential outcomes: y_i^0 represents the potential outcome for the i-th sample when not receiving treatment, and y_i^1 represents the potential outcome for the i-th sample when receiving treatment. The actually observed outcome $y_i^{t_i} \in \mathcal{Y}$ only reflects the result under the sample's actual assigned treatment status (*i.e.*, the factual outcome), while the outcome under the unassigned status (the counterfactual outcome $y_i^{1-t_i}$) cannot be directly observed. The observed outcome can be expressed as: $y_i = (1 - t_i)\, y_i^0 + t_i y_i^1$.

We illustrate with a drug development example that analyzes the efficacy of a newly developed medication for specific patients. In this context, treatment assignment t_i indicates whether a patient received the new medication ($t_i = 1$) or no treatment ($t_i = 0$). The patient's covariates x_i include baseline clinical characteristics such as sex, age, weight, etc. The outcomes y_i^1 and y_i^0 represent the patient's blood sugar levels with and without the new medication, respectively.

The individual treatment effect (ITE) of sample i is defined as the difference between the potential treatment and control outcomes:

$$\text{ITE}_i = y_i^1 - y_i^0. \tag{1}$$

We made the following assumptions to ensure that treatment effects are identifiable:

Assumption 1 (Consistency)**.** *For a unit with treatment assignment t, the observed outcome equals potential outcome y^t.*

Assumption 2 (Ignorability)**.** *The potential outcomes are independent of the treatment conditioning on covariates, such that $(y^1, y^0) \perp\!\!\!\perp t|x$.*

Assumption 3 (Overlap)**.** *For any x, the probability of receiving treatment is positive. That is, $0 < P(t = 1|x) < 1$, for $\forall x \in \mathcal{X}$.*

Methodology

We propose a novel Multi-**P**rototype Alignment framework for **I**ndividual **T**reatment **E**ffect Estimation (PITE), which integrates three key techniques: (1) Within-group Prototype Matching, which performs instance-to-prototype matching to assign individuals to the nearest prototype; (2) Cross-group Prototype Alignment, which enforces correspondence between matched prototypes across treatment arms; (3) Two-head prediction networks, which predict potential outcomes for treatment and control groups separately based on the learned balanced representations. The overall model architecture is presented in Figure 2.

Within-Group Prototype Matching

Prototypes serve as representative embeddings of semantically similar samples (Yue et al. 2021; An et al. 2024), providing a stable representation that is less sensitive to individual heterogeneity. Therefore, we first define prototypes and leverage prototypes to identify the natural clustering structures among individuals, thereby reducing bias caused by subgroup differences.

Definition 1 (Prototype)**.** *A prototype is defined as a learnable cluster centroid that represents a group of individuals with similar hidden representations under the same treatment condition. Formally, for each group $t \in \{0, 1\}$, PITE maintains a set of K prototypes:*

$$\mu_t = \{\mu_{t,k}\}_{k=1}^K \in \mathbb{R}^{K \times d_h}, \tag{2}$$

where d_h is the dimension of the hidden representation space.

During training, each sample is assigned to its nearest prototype based on Euclidean distance:

$$k^i = \arg\min_{k \in [1,K]} \|\phi_i - \mu_{t,k}\|_2^2, \tag{3}$$

where k^i is the assigned prototype index for sample i, ϕ_i is the feature representation of sample i, and $\mu_{t,k}$ is the k-th prototype of group t.

We define the clustering loss as:

$$\mathcal{L}_{\text{cluster}} = \sum_{t \in \{0,1\}} \sum_{k=1}^{K} \sum_{i \in \mathcal{S}_{t,k}} \|\phi_i - \mu_{t,k}\|_2^2, \tag{4}$$

where ϕ_i denotes the representation of instance i, and $\mu_{t,k}$ is the prototype for cluster k in group t. Each sample i is assigned to a prototype k^i, and the assignment set is:

