VISIBLE AND INVISIBLE: CAUSAL VARIABLE LEARNING AND ITS APPLICATION IN A CANCER STUDY

Anonymous authors
Paper under double-blind review

Abstract

Causal visual discovery is a fundamental yet challenging problem in many research fields. Given visual data and the outcome of interest, the goal is to infer the cause-effect relation. Aside from rich visual ('visible') variables, oftentimes, the outcome is also determined by 'invisible' variables, *i.e.* the variables from non-visual modalities that do not have visual counterparts. This (**visible**, **invisible**) combination is particularly common in the clinical domain.

Built upon the promising invariant causal prediction (ICP) framework, we propose a novel ε-ICP algorithm to resolve the (visible, invisible) setting. To efficiently discover ε-plausible causal variables and to estimate the cause-effect relation, the ε-ICP is learned under a min-min optimisation scheme. Driven by the need for clinical reliability and interpretability, the ε-ICP is implemented with a typed neural-symbolic functional language. With the built-in program synthesis method, we can synthesize a type-safe program that is comprehensible to the clinical experts.

For concept validation of the ε-ICP, we carefully design a series of synthetic experiments on the type of visual-perception tasks that are encountered in daily life. To further substantiate the proposed method, we demonstrate the application of ε-ICP on a real-world cancer study dataset, Swiss CRC. This population-based cancer study has spanned over two decades, including $25k$ fully annotated tissue micro-array (TMA) images with at least $3k \times 3k$ resolution and a broad spectrum of clinical meta data for 533 patients. Both the synthetic and clinical experiments demonstrate the advantages of ε-ICP over the state-of-the-art methods. Finally, we discuss the limitations and challenges to be addressed in the future.

1 Introduction

Causal discovery is of key importance to determine cause-effect relations in different scientific disciplines (Spirtes et al. (2000); Pearl et al. (2009); Pearl & Mackenzie (2018); Schölkopf (2019)). Lake et al. (2017) pointed out that the next generation of artificial intelligence (AI) should be endowed with causal reasoning to support explainability and human understanding. In the clinical domain, randomized controlled trials have been widely applied to investigate the causal dependence between treatment and recovery (Peters et al. (2017)).

1.1 Causal Discovery

Given either purely observational data or by inclusion of additional experimental (interventional) data, we aim to infer the underlying causal structure. In causal discovery, we refer to this question as **structure identifiability**. (Spirtes et al. (2000); Pearl et al. (2009); Peters et al. (2017)) presented a comprehensive review on this topic. The goal of structure identifiability is often to learn the entire causal structure (Verma & Pearl (1991); Spirtes et al. (2000); Tian & Pearl (2013); Hauser & Bühlmann (2014)). In contrast, invariant causal prediction (ICP) (Peters et al. (2016)) aims to learn a set of identifiable causal variables given an outcome of interest. Thus, this goal is easier to achieve. The idea of ICP originates from the autonomy assumption (Aldrich (1989); Hendry & Morgan (1997)) [1] of causality, *if we intervene variables other than the outcome variable in a causal model, the conditional probability of the outcome given its causal variables should remain identical.*

[1] This assumption is also known as modularity or stability (Pearl et al. (2009); Schölkopf et al. (2012); Dawid et al. (2010))

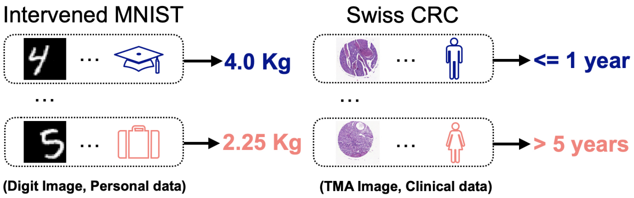
Recently, several ICP-based algorithms have been proposed. Peters et al. (2016) first presented vanilla ICP and verified its efficiency on linear cause-effect relations. Next, Pfister et al. (2017) applied several choices of test statistics for sequential data. Rojas-Carulla et al. (2018) proposed (greedy) subset search algorithms (GSS) for domain adaptation and multi-task learning. Heinze-Deml et al. (2018) utilized multiple conditional independence tests (CIT) and extended the vanilla ICP to nonlinear settings. To quickly discover the direct causes of an outcome variable, Gamella & Heinze-Deml (2020) proposed several intervention selection policies for the active ICP method. On one hand, the existing ICPs demonstrated themselves to be conceptually simple and theoretically sound. Numerical experiments validated their advantages on identifying causal variables. On the other hand, how to accurately predict the outcome remains an open question. Besides, the experiments were mostly conducted on numerical datasets. The strength of the ICP has not been fully explored on real-world vision datasets at scale.

1.2 Causal Visual Discovery

Integrating causal discovery in vision-related tasks has emerged as an important direction of computer vision research. Topics include fundamental vision tasks, *e.g.* classification (Lopez-Paz et al. (2017); Chang et al. (2018); Goyal et al. (2019a)), tracking (Lebeda et al. (2015); Xu et al. (2018)), segmentation (Yang et al. (2015); Taylor et al. (2015); Bideau & Learned-Miller (2016)), and their downstream problems, *i.e.* visual question answering (Chen et al. (2020); Abbasnejad et al. (2020); Niu et al. (2020)), scene graph generation (Chen et al. (2019); Tang et al. (2020)), visual physics reasoning (Zhang et al. (2016); Baradel et al. (2019); Yi et al. (2019)), visual explanations (Hendricks et al. (2018); Goyal et al. (2019b); Kanehira et al. (2019); Fang et al. (2019); Wang & Vasconcelos (2020)) and visual navigation (Nguyen et al. (2019); Fu et al. (2019)). At the same time, a series of studies have shed light on how causal inference solves medical imaging tasks, including cancer classification (Major et al. (2020)), brain imaging analysis (Friston et al. (2003); Marreiros et al. (2008); Liao et al. (2009); Weichwald et al. (2015); Kassani et al. (2020)), medical image synthesis (Xu et al. (2020)), and medical report synthesis (Han et al. (2020)). To solve the listed tasks, the combination of textual and visual information has become a popular trend. In many of these instances, it is common that a visual object extracted from an image is sufficiently explained by its textual counterpart. In other cases, the combination of the visual ('visible') variables and 'invisible' variables, *i.e.* the variables from non-visual modalities that do not have visual counterparts determine the outcome of interest. This investigation of (**visible**, **invisible**) variables in relation to the outcome bears great importance in scientific research, especially in the clinical domain.

2 Application Examples

To instantiate the concept of (in)visible causal variables, we elaborate two concrete applications (See Tab. 1 (left)). For proof of concept, we first derive a synthetic dataset to illustrate the application on a canonical visual-perception task, inspired by the first chapter of Peters et al. (2017). Then, we apply the proposed method for the identification of prognostic (causal) variables in a real-world clinical cohort of colorectal cancer patients (Swiss CRC). The Swiss CRC cohort is one of the world's largest colorectal cancer collections with availability of full clinicopathological information and digital histopathology images containing rich visual information with biological and clinical relevance for each patient (see Tab. 1 (right) for the dataset comparison). In the experimental section, we demonstrate the advantages of the proposed ε-ICP on both examples.



Reference	Patients	TMA images
Knösel et al. (2005)	270	351
Hashimoto et al. (2006)	131	374
Kume et al. (2014)	-	1150
Bychkov et al. (2018)	420	420
Uttam et al. (2020)	432	694
Swiss CRC (present study)	**533**	**~25000**

Table 1: Left: Conceptual illustrations of two application examples: Intervened MNIST and Swiss CRC. Right: Comparison to existing tissue micro-array (TMA) colorectal datasets.

Intervened MNIST: (, Personal Data) → Consumption Quantity.

One of the canonical visual-perception tasks is *optical character recognition*, meaning the recognition of the number on an image for a variety of downstream tasks. In many scenarios, how different individuals interpret the number matters. Recently, (Yadav & Bottou (2019)) re-discovered an interesting fact of the seminal MNIST dataset (LeCun (1998)): The digits were written by two groups of people, Census Bureau employees and high school students in the early 1990s. Here, we assume that occupations impact the way that people interpret numbers (visual information) in the MNIST dataset, and design several synthetic yet plausible scenarios to explore how this subtle difference in interpretation could impact consuming behavior.

Problem. *Assume an individual is asked to purchase a certain quantity of food and the quantity is determined by the digit image she/he perceives. While a high school student applies SI units,* i.e. *kilogram (kg), a Census Bureau staff uses pound (lb). Then we have the following outcomes:*

1. $Y_0 = \begin{cases} 0.45kg \times \mathsf{Digit} & \textit{if Census Bureau staff} \\ 1kg \times \mathsf{Digit} & \textit{if high school student} \end{cases}$

2. *We further assume that* $f(\mathsf{Age}, \mathsf{Gender})$ *additionally impacts the meat consumption pattern*

   $Y_1 = Y_0 + f(\mathsf{Age}, \mathsf{Gender})$

3. *For the purpose of exploring visual causal variables, we assume image shifts* $\mathsf{abs}(\mathsf{Shift_x}), \mathsf{abs}(\mathsf{Shift_y})$ *implicitly impact the buying pattern as well, that is*

   $Y_2 = Y_0 + f(\mathsf{Age}, \mathsf{Gender}) * \mathsf{abs}(\mathsf{Shift_x}) * \mathsf{abs}(\mathsf{Shift_y})/28$

*As a sanity check, we plug the two non-causal variables rotation (*Rot*) and income (*Income*) in* $Y_{0,1,2}$*. In summary, we create a dataset pairing a digit image* *with* Personal data*, where* *encodes visible variables* (Digit, Shift$_x$, Shift$_y$, Rot) *and* Personal data *includes invisible variables* (Occp, Gender, Age, Income)*. Concretely, the paired data is sampled from either an observational or an interventional distribution. Eventually, our goals are to discover the (in)visible causal variables and to predict the* $Y_{0,1,2}$*. For data statistics and function* f *please check Appendix A.*

Swiss CRC: (, Clinical Data) → Disease Free Survival.

Colorectal cancer (CRC) is the third most common malignancy and fourth leading cause of cancer deaths worldwide (`https://gco.iarc.fr/` WHO database). Precise diagnosis and prognostic stratification for CRC patients is critical for personalised treatment and optimal survival (Sirinukunwattana et al. (2020)). To identify prognostic (causal) variables and their relevance for the survival outcomes, we propose a novel large-scale clinicopathological dataset from the population-based Swiss CRC study that has spanned over two decades, including 533 patients with complete clinicopathological data and survival outcome. Importantly, the Swiss CRC dataset contains $25k$ high quality digital tissue micro-array (TMA) images with at least $3k \times 3k$ resolution (Kononen et al. (1998)). The TMA images stained by hematoxylin and eosin (H&E) provide access to inexpensive and essential morphological features that are informative for cancer diagnosis and prognostic stratification (Srinidhi et al. (2019)). All the TMA images were carefully annotated and reviewed by two experienced gastrointestinal pathologists. Here, we address the following clinical problem:

Problem. *Given a patient suffering from CRC, depending on the variables extracted from TMA images of the surgical tumor resection specimen and from clinical data, the doctors need to determine the treatment plan. After surgical resection, each patient undergoes examinations at regular intervals and the duration of disease free survival (*DFS*) is recorded. This process can be formalized as follows:*

$$\begin{aligned} Y_{\mathsf{DFS}} = g(&\mathsf{Gender}, \mathsf{Age}, \mathsf{Weight}, \mathsf{Height}, \mathsf{preOP}, \mathsf{postOP}, \mathsf{pT}, \mathsf{pN}, \mathsf{pM}; \\ &\mathsf{Grade}, \mathsf{Immune}), \end{aligned} \tag{1}$$

where Grade, Immune *are visible variables encoded in* *, and the rest are invisible clinical variables. Our goals are to discover strongly informative (causal) variables and classify the* Y_{DFS} *of each patient, where* Y_{DFS} *is the ordinal label converted from monthly* DFS*. As a sanity check, we plug a non-causal variable* Normal *(the binary label of Tumor vs Normal TMA images) into* $g(\ldots)$*. For data statistics and clinical variable definition please check Appendix A.*

Contributions

Motivated by the above examples, we propose a novel ε-ICP algorithm to resolve the causal discovery problem under the (visible, invisible) setting. Our contributions can be summarized as follows

- To efficiently discover ε-plausible causal variables and to estimate the cause-effect relation, the proposed ε-ICP is learned under a min-min optimisation scheme. Concretely, we learn the gradients to reveal the ε-plausible causal variables under multiple experimental environments, and estimate the cause-effect under one ensemble environment.
- Driven by the clinical reliability and interpretability, the ε-ICP is implemented with a typed neural-symbolic functional language HOUDINI (Valkov et al. (2018)). With the built-in program synthesis method, we can synthesize a type-safe program that is comprehensible to clinical experts.
- To thoroughly examine the ε-ICP, not only do we design a series of synthetic experiments that reflect the type of visual-perception tasks we encounter in our daily life, but we investigate a unique large-scale real-world dataset collected from the Swiss colorectal cancer study, with full clinicopathological data, *i.e.* clinical data and tissue micro-array (TMA) images for each patient. A dataset of this scale is not yet publicly available.

3 Invariant Causal Prediction (ICP)

Following the terminologies introduced in ICP (Peters et al. (2016)), we first assume the set of experimental environments U from which the data are collected. The experimental environments arise via one or several interventions, including but not limited to *do intervention* (Pearl et al. (2009)), *noise intervention* and *simultaneous noise intervention*. Now we introduce the key assumption of ICP

Assumption 1. *(**Plausible Causal Variables**) Given a random vector $\boldsymbol{X}^u = (x^u_1, x^u_2, \dots, x^u_n)$, an outcome Y^u for all $u \in U$, there exists a $\boldsymbol{X}^u_{S^*} = (x^u_{S^*_1}, \dots, x^u_{S^*_j})$ with indices $S^* \subseteq \{1, \dots, n\}$ such that*

$$Y^u = f^*(\boldsymbol{X}^u_{S^*}) + \delta^u, \text{ where } f^* : \mathbb{R}^{|S^*|} \mapsto \mathbb{R}. \tag{2}$$

$$\delta^u \text{ are} \begin{cases} \textit{identically distributed} & \textit{if } \exists \textit{ hidden confounders} \\ \textit{identically distributed and } \delta^u \perp\!\!\!\perp \boldsymbol{X}^u_{S^*} & \textit{else} \end{cases} \tag{3}$$

Here, $\boldsymbol{X}^u_{S^*}$ are the plausible causal variables. Then, we have the following core theorem of ICP

Theorem 1. *We define **Identifiable Causal Variables** $S(U)$ as follows*

$$S(U) := \bigcap \{S \subseteq \{1, \dots, n\} \mid H_{0,S}(U) \textit{ is true.}\} \tag{4}$$

$$H_{0,S}(U): \quad \exists f : \mathbb{R}^{|S|} \mapsto \mathbb{R} \textit{ s.t. } Y^u = f(\boldsymbol{X}^u_S) + \delta^u, \ \forall u \in U, \tag{5}$$

where δ^u satisfy (3). Then we have $S(U) \subseteq S^*$.

4 Proposed Method

Proposed ε-ICP Theorem: Noting that Theorem 1 does not make assumptions on the f. In real-world applications, it is reasonable to specify the search space of f. One key observation is that if f is differentiable, then the gradient w.r.t an input variable can reflect the 'importance' of the variable. Therefore, we propose a modified version of Assumption 1.

Assumption 2. *(**ε-plausible Causal Variables**) Given $\varepsilon > 0$, $\boldsymbol{X}^u = (x^u_1, x^u_2, \dots, x^u_n)$ and an outcome Y^u for all $u \in U$, there exists a subset of indices $S^*_\varepsilon \subseteq \{1, \dots, n\}$ such that*

$$Y^u = f^*_\varepsilon(\boldsymbol{X}^u) + \delta^u, \ \|\nabla_{S^{*c}_\varepsilon} f^*_\varepsilon\| \le \varepsilon, \text{ where } f^*_\varepsilon : \mathbb{R}^n \mapsto \mathbb{R} \text{ differentiable.} \tag{6}$$

$$\delta^u \text{ are} \begin{cases} \textit{identically distributed} & \textit{if } \exists \textit{ hidden confounders} \\ \textit{identically distributed and } \delta^u \perp\!\!\!\perp \boldsymbol{X}^u_{S^*_\varepsilon} & \textit{else} \end{cases} \tag{7}$$

Here, $\boldsymbol{X}^u_{S^*_\varepsilon}$ are the ε-plausible causal variables. The ε condition seeks for insignificant gradient norm $\|\nabla_{S^{*c}_\varepsilon}\|$ to exclude the non ε-plausible causal variables. If $\varepsilon = 0$ and f is differentiable in Eq. (5), then this is reduced to Assumption 1. With minor changes of Eq. (4,5), we conclude the following

Theorem 2. *We define **ε-identifiable Causal Variables** $S_\varepsilon(U)$ as follows*

$$S_\varepsilon(U) := \bigcap\{S_\varepsilon \subseteq \{1,\dots,n\} \mid \ H^\varepsilon_{0,S_\varepsilon}(U) \textit{ is true.}\} \tag{8}$$

$$H^\varepsilon_{0,S_\varepsilon}(U): \quad \exists \text{ differentiable } f_\varepsilon : \mathbb{R}^n \mapsto \mathbb{R} \textit{ s.t. } Y^u = f_\varepsilon(\boldsymbol{X}^u) + \delta^u,\ \|\nabla_{S^c_\varepsilon} f_\varepsilon\| \le \varepsilon,\ \forall u \in U, \tag{9}$$

where δ^u satisfy (7). If f in Eq. (5) is differentiable, for all $\varepsilon > 0$ it holds $S(U) \subseteq S_\varepsilon(U) \subseteq S^*_\varepsilon$. Moreover, the equality holds for $S(U)$ and $S_\varepsilon(U)$ if $\varepsilon = 0$ (See Appendix B for the proof).