$$\mathcal{S}_{t,k} = \{i \mid k_i^* = k,\ t_i = t\}. \tag{5}$$

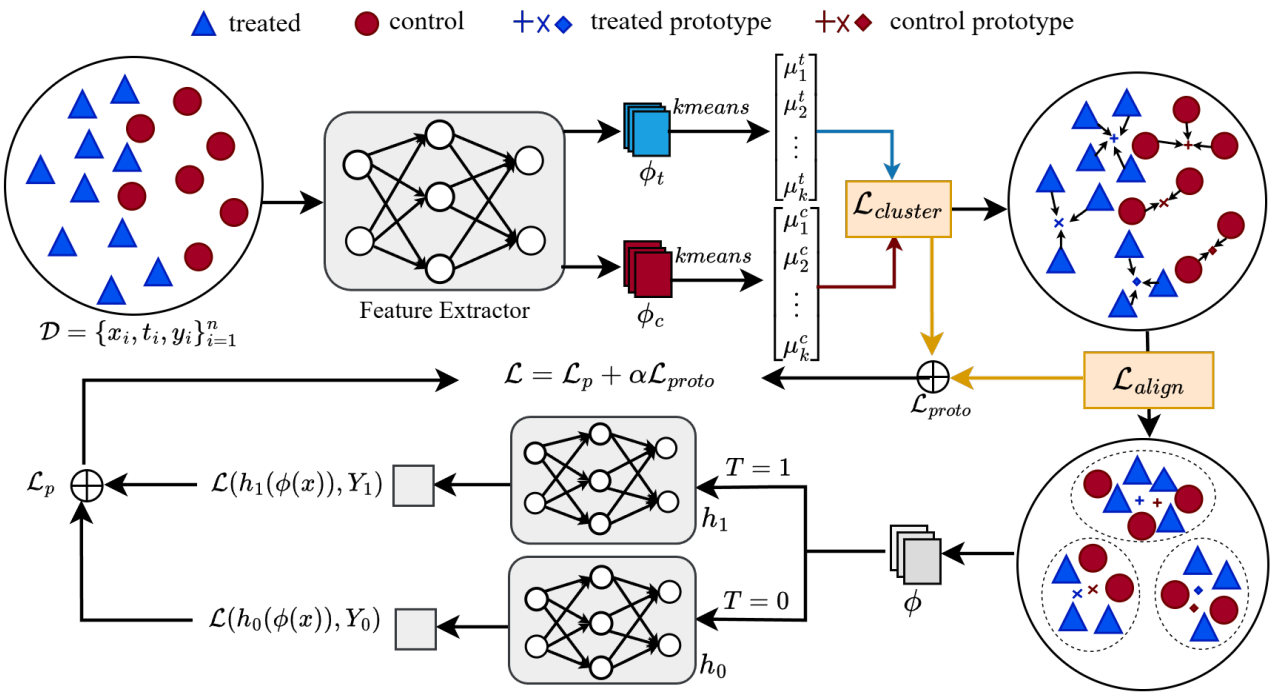

Figure 2: An overview of the PITE framework. We perform prototype learning on the representations ϕ_t and ϕ_c for treated and control groups respectively via k-means to capture local structures within each group. Cross-treatment prototype alignment ($\mathcal{L}_{align}$) enforces correspondence between treated and control prototypes to reduce distribution shift. Finally, two separate neural networks, $h_1(\phi(x))$ and $h_0(\phi(x))$, are used to estimate potential outcomes under different treatments.

The gradient with respect to the prototype $\mu_{t,k}$ is given by:

$$\frac{\partial \mathcal{L}_{\text{cluster}}}{\partial \mu_{t,k}} = \frac{2}{|\mathcal{S}_{t,k}|} \sum_{i \in \mathcal{S}_{t,k}} (\mu_{t,k} - \phi_i). \tag{6}$$

This objective encourages instance representations to stay close to their corresponding prototypes, thereby preserving the local structure in the representation space.

PITE defines prototypes as learnable cluster centroids of hidden representations within each group, serving as stable and representative anchors for each subgroup. During training, each sample is assigned to the nearest prototype based on Euclidean distance, which ensures clear clustering boundaries and helps identify the natural clustering structures among individuals. The gradient update mechanism ensures that prototypes converge toward the centroids of their assigned samples, obtaining stable and representative cluster centers. Thus, within-group prototype matching provides a stable structural constraint, which are subsequently utilized in cross-group prototype alignment to achieve robust counterfactual estimation.

Cross-Group Prototype Alignment

To address the distribution shift between treated and control groups, PITE performs a pairwise prototype alignment strategy through meaningful cross-group prototype matching in the latent space. Unlike global alignment methods that average over the entire group, we align prototypes—each representing a distinct local cluster, based on the motivation that they capture the local structure in the representation space that represents the natural clustering among individuals. By enforcing proximity between matched prototypes across groups, this loss effectively reduces the distribution mismatch, encourages cross-group correspondence at the prototype-level alignment, and facilitates more reliable counterfactual estimation at a finer granularity. Formally, the alignment objective is defined as:

$$\mathcal{L}_{\text{align}} = \frac{1}{K} \sum_{k=1}^{K} \|\mu_{1,k} - \mu_{0,k}\|_2^2, \tag{7}$$

where $\mu_{1,k}$ and $\mu_{0,k}$ are the k-th prototypes of the treated and control groups, respectively.