Proposed ε-ICP Algorithm: To discover the identifiable causal variables, the existing ICPs proposed to exhaustively estimate each $H_{0,S}(U)$ (Theorem 1) related function (Eq. (5)). This is less efficient and memory consuming when dealing with large-scale datasets. Instead, we relax the strict verification. More specifically, we propose to approximate the single differentiable $f^*_\varepsilon : \mathbb{R}^n \mapsto \mathbb{R}$ (Eq. (6)), and lower the ε w.r.t the gradient norm of non ε-plausible causal variables in parallel. Concretely, given data from multiple environments, we learn a parameterized differentiable function $f(,;\boldsymbol{\theta})$ in a min-min optimisation fashion (See Alg. 1). Under the single environment that assembles all the data, we estimate the cause-effect relation f^*_ε. Under the experimental environments, we not only approximate f^*_ε but also suppress the gradients of non ε-plausible causal variables. Consequently, the final regularized loss function is

$$\frac{1}{b}\sum_i \lambda_l L(y_i, f(\boldsymbol{x_i};\boldsymbol{\theta}))) + \lambda_r R(\nabla_{\boldsymbol{x_i}} f(\boldsymbol{x_i};\boldsymbol{\theta})), \tag{10}$$

where L is a loss function, R is a regularizer of the gradient norm, b is the batch size.

Algorithm 1 ε-ICP

```
Require: Differentiable function (program) $f(;\boldsymbol{\theta})$, data from experimental environments $E$, batch size $b$, training steps $h$, etc.
for $t \leftarrow 1, h$ do
    Randomly sample data $(\boldsymbol{x}_1, y_1), \dots, (\boldsymbol{x}_b, y_b)$ from the entire dataset
    Update the weights $\boldsymbol{\theta}$ by minimizing $\frac{1}{b}\sum_i \lambda_l L(y_i, f(\boldsymbol{x}_i;\boldsymbol{\theta})))$ of Eq. (10)
    Randomly sample an environment $e$ from $E$
    Randomly sample data $(\boldsymbol{x}^e_1, y^e_1), \dots, (\boldsymbol{x}^e_b, y^e_b)$ from $e$
    Update the weights $\boldsymbol{\theta}$ by minimizing $\frac{1}{b}\sum_i \lambda_l L(y^e_i, f(\boldsymbol{x^e_i};\boldsymbol{\theta}))) + \lambda_r R(\nabla_{\boldsymbol{x^e_i}} f(\boldsymbol{x^e_i};\boldsymbol{\theta}))$ of Eq. (10)
end for
```

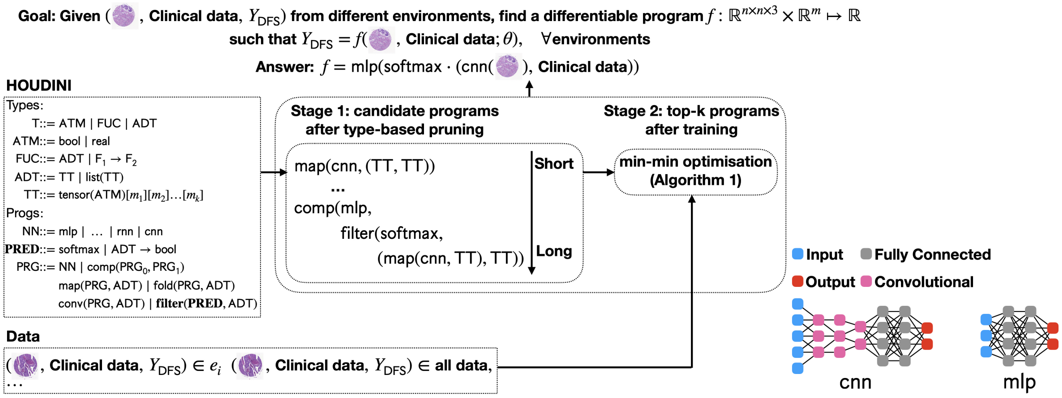

Goal: Given (, Clinical data, Y_{DFS}) from different environments, find a differentiable program $f : \mathbb{R}^{n\times n\times 3} \times \mathbb{R}^m \mapsto \mathbb{R}$
such that $Y_{\mathrm{DFS}} = f($, Clinical data; $\theta)$, ∀environments
Answer: f = mlp(softmax · (cnn( ), Clinical data))
HOUDINI
Types:
T::= ATM | FUC | ADT
ATM::= bool | real
FUC::= ADT | $F_1 \to F_2$
ADT::= TT | list(TT)
TT::= tensor(ATM)$[m_1][m_2]\dots[m_k]$
Progs:
NN::= mlp | … | rnn | cnn
PRED::= softmax | ADT → bool
PRG::= NN | comp(PRG_0, PRG_1)
map(PRG, ADT) | fold(PRG, ADT)
conv(PRG, ADT) | filter(PRED, ADT)
Stage 1: candidate programs after type-based pruning
map(cnn, (TT, TT))
...
comp(mlp,
filter(softmax,
(map(cnn, TT), TT))
Short
Long
Stage 2: top-k programs after training
min-min optimisation (Algorithm 1)
Data
(, Clinical data, Y_{DFS}) ∈ e_i (, Clinical data, Y_{DFS}) ∈ all data,
...
Input
Fully Connected
Output
Convolutional
cnn
mlp

Figure 1: The illustrative scheme of the ε-ICP integrated program synthesis method.

Modified HOUDINI: To deliver a differentiable and human-understandable function f in Eq. (10), we implement the proposed ε-ICP with a typed neural-symbolic functional language HOUDINI (Valkov et al. (2018)). Compared to other neural-symbolic languages (Gaunt et al. (2017); Mao et al. (2018); Vedantam et al. (2019); Ellis et al. (2020)), HOUDINI presents itself as an ideal candidate for a clinical algorithm implementation (See Appendix C for the language comparison and discussion). Relying on the built-in *top-down iterative refinement* strategy for program search, HOUDINI allows us to synthesize a reliable clinical-specific program. Although HOUDINI provides rich types, *i.e.* atomic bool and real, abstract data type (ADT), tensor, and so on, one issue remains. It lacks the explicit branching mechanism. To address this problem, we introduce the predicate module (**PRED**) containing both the soft (softmax) and hard flow control (boolean function) (See Fig. 1). Together with the novel higher-order **filter**, it is expected that the modified HOUDINI can synthesize more expressive programs with explicit flow control. Furthermore, we adapt the modified HOUDINI to the (visible, invisible) setting. For the visible variables, we consider convolutional neural networks (cnn) as the existing building blocks that process the input image. At the same time, we do not assume prior knowledge on processing invisible variables.

Evaluation Metric: After optimising each program candidate, we sort out the programs with top performances (See Fig. 1). As the loss L merely reflects the overall prediction accuracy, it is not comprehensive for evaluating programs in causal discovery. Under the ICP framework, whether the prediction errors are identically distributed among environments matters. Complementary to L, we utilize the Fréchet inception distance (FID) (Heusel et al. (2017)) for measuring the distribution difference between prediction errors from different environments. The FID score, derived from the Wasserstein distance (Villani (2008)), has turned out to be a reliable metric in measuring the distribution difference between a large amount of generated and real images (Lucic et al. (2017)), it is thus a good fit for our use case when dealing with large-scale datasets. Given two distributions μ_0, μ_1 with mean m_0, m_1 and covariance C_0, C_1, we have $\text{FID}(\mu_0, \mu_1) = \|m_0 - m_1\|_2^2 + \text{Tr}(C_0 - C_1 + 2(C_0 C_1)^{0.5})$.

5 Experiments

In both Intervened MNIST and Swiss CRC experiments, we split the data to training, validation and testing. Specifically, we keep the $60k$ training data of MNIST for training, and split the re-discovered $60k$ testing data (Yadav & Bottou (2019)) to $30k$ validation and $30k$ testing data. Based on the patient ID, we split Swiss CRC. The training, validation and testing data include roughly 80%, 10%, 10% of patients with the related TMA images and clinical data. We configure the experimental environments by taking control of (in)visible variables. Some of the variables are held at either minimum or maximum value for each environment. Then, we learn the gradients on those data samples stratified by extreme values. As the amount of environments increase exponentially with the growing number of controlled variables, we limit the maximum number of controlled variables to be 2. Noting that control on (ε-plausible) causal variables tends to perturb the standard deviation of the outcome, *e.g.* if we stratify Occp, Digit with fixed extreme values, then the standard deviation of Y_0 should be close to zero. Motivated by this observation, we instantiate the regularizer R in Eq. (10) as follows

$$R(D_{\boldsymbol{x_i}} f(\boldsymbol{x_i}; \boldsymbol{\theta})) = \hat{k}\, \|\tanh(\text{std}(\boldsymbol{x_1}, \ldots, \boldsymbol{x_b})) \cdot \nabla_{\boldsymbol{x_i}} f(\boldsymbol{x_i}; \boldsymbol{\theta})\|_2, \tag{11}$$

where the multiplication is element-wise, and $\hat{k} = 1 - \tanh(\text{std}(y_1, \ldots, y_b)/k)$. In plain words, if the standard deviation of y_i is insignificant with respect to a batch of training data sampled from an environment, we decrease the gradients w.r.t. the variables that are not controlled. Accordingly, we instantiate the loss L in Eq. (10) to be the l_1 loss for $Y_{0,1,2}$ and cross-entropy loss for Y_{DFS}. Meanwhile, we extract visible variables with standard cnn backbones. All the models are trained four times and the average performances with standard deviation are recorded. As a result, we determine ResNet (He et al. (2016)) and DenseNet (Huang et al. (2017)) to be the optimal backbone for Intervened MNIST and Swiss CRC (See more experimental details in Appendix D).

We compare the proposed ε-ICP to ε-ICP (GT) with availability of ground-truth visible data and to the baseline without experimental environments. Both ε-ICP (GT) and baseline share the identical model settings as ε-ICP. Such comparisons form the core of our experiments. Complementary to the core experiments, we also investigate the performances of vanilla ICP (Peters et al. (2016)), non-linear ICP (Heinze-Deml et al. (2018)) and active ICP (Gamella & Heinze-Deml (2020)). As existing ICPs discover identifiable causal variables via statistical testing, we discuss the statistical significance (p value) for indirect comparison[2] (See Appendix E for more details). Among all the synthesized programs, we keep the one with the top performance, *i.e.* with the best loss and FID score. For balancing the computational efficiency and accuracy, the FID score is computed with the ensemble training, validation and testing data. As Intervened MNIST and Swiss CRC share an identical data structure, the built-in program synthesis method discovers the same type-safe program

$$f = \mathsf{mlp}(\mathsf{softmax} \cdot (\mathsf{cnn}(\textbf{visible data}),\ \textbf{invisible data})). \tag{12}$$

For more hyperparameter settings and results of synthesized programs please check Appendix F.

[2] From now on, we refer 'causal variable' as **identifiable causal variable** for existing ICPs and **ε-plausible causal variable** for ε-ICP for the sake of simplifying terminology.

Intervened MNIST	Methods	Train Accuracy	Val Accuracy	Test Accuracy	FID(Train, Val)	FID(Train, Test)
Y_0	ε-ICP (GT)	**2.01 ± 0.40**	**2.11 ± 0.04**	**2.20 ± 0.05**	3.81 ± 0.11	4.04 ± 0.11
	ε-ICP	2.08 ± 0.30	2.14 ± 0.02	2.26 ± 0.02	**3.66 ± 0.04**	**3.98 ± 0.04**
	baseline	2.11 ± 0.61	2.38 ± 0.07	2.53 ± 0.07	4.05 ± 0.07	4.45 ± 0.07
Y_1	ε-ICP (GT)	2.99 ± 0.29	**2.91 ± 0.13**	**2.97 ± 0.13**	**4.81 ± 0.40**	**5.04 ± 0.42**
	ε-ICP	**2.97 ± 0.14**	3.05 ± 0.03	3.10 ± 0.06	5.28 ± 0.17	5.45 ± 0.21
	baseline	3.05 ± 0.35	3.15 ± 0.08	3.14 ± 0.09	5.57 ± 0.41	5.97 ± 0.53
Y_2	ε-ICP (GT)	**4.23 ± 0.83**	**4.43 ± 0.03**	**4.62 ± 0.04**	**14.07 ± 0.10**	15.53 ± 0.19
	ε-ICP	4.33 ± 0.75	4.50 ± 0.04	4.64 ± 0.04	14.39 ± 0.50	**14.51 ± 0.38**
	baseline	4.40 ± 0.95	4.54 ± 0.08	4.69 ± 0.09	15.24 ± 1.07	18.01 ± 1.49
Swiss CRC	Methods	Train Accuracy	Val Accuracy	Test Accuracy	FID(Train, Val)	FID(Train, Test)
Y_{DFS}	ε-ICP (GT)	**57.81 ± 5.52 %**	**51.60 ± 0.71 %**	**55.39 ± 2.11 %**	**0.049 ± 0.057**	0.054 ± 0.028
	ε-ICP	54.69 ± 8.77 %	51.32 ± 1.70 %	54.90 ± 0.99 %	0.064 ± 0.049	**0.053 ± 0.034**
	baseline	49.22 ± 4.21 %	46.63 ± 0.91 %	45.65 ± 3.19 %	0.082 ± 0.038	0.113 ± 0.030

Table 2: The results comparison of $Y_{0,1,2}$ for Intervened MNIST and Y_{DFS} for Swiss CRC.

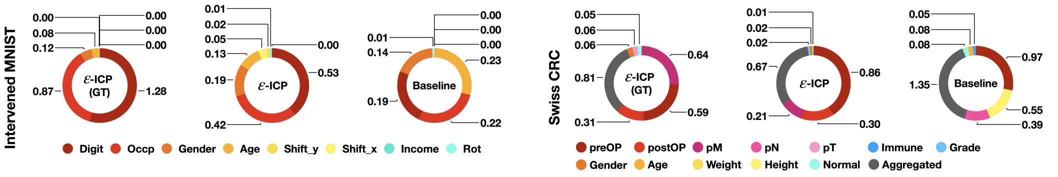


Figure 2: The average gradient norms w.r.t all the variables for Y_0 (Left), and the three most and least important variables for Y_{DFS} (Right), identified by ε-ICP (GT), ε-ICP and baseline.

5.1 Evaluation on Intervened MNIST

Compared to the baseline, Tab. 2 (Top three) presents overall better performances achieved by the proposed ε-ICP in terms of prediction accuracy and FID. Such numerical advantages demonstrate the clear difference between learning with and without experimental environments. As expected, we witness further improvements achieved by ε-ICP (GT), *i.e.* by training the model on the ground-truth visible variables instead of predicted ones. Both ε-ICP (GT) and ε-ICP identify Occp, Digit as the two causal variables with the highest gradient norms, passing the sanity test by ranking Income, Rot as the two least important ones (Fig. 2 (left)). See also Appendix F for $Y_{1,2}$ gradient norms visualization. In contrast, the existing ICPs do not accurately differentiate the causal and non-causal variables for $Y_{0,1,2}$. Take Y_0 as an example, for vanilla and non-linear ICP, the obtained p values are overall smaller than 0.01, suggesting the rejection of $H_{0,S}$ (Eq. 4) for all subsets S. The active ICP identifies Age, Occp, Digit as the causal variables with $p > 0.95$ and rejects the remaining variables with $p < 0.01$, while Age is supposed to be non-causal for Y_0.

Hidden confounders and non-causal variables: To investigate the hidden confounders, we intentionally exclude Digit and Digit + Occp for the ε-ICP. Fig. 3 (Top left) reports the decreasing accuracy and FID scores with the growing number of hidden variables. It is less likely to deliver a meaningful prediction with more hidden causal variables. Meanwhile, the performance achieved by excluding Income and Income + Rot remain competitive; the FID scores present less distribution difference of prediction errors between Train, Val and Test data compared to the case of including non-causal variables. The gradient norm of Occp, Digit stay dominant when excluding Income + Rot, whereas no meaningful causal variables can be identified if Digit, Occp are hidden (Fig. 3 (Top right)).

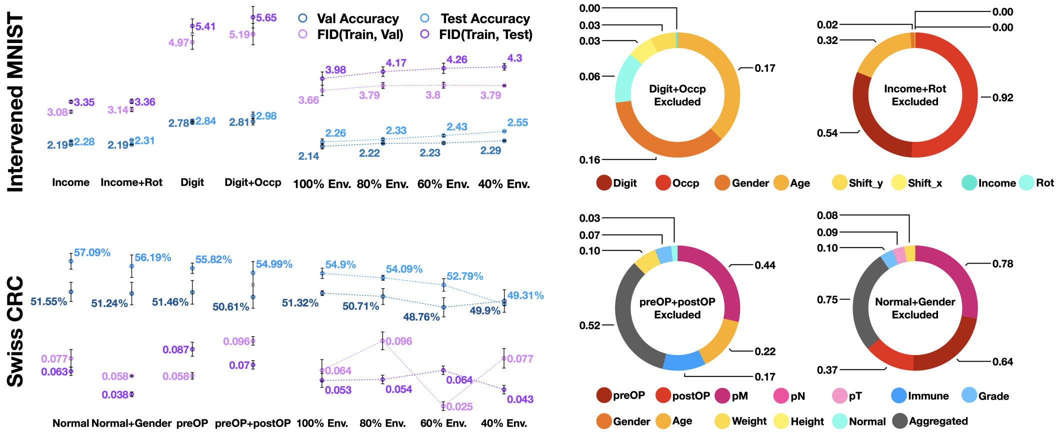


Figure 3: The ablation study of Y_0 for Intervened MNIST (Top) and Y_{DFS} for Swiss CRC (Bottom).