However, aggressive alignment may cause prototype collapse. To preserve diversity, PITE introduces a diversity regularization term:

$$\mathcal{L}_{\text{div}} = -\frac{1}{K(K-1)} \sum_{t \in \{0,1\}} \sum_{i \neq j} \|\mu_{t,i} - \mu_{t,j}\|_2^2, \tag{8}$$

where $t \in \{0, 1\}$ denotes the treatment and control groups, $\mu_{t,i}$ represents the i-th prototype in group t, K is the number of prototypes each group. This regularization encourages each prototype to capture distinct feature patterns within each group. This design balances two key objectives: cross-group prototype alignment, which is essential for accurate

individual treatment effect estimation, and intra-group diversity preservation, which prevents information redundancy. By maintaining rich and heterogeneous representations, it ultimately enhances both the accuracy and robustness of causal effect estimation.

The overall prototype loss combines clustering, alignment, and diversity objectives:

$$\mathcal{L}_{\text{proto}} = \mathcal{L}_{\text{cluster}} + \beta\mathcal{L}_{\text{align}} + \gamma\mathcal{L}_{\text{div}}, \tag{9}$$

where β and γ are hyperparameters that weight the alignment and diversity terms relative to the clustering objective.

Prediction Head

The learned balanced representations $\phi(x_i)$ are fed into two neural networks to predict potential outcomes for treatment ($t = 1$) and control ($t = 0$) (Assaad et al. 2021; Huang et al. 2023). The predicted outcomes are defined as $T_{\text{out}} = h(\phi(x_i), t_i = 1)$ and $C_{\text{out}} = h(\phi(x_i), t_i = 0)$, respectively. The predictive loss is given by:

$$\mathcal{L}_p = \frac{1}{n}\sum_{i=1}^{n} w_i \cdot \mathcal{L}\left(h\left(\phi\left(x_i\right), t_i\right), y_i\right), \tag{10}$$

where $w_i = \frac{t_i}{2u} + \frac{1-t_i}{2(1-u)}$, and $u = \frac{1}{n}\sum_{i=1}^{n} t_i$.

The end-to-end prototype-level alignment method reduces distribution discrepancy across groups while preserving the intrinsic clustering structure of the data by ensuring within-group cohesion and cross-group alignment.The total loss $\mathcal{L}_t$ combines predictive loss, prototype loss, and regularization:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{p}} + \alpha\mathcal{L}_{\text{proto}} + \lambda\|W\|_2, \tag{11}$$

where α and β are adjustable hyper-parameters that control the contributions of prototype loss and regularization loss $\|\cdot\|_2$ on model weights W to prevent overfitting.

We train our model by minimizing Equation (11) and provide the detailed multi-prototype alignment strategy for ITE estimation in Algorithm 1 in the Appendix. This formulation ensures that the learned representations achieve cross-group prototype alignment while accurately predicting potential outcomes, ultimately reducing ITE estimation error.

Experiments

Datasets

Synthetic. We generate covariates from the multivariate normal distribution $\mathcal{N}\left(\mathbf{0}, \gamma \cdot \sigma^2 \cdot \left[\rho \mathbf{1}_p \mathbf{1}_p^\top + (1-\rho)\mathbf{I}_p\right]\right)$, where the covariance matrix combines an all-ones matrix $\mathbf{1}_p\mathbf{1}_p^\top$ and an identity matrix $\mathbf{I}_p$. The scaling parameter $\gamma \in \{0.4, 0.7, 1.0, 1.2\}$ controls the degree of covariate dispersion. We sample 800 units with parameters $p = 10$, $\rho = 0.2$, $\sigma^2 = 3$, $\beta_0 = [0.2, 0.2, ..., 0.2]$, and $\beta_1 = [1.2, 1.2, ..., 1.2]$. For each γ, we generate 30 independent datasets, dividing them into training, validation, and test sets with ratios of 63%, 27%, and 10%, respectively. The data generation process is outlined as follows:

$$\begin{aligned}
&\mathbf{X}_i \sim \mathcal{N}\left(\mathbf{0}, \gamma \cdot \sigma^2 \cdot \left[\rho \mathbf{1}_p \mathbf{1}_p^\top + (1-\rho)\mathbf{I}_p\right]\right), \\
&T_i \mid \mathbf{X}_i \sim \text{Bernoulli}\left(1/(1+\exp\left(-\mathbf{1}_p^\top \mathbf{X}_i\right))\right), \\
&Y_i^0 = \boldsymbol{\beta_0}\mathbf{X}_i + \xi_i, \quad Y_i^1 = \boldsymbol{\beta_1}\mathbf{X}_i + \xi_i, \quad \xi_i \sim \mathcal{N}(0,1).
\end{aligned}$$

Semi-synthetic (IHDP). The IHDP dataset, introduced by Hill (Hill 2011) based on the Infant Health and Development Program, is a randomized control trail to assess whether there is influence of specialist visit (treatment) on children's cognitive scores (outcome). Hill excluded a subpopulation with non-white mothers from the treatment group to cause selection bias. The IHDP dataset consists of 747 samples, comprising 139 treated samples and 608 controlled samples. We use the same 100 datasets, following the standard practice in the field.

Real-world (Jobs). The Jobs dataset, combined Lalonde and a randomized study, investigated the causal effect of job training (treatment) on income and employment status after training (Dehejia and Wahba 2002). This research constructed a binary classification task, where the goal is to predict unemployment using the feature sets. We use the same 10 datasets as used in (Shalit, Johansson, and Sontag 2017), comprising 297 treated samples and 2915 controlled samples with train /validation/test splits with ratios 56/24/20.

Metrics

On IHDP dataset where the true treatment effect for individual is known, we adopt two commonly evaluation metrics, namely the *Precision in Estimation of Heterogeneous Effect* (ϵ_{PEHE}) and the *absolute error of Average Treatment Effect* (ϵ_{ATE}) defined as:

$$\epsilon_{PEHE} = \frac{1}{n}\sum_{i=1}^{n}\left(\tau(x_i) - \hat{\tau}(x_i)\right)^2, \tag{12}$$

$$\epsilon_{ATE} = \left|\hat{ATE} - ATE\right| = \frac{1}{n}\left|\sum_{i=1}^{n}\left(\tau_i - \hat{\tau}_i\right)\right|, \tag{13}$$

where τ_i refer to the ground truth treatment effect, $\hat{\tau}_i$ is the estimated treatment effect.

On Jobs dataset, we adopt the *policy risk* $\mathcal{R}_{pol}(\pi_{\hat{\tau}})$ and the bias of *Average Treatment Effect on the Treated* prediction ϵ_{ATT}.

$$\begin{aligned}
\mathcal{R}_{\text{pol}}(\pi_{\hat{\tau}}) = 1 - \Big[&\Pr(\pi_{\hat{\tau}}(x) = 1) \cdot \mathbb{E}\left[Y_1 \mid \pi_{\hat{\tau}}(x) = 1\right] \\
&+ \Pr(\pi_{\hat{\tau}}(x) = 0) \cdot \mathbb{E}\left[Y_0 \mid \pi_{\hat{\tau}}(x) = 0\right] \Big],
\end{aligned} \tag{14}$$

where $\pi_{\hat{\tau}} : \mathcal{X} \to \{0, 1\}$ is an policy induced from an ITE estimator $\hat{\tau}(\cdot)$ with $\pi_{\hat{\tau}}(x) = 1$ if $\hat{\tau}(x) > 0$ and $\hat{\tau}(x) = 0$ otherwise.

$$\epsilon ATT = \left| \left| \frac{\sum_{i=1}^{|\mathcal{T}_1|} y_i^1}{|\mathcal{T}_1|} - \frac{\sum_{i=1}^{|\mathcal{T}_0|} y_i^0}{|\mathcal{T}_0|} \right| - \left| \frac{\sum_{i=1}^{|\mathcal{T}_1|} \left(\hat{y_i^1} - \hat{y_i^0}\right)}{|\mathcal{T}_1|} \right| \right|, \tag{15}$$

where $|\mathcal{T}_1|$ and $|\mathcal{T}_0|$ are the number of the units in the treatment and the control groups, respectively.