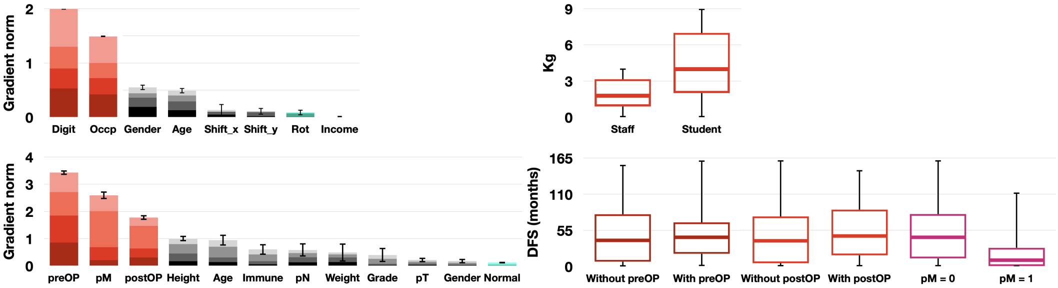


Figure 4: Left: The sum of the average gradient norms of 100% (proposed), 80%, 60%, 40% environments for Y_0 (Top) and Y_{DFS} (Bottom). The error bar presents the standard deviation of ranking. Right: The box plot of Occp for Y_0 and preOP, postOP, pM for Y_{DFS}.

ε-interpretation under environments: With the ensemble gradient norms under environments, we can interpret the synthesized model in more detail (Fig. 4 (Top)). Not only outweigh Digit, Occp the rest of the candidate variables with a clear margin in terms of gradient norm, but both variables remain the two most important causal variables with ranking std = 0, while the remaining variables show variance in ranking. As a result, we can determine $\varepsilon = 1$. Besides, the performances can be mildly improved with the increasing amount of environments (Fig. 3 (Top left)).

5.2 Evaluation on Swiss CRC

As displayed in Tab. 2 (Bottom), the proposed ε-ICP outperforms the baseline with clear margin in terms of better Y_{DFS} accuracy and FID score. The results confirm the necessity of learning the model on experimental environments for causal discovery. Unsurprisingly, ε-ICP (GT) improves the prediction accuracy over proposed ε-ICP when learning on ground-truth visible variables. Both ε-ICP and ε-ICP (GT) identify pM, preOP, postOP as the three most important causal variables (Fig. 2 (right)), all of which are considered to be strongly informative variables based on the pathologist's domain knowledge. Further, ε-ICP and ε-ICP (GT) consistently identify the Normal at the bottom end of the importance scale, validating the consistent ranking of the variables in their importance for the determination of outcomes. As to the existing ICPs, vanilla ICP rejects all subsets of clinical variables with $p < 0.01$. Non-linear ICP identifies preOP as the causal variable with $p = 0.16$ while rejecting the rest variables with $p < 0.01$. Besides, active ICP identifies all the variables (including the non-causal Normal) as causes to the Y_{DFS} with $p > 0.95$. Such identification achieved by existing ICPs is not well supported based on clinical domain knowledge.

Hidden confounders and non-causal variables: We deliberately exclude preOP and preOP + postOP for the study of hidden confounders. As shown in Fig. 3 (bottom), the decreasing performance come as no surprise. Nevertheless, the Y_{DFS} accuracy consistently outperform the baseline without experimental environments. Interestingly, Age, Immune arise as alternatives for preOP, postOP (Fig. 3 (bottom)). The importance of Age, Immune are consistent to the understanding obtained in clinical practice. In parallel, we also investigate how less informative variables impact the ε-ICP. To this end, we exclude Normal and Normal + Gender. By excluding the non-causal variable Normal we achieve mild improvement in terms of Y_{DFS} accuracy (Fig. 3 (bottom)). The effect of excluding Normal and Gender is inconclusive, likely due to the unclear association between Gender and Y_{DFS}.

ε-interpretation under environments: The improved Y_{DFS} accuracy confirms the importance of increasing experimental environments (Fig. 3 (Bottom)). Further, the sum of the average gradient norms under environments offers a finer interpretation of the synthesized model, *i.e.* the ε-ICP agrees on the importance of preOP, postOP, pM (Fig. 4 (Bottom)) with low ranking standard deviation. With the combination of clinical knowledge, $\varepsilon = 1$ can be a meaningful threshold indicating strong informative (causal) variables. Meanwhile, Normal is consistently identified as the least important variable. Lastly, we observe a third group of variables that are assigned less stable associations with Y_{DFS}. Due to the challenging nature of causal inference and the limitations of the present dataset we stay conservative and no clinical recommendation is drawn for these variables.

6 Limitations and future work

As some clinical variables remain inconclusive in Swiss CRC, a follow-up work in a larger, randomized cohort with a more accurate and automatic differentiation becomes necessary. Despite the reasonable Y_{DFS} accuracy achieved in this work, a more precise and personalized DFS (months) prediction is required to reach the goal of precise diagnosis and prognosis. Such a fine-grained stratification study will be conducted on a larger-scale clinical dataset in future work.

References

Ehsan Abbasnejad, Damien Teney, Amin Parvaneh, Javen Shi, and Anton van den Hengel. Counterfactual vision and language learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10044–10054, 2020.

John Aldrich. Autonomy. *Oxford Economic Papers*, 41(1):15–34, 1989.

Fabien Baradel, Natalia Neverova, Julien Mille, Greg Mori, and Christian Wolf. Cophy: Counterfactual learning of physical dynamics. *arXiv preprint arXiv:1909.12000*, 2019.

Pia Bideau and Erik Learned-Miller. It's moving! a probabilistic model for causal motion segmentation in moving camera videos. In *European Conference on Computer Vision*, pp. 433–449. Springer, 2016.

Dmitrii Bychkov, Nina Linder, Riku Turkki, Stig Nordling, Panu E Kovanen, Clare Verrill, Margarita Walliander, Mikael Lundin, Caj Haglund, and Johan Lundin. Deep learning based tissue analysis predicts outcome in colorectal cancer. *Scientific reports*, 8(1):1–11, 2018.

Chun-Hao Chang, Elliot Creager, Anna Goldenberg, and David Duvenaud. Explaining image classifiers by counterfactual generation. *arXiv preprint arXiv:1807.08024*, 2018.

Long Chen, Hanwang Zhang, Jun Xiao, Xiangnan He, Shiliang Pu, and Shih-Fu Chang. Counterfactual critic multi-agent training for scene graph generation. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 4613–4623, 2019.

Long Chen, Xin Yan, Jun Xiao, Hanwang Zhang, Shiliang Pu, and Yueting Zhuang. Counterfactual samples synthesizing for robust visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10800–10809, 2020.

A Philip Dawid, Vanessa Didelez, et al. Identifying the consequences of dynamic treatment strategies: A decision-theoretic overview. *Statistics Surveys*, 4:184–231, 2010.

Kevin Ellis, Catherine Wong, Maxwell Nye, Mathias Sable-Meyer, Luc Cary, Lucas Morales, Luke Hewitt, Armando Solar-Lezama, and Joshua B Tenenbaum. Dreamcoder: Growing generalizable, interpretable knowledge with wake-sleep bayesian program learning. *arXiv preprint arXiv:2006.08381*, 2020.

Zhiyuan Fang, Shu Kong, Charless Fowlkes, and Yezhou Yang. Modularized textual grounding for counterfactual resilience. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6378–6388, 2019.

Karl J Friston, Lee Harrison, and Will Penny. Dynamic causal modelling. *Neuroimage*, 19(4): 1273–1302, 2003.

Tsu-Jui Fu, Xin Wang, Matthew Peterson, Scott Grafton, Miguel Eckstein, and William Yang Wang. Counterfactual vision-and-language navigation via adversarial path sampling. *arXiv preprint arXiv:1911.07308*, 2019.

Juan L Gamella and Christina Heinze-Deml. Active invariant causal prediction: Experiment selection through stability. *arXiv preprint arXiv:2006.05690*, 2020.

Alexander L Gaunt, Marc Brockschmidt, Nate Kushman, and Daniel Tarlow. Differentiable programs with neural libraries. In *International Conference on Machine Learning*, pp. 1213–1222, 2017.

Yash Goyal, Uri Shalit, and Been Kim. Explaining classifiers with causal concept effect (cace). *arXiv preprint arXiv:1907.07165*, 2019a.

Yash Goyal, Ziyan Wu, Jan Ernst, Dhruv Batra, Devi Parikh, and Stefan Lee. Counterfactual visual explanations. *arXiv preprint arXiv:1904.07451*, 2019b.

Sumit Gulwani, Oleksandr Polozov, Rishabh Singh, et al. Program synthesis. *Foundations and Trends® in Programming Languages*, 4(1-2):1–119, 2017.

Zhongyi Han, Benzheng Wei, Yilong Yin, and Shuo Li. Unifying neural learning and symbolic reasoning for spinal medical report generation. *arXiv preprint arXiv:2004.13577*, 2020.

Yosuke Hashimoto, Marek Skacel, Ian C Lavery, Abir L Mukherjee, Graham Casey, and Josephine C Adams. Prognostic significance of fascin expression in advanced colorectal cancer: an immunohistochemical study of colorectal adenomas and adenocarcinomas. *BMC cancer*, 6(1):1–11, 2006.

Alain Hauser and Peter Bühlmann. Two optimal strategies for active learning of causal models from interventional data. *International Journal of Approximate Reasoning*, 55(4):926–939, 2014.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Christina Heinze-Deml, Jonas Peters, and Nicolai Meinshausen. Invariant causal prediction for nonlinear models. *Journal of Causal Inference*, 6(2), 2018.

Lisa Anne Hendricks, Ronghang Hu, Trevor Darrell, and Zeynep Akata. Grounding visual explanations. In *European Conference on Computer Vision*, pp. 269–286. Springer, 2018.

David F Hendry and Mary S Morgan. *The foundations of econometric analysis*. Cambridge University Press, 1997.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in neural information processing systems*, pp. 6626–6637, 2017.

Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4700–4708, 2017.

Atsushi Kanehira, Kentaro Takemoto, Sho Inayoshi, and Tatsuya Harada. Multimodal explanations by predicting counterfactuality in videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8594–8602, 2019.

Peyman Hosseinzadeh Kassani, Li Xiao, Gemeng Zhang, Julia M Stephen, Tony W Wilson, Vince D Calhoun, and Yu Ping Wang. Causality based feature fusion for brain neuro-developmental analysis. *IEEE Transactions on Medical Imaging*, 2020.

Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Thomas Knösel, Anna Emde, Karsten Schlüns, Yuan Chen, Karsten Jürchott, Matthias Krause, Manfred Dietel, and Iver Petersen. Immunoprofiles of 11 biomarkers using tissue microarrays identify prognostic subgroups in colorectal cancer. *Neoplasia (New York, NY)*, 7(8):741, 2005.

Juha Kononen, Lukas Bubendorf, Anne Kallionimeni, Maarit Bärlund, Peter Schraml, Stephen Leighton, Joachim Torhorst, Michael J Mihatsch, Guido Sauter, and Olli-P Kallionimeni. Tissue microarrays for high-throughput molecular profiling of tumor specimens. *Nature medicine*, 4(7): 844–847, 1998.

Hideaki Kume, Satoshi Muraoka, Takahisa Kuga, Jun Adachi, Ryohei Narumi, Shio Watanabe, Masayoshi Kuwano, Yoshio Kodera, Kazuyuki Matsushita, Junya Fukuoka, et al. Discovery of colorectal cancer biomarker candidates by membrane proteomic analysis and subsequent verification using selected reaction monitoring (srm) and tissue microarray (tma) analysis. *Molecular & Cellular Proteomics*, 13(6):1471–1484, 2014.

Brenden M Lake, Tomer D Ullman, Joshua B Tenenbaum, and Samuel J Gershman. Building machines that learn and think like people. *Behavioral and brain sciences*, 40, 2017.

Karel Lebeda, Simon Hadfield, and Richard Bowden. Exploring causal relationships in visual object tracking. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3065–3073, 2015.

Yann LeCun. The mnist database of handwritten digits. *http://yann. lecun. com/exdb/mnist/*, 1998.

Wei Liao, Daniele Marinazzo, Zhengyong Pan, Qiyong Gong, and Huafu Chen. Kernel granger causality mapping effective connectivity on fmri data. *IEEE transactions on medical imaging*, 28 (11):1825–1835, 2009.

David Lopez-Paz, Robert Nishihara, Soumith Chintala, Bernhard Scholkopf, and Léon Bottou. Discovering causal signals in images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6979–6987, 2017.

Mario Lucic, Karol Kurach, Marcin Michalski, Sylvain Gelly, and Olivier Bousquet. Are GANs created equal? a large-scale study. *arXiv preprint arXiv:1711.10337*, 2017.

Ningning Ma, Xiangyu Zhang, Hai-Tao Zheng, and Jian Sun. Shufflenet v2: Practical guidelines for efficient cnn architecture design. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 116–131, 2018.

David Major, Dimitrios Lenis, Maria Wimmer, Gert Sluiter, Astrid Berg, and Katja Bühler. Interpreting medical image classifiers by optimization based counterfactual impact analysis. In *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, pp. 1096–1100. IEEE, 2020.

Jiayuan Mao, Chuang Gan, Pushmeet Kohli, Joshua B Tenenbaum, and Jiajun Wu. The neuro-symbolic concept learner: Interpreting scenes, words, and sentences from natural supervision. In *International Conference on Learning Representations*, 2018.

André C Marreiros, Stefan J Kiebel, and Karl J Friston. Dynamic causal modelling for fmri: a two-state model. *Neuroimage*, 39(1):269–278, 2008.

Khanh Nguyen, Debadeepta Dey, Chris Brockett, and Bill Dolan. Vision-based navigation with language-based assistance via imitation learning with indirect intervention. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 12527–12537, 2019.

Yulei Niu, Kaihua Tang, Hanwang Zhang, Zhiwu Lu, Xian-Sheng Hua, and Ji-Rong Wen. Counterfactual vqa: A cause-effect look at language bias. *arXiv preprint arXiv:2006.04315*, 2020.

Judea Pearl and Dana Mackenzie. *The book of why: the new science of cause and effect*. Basic Books, 2018.

Judea Pearl et al. Causal inference in statistics: An overview. *Statistics surveys*, 3:96–146, 2009.

Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. Causal inference by using invariant prediction: identification and confidence intervals series b statistical methodology. 2016.

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference*. The MIT Press, 2017.

Niklas Pfister, Peter Bühlmann, and Jonas Peters. Invariant causal prediction for sequential data. *arXiv preprint arXiv:1706.08058*, 2017.

William A Reupke, E Srinivasan, Paul V Rigterink, and David N Card. The need for a rigorous development and testing methodology for medical software. In *Proceedings of the Symposium on the Engineering of Computer-Based Medical*, pp. 15–20. IEEE, 1988.

Mateo Rojas-Carulla, Bernhard Schölkopf, Richard Turner, and Jonas Peters. Invariant models for causal transfer learning. *The Journal of Machine Learning Research*, 19(1):1309–1342, 2018.

Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4510–4520, 2018.

Bernhard Schölkopf. Causality for machine learning. *arXiv preprint arXiv:1911.10500*, 2019.

Bernhard Schölkopf, Dominik Janzing, Jonas Peters, Eleni Sgouritsa, Kun Zhang, and Joris Mooij. On causal and anticausal learning. *arXiv preprint arXiv:1206.6471*, 2012.

K Sirinukunwattana, E Domingo, S Richman, K Redmond, A Blake, C Verrill, S Leedham, A Chatzipli, C Hardy, C Whalley, C Wu, A Beggs, U McDermott, P Dunne, A Meade, S Walker, G Murray, L Samuel, M Seymour, I Tomlinson, P Quirke, T Maughan, J Rittscher, and VH Koelzer. Image-based consensus molecular subtype classification (imcms) of colorectal cancer using deep learning. *Gut*, 2020.

Peter Spirtes, Clark N Glymour, Richard Scheines, and David Heckerman. *Causation, prediction, and search*. MIT press, 2000.

Chetan L Srinidhi, Ozan Ciga, and Anne L Martel. Deep neural network models for computational histopathology: A survey. *arXiv preprint arXiv:1912.12378*, 2019.

Kaihua Tang, Yulei Niu, Jianqiang Huang, Jiaxin Shi, and Hanwang Zhang. Unbiased scene graph generation from biased training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3716–3725, 2020.

Brian Taylor, Vasiliy Karasev, and Stefano Soatto. Causal video object segmentation from persistence of occlusions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4268–4276, 2015.

Jin Tian and Judea Pearl. Causal discovery from changes. *arXiv preprint arXiv:1301.2312*, 2013.

Shikhar Uttam, Andrew M Stern, Christopher J Sevinsky, Samantha Furman, Filippo Pullara, Daniel Spagnolo, Luong Nguyen, Albert Gough, Fiona Ginty, D Lansing Taylor, et al. Spatial domain analysis predicts risk of colorectal cancer recurrence and infers associated tumor microenvironment networks. *Nature communications*, 11(1):1–14, 2020.

Lazar Valkov, Dipak Chaudhari, Akash Srivastava, Charles Sutton, and Swarat Chaudhuri. Houdini: Lifelong learning as program synthesis. In *Advances in Neural Information Processing Systems*, pp. 8687–8698, 2018.