Method	$\gamma = 0.4$		$\gamma = 0.7$		$\gamma = 1.0$		$\gamma = 1.2$	
	$\sqrt{\epsilon_{PEHE}^{within}}$	$\sqrt{\epsilon_{PEHE}^{out-of}}$	$\sqrt{\epsilon_{PEHE}^{within}}$	$\sqrt{\epsilon_{PEHE}^{out-of}}$	$\sqrt{\epsilon_{PEHE}^{within}}$	$\sqrt{\epsilon_{PEHE}^{out-of}}$	$\sqrt{\epsilon_{PEHE}^{within}}$	$\sqrt{\epsilon_{PEHE}^{out-of}}$
OLS-1	8.39(0.38)	8.41(0.84)	10.86(0.43)	10.85(1.34)	12.89(0.59)	13.00(1.68)	14.21(0.59)	14.32(1.55)
OLS-2	5.92(0.27)	5.94(0.60)	7.64(0.30)	7.64(0.96)	9.05(0.40)	9.13(1.18)	9.97(0.41)	10.07(1.11)
BART	4.04(0.22)	3.40(0.50)	4.70(0.20)	4.17(0.60)	5.36(0.31)	4.86(0.61)	5.81(0.26)	6.14(0.58)
KNN	4.55(0.33)	5.50(0.62)	6.48(0.37)	7.26(0.75)	7.91(0.38)	9.08(1.27)	8.76(0.46)	10.27(1.07)
CFR-Wass	2.48(0.05)	2.48(0.06)	3.73(0.05)	3.60(0.09)	4.68(0.07)	4.72(0.14)	5.34(0.07)	5.37(0.14)
CFR-MMD	2.54(0.05)	2.54(0.06)	3.75(0.05)	3.62(0.09)	4.70(0.07)	4.74(0.14)	5.37(0.08)	5.41(0.14)
GANITE	4.66(0.03)	4.69(0.06)	6.20(0.03)	6.16(0.07)	7.32(0.03)	7.33(0.07)	8.08(0.03)	8.11(0.08)
ABCEI	2.73(0.03)	2.75(0.06)	3.74(0.04)	3.57(0.09)	4.61(0.05)	4.73(0.13)	5.19(0.06)	5.19(0.12)
CBRE	2.91(0.03)	2.93(0.05)	4.01(0.04)	3.85(0.08)	4.95(0.05)	5.02(0.12)	5.77(0.06)	5.73(0.13)
DIGNet	3.17(0.09)	3.18(0.11)	4.09(0.10)	3.97(0.13)	5.05(0.10)	5.10(0.17)	5.78(0.09)	5.81(0.15)
SITE	2.68(0.11)	2.69(0.13)	4.17(0.16)	4.25(0.23)	5.98(0.34)	6.01(0.38)	6.19(0.15)	6.21(0.17)
CITE	2.69(0.04)	2.71(0.07)	3.81(0.06)	3.70(0.10)	4.68(0.06)	4.74(0.14)	5.39(0.07)	5.41(0.14)
FCCL	2.56(0.04)	2.58(0.06)	3.65(0.05)	3.50(0.09)	4.40(0.06)	4.49(0.13)	5.10(0.06)	5.12(0.12)
PITE	**2.22(0.09)**	**2.24(0.10)**	**3.42(0.08)**	**3.29(0.10)**	**4.30(0.08)**	**4.35(0.16)**	**4.99(0.10)**	**5.02(0.15)**

Table 1: Experimental results on Synthetic datasets. The best result in each row is highlighted in **bold**.

Method	$\sqrt{\epsilon_{PEHE}^{within}}$	ϵ_{ATE}^{within}	$\sqrt{\epsilon_{PEHE}^{out-of}}$	ϵ_{ATE}^{out-of}
OLS-1	5.83(0.39)	0.73(0.04)	5.91(0.27)	0.95(0.06)
OLS-2	2.42(0.16)	0.14(0.02)	2.55(0.16)	0.31(0.02)
BART	2.13(0.22)	0.24(0.05)	2.32(0.12)	0.35(0.03)
KNN	2.13(0.08)	0.15(0.05)	4.16(0.23)	0.80(0.05)
CFR-Wass	0.71(0.04)	0.27(0.03)	0.83(0.08)	0.28(0.03)
CFR-MMD	0.77(0.05)	0.25(0.04)	0.92(0.09)	0.28(0.04)
GANITE	1.92(0.29)	0.43(0.41)	2.43(0.46)	0.49(0.38)
ABCEI	0.79(0.06)	0.12(0.02)	1.00(0.13)	0.15(0.03)
CBRE	0.59(0.06)	0.11(0.02)	0.66(0.07)	0.13(0.02)
DIGNet	0.60(0.04)	0.15(0.02)	0.67(0.07)	0.16(0.02)
SITE	0.84(0.05)	0.30(0.04)	0.98(0.07)	0.32(0.05)
CITE	0.59(0.06)	0.11(0.02)	0.67(0.14)	0.14(0.02)
FCCL	0.53(0.04)	0.09(0.01)	0.64(0.07)	0.12(0.02)
PITE	**0.51(0.02)**	**0.09(0.01)**	**0.60(0.04)**	**0.11(0.02)**

Table 2: Within-sample and out-of-sample estimation errors for the metrics (**Lower is better**) on IHDP dataset.

Comparison with Baseline Approaches

We compare PITE empirically against the following 13 baselines. These approaches can be mainly divided into two categories: traditional methods and deep learning. We further categorize deep learning methods into distribution-level alignment methods and instance-level alignment methods.

Traditional Methods: Ordinary least square **(OLS-1)** using treatment as a covariate; **(OLS-2)**, predicting outcomes separately for each group; Bayesian additive regression trees **(BART)** leveraging a sum-of-trees structure; K-nearest neighbor **(KNN)** matching samples using k-nearest neighbors. **Distribution-level alignment: CFR-Wass** (Shalit, Johansson, and Sontag 2017) and **CFR-MMD** (Shalit, Johansson, and Sontag 2017) are two methods using the Wasserstein and MMD metric for counterfactual regression, respectively; **GANITE** (Yoon, Jordon, and Van Der Schaar 2018) implicitly learns counterfactual distribution using GANs; **ABCEI** (Du et al. 2021) balances distributions using adversarial learning; **CBRE** (Zhou et al. 2022) constructs an information loop during adversarial training to minimize information loss; **DIGNet** (Huang et al. 2024) utilizes individual propensity confusion and group distance minimization. **Instance-level alignment: SITE** (Yao et al. 2018), which preserves local similarity in sample representations; **CITE** (Li and Yao 2022) learns representation based on propensity score; **FCCL** (Zhang et al. 2025) integrates diffeomorphic counterfactual generation and contrastive learning to achieve sample-level alignment.

Experimental Results

In this section, we compare and analyse the overall performance of PITE, focusing on robustness under different covariate dispersion conditions. Moreover, we conduct uniformity analysis and sensitivity analysis to validate the efficiency of PITE. Further results, including sensitivity analysis, are presented in the Appendix.

Performance Evaluation. We evaluate PITE against baseline methods on the Synthetic, IHDP and Jobs datasets, with the main results shown in Table 1 and Table 2, and additional results provided in the Appendix.

Synthetic Data. Table 1 presents the evaluation results of our PITE compared to baseline methods on the synthetic dataset under varying degrees of covariate dispersion ($\gamma = 0.4, 0.7, 1.0, 1.2$). When the covariance parameter γ increases from 0.4 to 1.2, PEHE estimation errors universally increase across all methods, indicating that higher data dispersion poses greater challenges for causal effect estimation. Key findings include: (1) PITE consistently achieves the lowest estimation errors across all γ settings, significantly outperforming existing methods. (2) Traditional approaches like KNN show dramatic performance degradation as γ increases ($\sqrt{\epsilon_{PEHE}^{within}}$ from 4.55 to 8.76), while deep learning methods represented by CFR-MMD also demonstrate poor performance ($\sqrt{\epsilon_{PEHE}^{within}}$ from 2.54 to

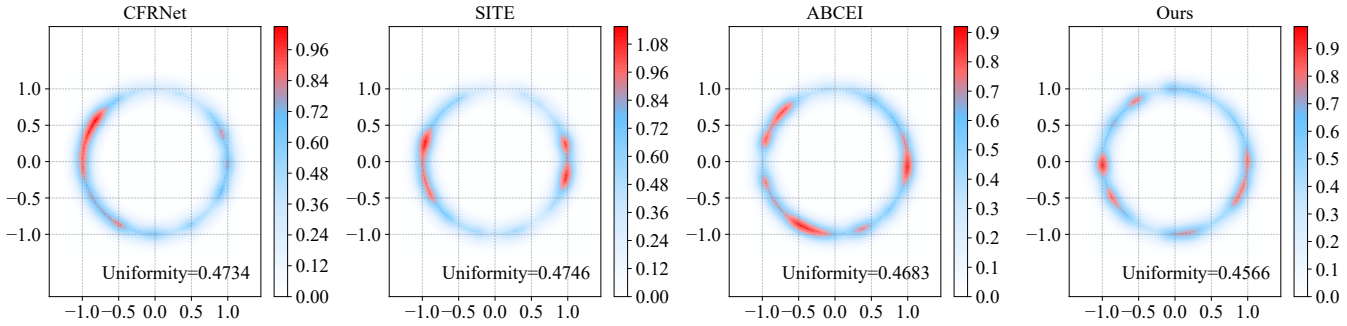

Figure 3: Visualization of representation uniformity of four typical methods on IHDP dataset. We visualize the overall feature distributions with Gaussian kernel density estimation (KDE) in $\mathbb{R}^2$, where the color gradient represents density levels from low (blue) to high (red). The uniformity metric is computed by measuring the pairwise distances between normalized representations on the hypersphere, with lower values indicating superior uniformity.

5.37). In contrast, PITE maintains remarkable stability with $\sqrt{\epsilon_{PEHE}^{within}}$ increasing only from 2.22 to 4.99. This superior robustness stems from the prototype learning mechanism, which represents natural grouping structures of similar individuals through cluster centroids, thereby avoiding outlier interference inherent in direct global matching and instance-to-instance matching. As data complexity increases, prototypes serve as stable representative points that are more resilient to noise and anomalous observations compared to samples, enabling more accurate and robust ITE estimation.