Ramakrishna Vedantam, Karan Desai, Stefan Lee, Marcus Rohrbach, Dhruv Batra, and Devi Parikh. Probabilistic neural symbolic models for interpretable visual question answering. In *International Conference on Machine Learning*, pp. 6428–6437, 2019.

Thomas Verma and Judea Pearl. *Equivalence and synthesis of causal models*. UCLA, Computer Science Department, 1991.

Cédric Villani. *Optimal transport: old and new*, volume 338. Springer Science & Business Media, 2008.

Li Wan, Matthew Zeiler, Sixin Zhang, Yann Le Cun, and Rob Fergus. Regularization of neural networks using dropconnect. In *International conference on machine learning*, pp. 1058–1066, 2013.

Pei Wang and Nuno Vasconcelos. Scout: Self-aware discriminant counterfactual explanations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8981–8990, 2020.

Sebastian Weichwald, Timm Meyer, Ozan Özdenizci, Bernhard Schölkopf, Tonio Ball, and Moritz Grosse-Wentrup. Causal interpretation rules for encoding and decoding models in neuroimaging. *NeuroImage*, 110:48–59, 2015.

Chenchu Xu, Lei Xu, Pavlo Ohorodnyk, Mike Roth, Bo Chen, and Shuo Li. Contrast agent-free synthesis and segmentation of ischemic heart disease images using progressive sequential causal gans. *Medical Image Analysis*, pp. 101668, 2020.

Yuanlu Xu, Lei Qin, Xiaobai Liu, Jianwen Xie, and Song-Chun Zhu. A causal and-or graph model for visibility fluent reasoning in tracking interacting objects. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

Chhavi Yadav and Léon Bottou. Cold case: The lost mnist digits. In *Advances in Neural Information Processing Systems*, pp. 13443–13452, 2019.

Gender/Age	7-14	15-29	30-44	≥ 45
Male	3.67	8.3	8.63	10.29
Female	3.13	3.89	5.64	8.72

Table 3: Red meat consumption (times/week)

Yanchao Yang, Ganesh Sundaramoorthi, and Stefano Soatto. Self-occlusions and disocclusions in causal video object segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 4408–4416, 2015.

Kexin Yi, Chuang Gan, Yunzhu Li, Pushmeet Kohli, Jiajun Wu, Antonio Torralba, and Joshua B Tenenbaum. Clevrer: Collision events for video representation and reasoning. *arXiv preprint arXiv:1910.01442*, 2019.

Renqiao Zhang, Jiajun Wu, Chengkai Zhang, William T Freeman, and Joshua B Tenenbaum. A comparative evaluation of approximate probabilistic simulation and deep neural networks as accounts of human physical scene understanding. *arXiv preprint arXiv:1605.01138*, 2016.

A Appendix

In this section, we present more details about Intervened MNIST and Swiss CRC.

Intervened MNIST: (, Personal Data) → Consumption Quantity.

- Visible variables:
  - Digit: $[0, 9]$ numbers, uniformly distributed (Yadav & Bottou (2019)).
  - Shift$_x$: $[-8, 8]$ pixels, uniformly distributed.
  - Shift$_y$: $[-8, 8]$ pixels, uniformly distributed.
  - Rot: $[-15, 15]$ degrees, uniformly distributed.
- Invisible variables:
  - Occp: 0, 1 (Census Bureau staff) or 4 (high school student), given by (Yadav & Bottou (2019))
  - Gender: 0 or 1, uniformly distributed.
  - Age: $[14, 18]$ for high school students, $[19, 57]$ for Census Bureau staff, both uniformly distributed
  - Income: $\mathcal{N}(4188.3, 2000)$, normal distribution (https://nces.ed.gov/programs/digest/d10/tables/dt10_025.asp (1990, monthly income))
- Outcomes:
  - $Y_{0,1,2}$: determined by (in)visible variables.

Noting that the observational data satisfy the above distributions. The interventional data of all the variables (except Digit) satisfy the atomic distribution, resulting from $\mathrm{do}(X = \{\text{minimum}, \text{maximum}\})$ for discrete variables and $\mathrm{do}(X = \{-\sigma, \sigma\})$ otherwise. Meanwhile, we do not intervene the Digit variable and the digit number in each MNIST image remain unchanged. The function $f(\text{age}, \text{gender})$ is based on https://www.nature.com/articles/ejcn201359/tables/1 (See also Tab. 3).

Swiss CRC: (, Clinical Data) → Disease Free Survival.

- Visible variables:
  - Tumor Grade: histopathological tumor grading as low or high grade malignancy, 1 (low grade) or 2 (high grade), median: 2, std: 0.49.
  - Immune (Infiltration): quantitative score of the local anti-tumoral immune response, [0, 3] (absent, low, moderate, strong), median: 1, std: 0.79.
  - Normal: 25% of TMA images are Normal, 75% are Tumor.

- Invisible variables:
  - Gender: 0 (female) or 1 (male), median: 1, std: 0.49.
  - Age: [19.15, 92.12] (years), mean: 69.38, std: 12.29.
  - Weight: [35.0, 139.0] (kilo), mean: 75.11 std: 15.48.
  - Height: [144.0, 196.0] (cm), mean: 169.82 std: 8.90.
  - preOP: pre-operative (neoadjuvant) treatment, 0 (absent) or 1 (present), median: 0 std: 0.35.
  - postOP: post-operative (adjuvant) treatment, 0 (absent) or 1 (present), median: 0, std: 0.44.
  - pT: tumor stage (anatomic extension of the primary tumor), [1, 5] (pT1, pT2, pT3, pT4a, pT4b), median: 3, std: 0.90.
  - pN: nodal stage (presence or absence of lymph node metastasis), [0, 5] (pN0, pN1a, pN1b, pN1c, pN2a, pN2b), median: 0, std: 1.79.
  - pM: presence or absence of distant metastasis, 0 (absent) or 1 (present), median: 0, std: 0.31.
- Outcome:
  - DFS: disease-free survival period after surgery, [0.03, 161.47] (months), mean: 46.85, std: 42.11

Due to large variance and missing data points, we convert the continuous variables Age, Weight, Height, DFS to discrete variables, *i.e.*

- Age: $\leq 65 \rightarrow 0, (65, 75] \rightarrow 1, > 75 \rightarrow 2$.
- Weight: $\leq 70 \rightarrow 0, (70, 80] \rightarrow 1, > 80 \rightarrow 2$.
- Height: $\leq 165 \rightarrow 0, (165, 175] \rightarrow 1, > 175 \rightarrow 2$.
- Y_{DFS}: $\leq 12 \rightarrow 0, (12, 60] \rightarrow 1, > 60 \rightarrow 2$.

B Appendix

Assumption 2. (***ε-plausible Causal Variables***) *Given $\varepsilon > 0$, $\boldsymbol{X}^u = (x_1^u, x_2^u, \dots, x_n^u)$ and an outcome Y^u for all $u \in U$, there exists a subset of indices $S_\varepsilon^* \subseteq \{1, \dots, n\}$ such that*

$$Y^u = f_\varepsilon^*(\boldsymbol{X}^u) + \delta^u, \ \|\nabla_{S_\varepsilon^{*c}} f_\varepsilon^*\| \leq \varepsilon, \ \text{where } f_\varepsilon^* : \mathbb{R}^n \mapsto \mathbb{R} \text{ differentiable.} \tag{13}$$

$$\delta^u \text{ are} \begin{cases} \textit{identically distributed} & \textit{if } \exists \textit{ hidden confounders} \\ \textit{identically distributed and } \delta^u \perp\!\!\!\perp \boldsymbol{X}^u_{S_\varepsilon^*} & \textit{else} \end{cases} \tag{14}$$

Theorem 2. *We define* ***ε-identifiable Causal Variables*** *$S_\varepsilon(U)$ as follows*

$$S_\varepsilon(U) := \bigcap \{S_\varepsilon \subseteq \{1, \dots, n\} \mid H^\varepsilon_{0,S_\varepsilon}(U) \textit{ is true.}\} \tag{15}$$

$$H^\varepsilon_{0,S_\varepsilon}(U): \quad \exists \text{ differentiable } f_\varepsilon : \mathbb{R}^n \mapsto \mathbb{R} \ \textit{s.t.} \ Y^u = f_\varepsilon(\boldsymbol{X}^u) + \delta^u, \ \|\nabla_{S_\varepsilon^c} f_\varepsilon\| \leq \varepsilon, \ \forall u \in U, \tag{16}$$

where δ^u satisfy (14). If f in Eq. (5) differentiable, for all $\varepsilon > 0$ it holds $S(U) \subseteq S_\varepsilon(U) \subseteq S_\varepsilon^*$. Moreover, the equality holds for $S(U)$ and $S_\varepsilon(U)$ if $\varepsilon = 0$.