Benchmark Data. It is worth noting that, PITE significantly outperforms distribution-level alignment and instance-level methods. Compared to *distribution-level* alignment methods such as CFRNet, PITE demonstrates superior performance, achieving substantial reductions in out-of-sample $\sqrt{\epsilon_{PEHE}^{out-of}}$ and ϵ_{ATE}^{out-of} by 34.8% and 60.7%, respectively. By performing multi-prototype alignment across groups and preserving the inherent structure during representation learning, PITE effectively mitigates distribution shift and enables more accurate estimation of counterfactual outcomes. CFRNet, GANITE, ABCEI, CBRE and DIGNet show limited performance because these methods generally use the first moment as the distribution discrepancy metric, ignoring the underlying structural constraint that represents the natural clustering among individuals. Compared to *instance-level* alignment methods, PITE outperforms these methods, achieving a 10.4% reduction in $\sqrt{\epsilon_{PEHE}^{out-of}}$ compared to CITE. PITE performs instance-to-prototype matching to preserve the local structure in a more robust manner. However, CITE depends heavily on the correct specification of the propensity score, which is usually difficult to obtain. Besides, SITE only achieves partial balance through selecting specific sample pairs for alignment, and therefore shows inferior performance ($\sqrt{\epsilon_{PEHE}^{within}} = 0.84$ and $\sqrt{\epsilon_{PEHE}^{out-of}} = 0.98$). Although FCCL demonstrates competitive performance compared with distribution-level methods, it suffers from high computational overhead and similarly overlooks local structure preservation, which ultimately compromises ITE estimation. Besides, we evaluate the contribution of the alignment loss and diversity regularization term in the prototype-level alignment in the Appendix.

Uniformity Analysis. Figure 3 evaluates the uniformity of four typical methods in the representation space. We observe that PITE obtains the lowest uniformity metric $uniformity = 0.4566$, which shows that PITE can make feature vectors roughly uniformly distributed on the unit hypersphere and preserve as much sample information as possible. PITE assigns instances to semantically meaningful cluster centroids via within-group prototype matching, promoting structured coverage of the representation space within each treatment group. Simultaneously, PITE establishes correspondence between treated and control prototypes through cross-group alignment, preventing isolated clusters and ensuring balanced distribution across treatment arms. By operating on stable cluster representatives rather than instances, PITE provides more robust alignment that effectively prevents representation collapse and achieves more uniform feature space utilization.

Conclusion

In this paper, we address the critical issue of neglecting local structure information that represents the natural clustering among individuals, which exists in both distribution-level and instance-level alignment methods for individual treatment effect estimation. To achieve this, we propose PITE, a novel prototype-level method for robust ITE estimation. PITE innovatively introduces prototypes and designs intra-group instance-to-prototype matching along with cross-group multi-prototype alignment strategies, effectively mitigating distribution shift while preserving the local structure of data, which provides meaningful constraints for ITE estimation. Compared to other baselines, comprehensive experiments across various datasets demonstrate that PITE achieves more accurate and robust ITE estimation. In future work, we will explore causal effect estimation in multimodal data settings, incorporating semantic information across different modalities to enhance ITE estimation performance.

Acknowledgements

This work is supported by the National Natural Science Foundation of China (U24A20323 and 62376145), the Key Technologies Program of Taihang Laboratory in Shanxi Province (THYF-JSZX-24010700), the Science and Technology Innovation Talent Team of Shanxi Province (202204051002016), and the Taiyuan City 'Double hundred Research action' of the first batch project about 'Leading the Charge with Open Competition' (2024TYJB0127).

References

An, W.; Tian, F.; Shi, W.; Chen, Y.; Wu, Y.; Wang, Q.; and Chen, P. 2024. Transfer and alignment network for generalized category discovery. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 10856–10864.

Assaad, S.; Zeng, S.; Tao, C.; Datta, S.; Mehta, N.; Henao, R.; Li, F.; and Carin, L. 2021. Counterfactual representation learning with balancing weights. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, 1972–1980.

Bica, I.; and Van der Schaar, M. 2022. Transfer learning on heterogeneous feature spaces for treatment effects estimation. In *Proceedings of the Advances in Neural Information Processing Systems*, 37184–37198.

Cheng, G.; Hardt, M.; and Mendler-Dünner, C. 2024. Causal inference out of control: Estimating performativity without treatment randomization. In *Proceedings of the International Conference on Machine Learning*, 8077–8103.

Chu, Z.; Ding, H.; Zeng, G.; Huang, Y.; Yan, T.; Kang, Y.; and Li, S. 2022. Hierarchical capsule prediction network for marketing campaigns effect. In *Proceedings of the ACM International Conference on Information & Knowledge Management*, 3043–3051.

Dehejia, R. H.; and Wahba, S. 2002. Propensity score-matching methods for nonexperimental causal studies. *Review of Economics and Statistics*, 84(1): 151–161.

Du, X.; Sun, L.; Duivesteijn, W.; Nikolaev, A.; and Pechenizkiy, M. 2021. Adversarial balancing-based representation learning for causal effect inference with observational data. *Data Mining and Knowledge Discovery*, 35(4): 1713–1738.

Feuerriegel, S.; Frauen, D.; Melnychuk, V.; Schweisthal, J.; Hess, K.; Curth, A.; Bauer, S.; Kilbertus, N.; Kohane, I. S.; and Van der Schaar, M. 2024. Causal machine learning for predicting treatment outcomes. *Nature Medicine*, 30(4): 958–968.

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

Huang, Y.; Leung, C. H.; Ma, S.; Yuan, Z.; Wu, Q.; Wang, S.; Wang, D.; and Huang, Z. 2023. Towards balanced representation learning for credit policy evaluation. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, 3677–3692.

Huang, Y.; Siyi, W.; Leung, C. H.; Qi, W.; Dongdong, W.; and Huang, Z. 2024. DIGNet: Learning decomposed patterns in representation balancing for treatment effect estimation. *Transactions on Machine Learning Research*.

Kong, I.; Park, Y.; Jung, J.; Lee, K.; and Kim, Y. 2023. Covariate balancing using the integral probability metric for causal inference. In *Proceedings of the International Conference on Machine Learning*, 17430–17461.

Li, Q.; Wang, Z.; Liu, S.; Li, G.; and Xu, G. 2021. Causal optimal transport for treatment effect estimation. *IEEE Transactions on Neural Networks and Learning Systems*, 34(8): 4083–4095.

Li, X.; and Yao, L. 2022. Contrastive individual treatment effects estimation. In *Proceedings of the International Conference on Data Mining*, 1053–1058.

Li, Y.; Leung, C. H.; Sun, X.; Wang, C.; Huang, Y.; Yan, X.; Wu, Q.; Wang, D.; and Huang, Z. 2024. The causal impact of credit lines on spending distributions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 180–187.

Liu, R.; Wei, L.; and Zhang, P. 2021. A deep learning framework for drug repurposing via emulating clinical trials on real-world patient data. *Nature Machine Intelligence*, 3(1): 68–75.

Rosenbaum, P. R.; and Rubin, D. B. 1983. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1): 41–55.

Rubin, D. B. 2005. Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, 100(469): 322–331.

Shalit, U.; Johansson, F. D.; and Sontag, D. 2017. Estimating individual treatment effect: generalization bounds and algorithms. In *Proceedings of the International Conference on Machine Learning*, 3076–3085. PMLR.

Wang, H.; Fan, J.; Chen, Z.; Li, H.; Liu, W.; Liu, T.; Dai, Q.; Wang, Y.; Dong, Z.; and Tang, R. 2023. Optimal transport for treatment effect estimation. *Proceedings of the Advances in Neural Information Processing Systems*, 36: 5404–5418.

Yan, Y.; Zhou, H.; Yang, Z.; Chen, W.; Cai, R.; and Hao, Z. 2024. Reducing balancing error for causal inference via optimal transport. In *Proceedings of the International Conference on Machine Learning*.

Yao, L.; Li, S.; Li, Y.; Huai, M.; Gao, J.; and Zhang, A. 2018. Representation learning for treatment effect estimation from observational data. In *Proceedings of the Advances in Neural Information Processing Systems*, 2638–2648.

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

Yue, X.; Zheng, Z.; Zhang, S.; Gao, Y.; Darrell, T.; Keutzer, K.; and Vincentelli, A. S. 2021. Prototypical cross-domain self-supervised learning for few-shot unsupervised domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13834–13844.

Zhang, J.; Eldele, E.; Wang, Y.; Li, X.; Liang, J.; et al. 2025. Counterfactual contrastive learning with normalizing flows

for robust treatment effect estimation. In *Proceedings of the International Conference on Machine Learning*.

Zhou, G.; Yao, L.; Xu, X.; Wang, C.; and Zhu, L. 2022. Cycle-balanced representation learning for counterfactual inference. In *Proceedings of the 2022 SIAM International Conference on Data Mining*, 442–450.