Proof If there exists a differentiable $\bar{f} : \mathbb{R}^{|\bar{S}|} \mapsto \mathbb{R}$ for $\bar{S} \subseteq \{1, \dots, n\}$ s.t. $H_{0,\bar{S}}(U)$ is true, then we can extend $\bar{f}$ to a map f where

$$\begin{aligned} f: \quad & \mathbb{R}^n \mapsto \mathbb{R} \\ & \underbrace{x^u_{i_0}, \dots, x^u_{i_{|\bar{S}|}}}_{\boldsymbol{X}^u_S}, \underbrace{x^u_{i_{|\bar{S}|+1}}, \dots, x^u_{i_n}}_{\boldsymbol{X}^u_{S^c}} \rightarrow \bar{f}(\boldsymbol{X}^u_{\bar{S}}), \quad \forall u \in U, \end{aligned} \tag{17}$$

then f is differentiable and $\|\nabla_{\bar{S}^c} f\| = 0 \leq \varepsilon$. Hence Eq. (16) holds true for f. Since $\bar{S}$ is arbitrarily selected, we have $S(U) \subseteq S_\varepsilon(U) \subseteq S_\varepsilon^*$. If $\varepsilon = 0$, then for all f_ε in Eq. (15) we can discard the variables with indices in S_ε^c. This brings us the differentiable f_ε required in Eq. (4), hence the equality holds for $S(U)$ and $S_\varepsilon(U)$. □

C Appendix

Unlike generic algorithms, clinical algorithms must deal with unique ethical, legal and technical challenges (Reupke et al. (1988)), because the algorithm output will potentially be directly informative for critical decisions. Error-prone clinical algorithms can be a danger to patient safety. HOUDINI is a typed language, the type-based pruning not only reduces the program search-space dramatically, it also rules out the type errors occurred during evaluation. Due to the high-level nature of a functional language, the synthesized program is more understandable. Of course, neural-symbolic program learning is a far-reaching sub-domain of program synthesis (Gulwani et al. (2017)). Due to the page limit, it is beyond the scope of this paper to thoroughly discuss this problem. We leave it for future work.

	Task	Functional	Typed	Code
NTPT (Gaunt et al. (2017))	Misc.	✗	✗	✗
NS-CL (Mao et al. (2018))	VQA	✓	✗	✓
Prob-NMN (Vedantam et al. (2019))	VQA	✓	✗	✓
DreamCoder (Ellis et al. (2020))	Misc.	✓	✗	✗
HOUDINI (Valkov et al. (2018))	**Misc.**	✓	✓	✓

Table 4: The comparison between existing neural symbolic languages.

D Appendix

For Intervened MNIST, we apply DropConn (Wan et al. (2013)), ResNet50 (He et al. (2016)), MobileNetV2 (Sandler et al. (2018)), ShuffleNetV2 (Ma et al. (2018)) and predict Digit, Rot, Shift_x, Shift_y. After comparing the overall performance, we determine ResNet50 to be the optimal backbone (Fig. 5). For Swiss CRC, we apply DenseNet121 (Huang et al. (2017)), ResNet50, MobileNetV2, ShuffleNetV2 and predict the level of Grade, where we treat level $0, 1, 2$ as the Normal TMA, low and high Grade of tumor TMA. The TMA image is resized to 512×512 before feeding into the models. Accordingly, we determine the DenseNet121 to be the optimal backbone (Fig. 6). As Immune (Infiltration) requires laborious fine-grained annotation for achieving reasonable prediction accuracy, we apply the ground-truth data in this work.

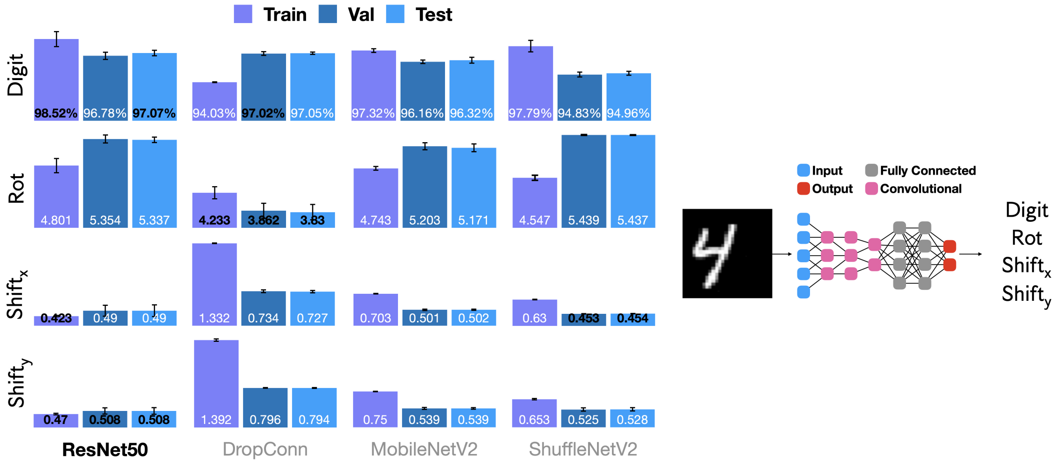


Figure 5: The prediction accuracy of visible variables Digit, Rot, Shift_x, Shift_y achieved by standard cnn backbones. The preferred model and best performances are highlighted with bold black.

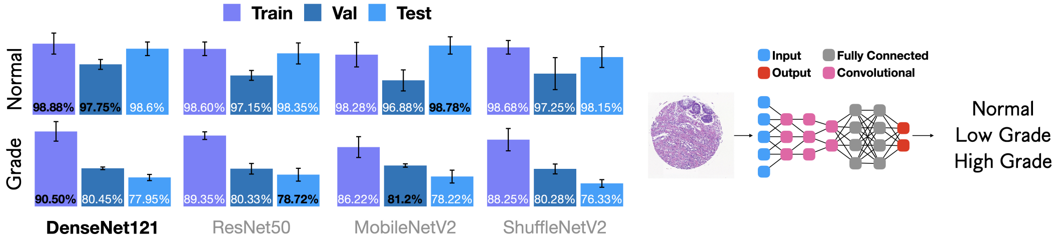


Figure 6: The prediction accuracy of visible variables Normal, Grade achieved by standard cnn backbones. The preferred model and best performances are highlighted with bold black.

E Appendix

We compare the proposed ε-ICP to vanilla ICP (Peters et al. (2016)), non-linear ICP (Heinze-Deml et al. (2018)) and active ICP (Gamella & Heinze-Deml (2020)). The codes are fine-tuned and customized to fit the experimental settings in this paper. For non-linear ICP (Heinze-Deml et al. (2018)), we convert the original R code to Python while keeping the performance comparable. Also, we apply the invariant residual distribution test (E)(ii) using GAM with Levene's test and Wilcoxon test as the benchmark, as this variant presented consistent top performances for different experimental settings. For active ICP (Gamella & Heinze-Deml (2020)), we apply the empty-set strategy as our benchmark, as it showed robustness to different sample sizes and a large number of interventions. For the three ICP methods, the p values w.r.t each $H_{0,S}$ (Eq. 4) are computed and the maximum of all the p values for each variable is taken.

Methods	URL
vanilla ICP	https://github.com/jan-glx/ICPy.git
non-linear ICP	https://github.com/christinaheinze/nonlinearICP-and-CondIndTests.git
active ICP	https://github.com/juangamella/aicp.git

Table 5: The source code of existing ICPs.

F Appendix

In general, the relevant hyperparamters are tuned on the validation data. For training the ε-ICP, we use the standard Adam optimization algorithm (Kingma & Ba (2014)) with the learning rate $0.03, 0.004$ for $Y_{0,1,2}$ and Y_{DFS} (See step 3, 4 in Alg. 1). It suffices to reach top performances within one epoch. Besides, the $\lambda_{\hat{l}}, \lambda_l, \lambda_r$ is determined to be $(2, 2, 5), (10, 0.1, 20)$ for $Y_{0,1,2}$ and Y_{DFS}. Similarly, the k are selected to be $1, 8, 8, 1$ for $Y_0, Y_1, Y_2, Y_{\mathsf{DFS}}$ resp.

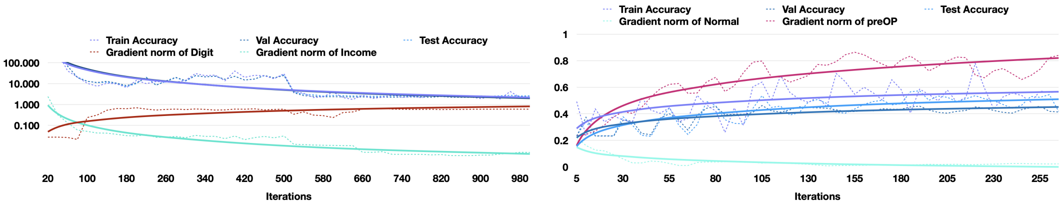


Figure 7: The training curve for Y_0 (Top) and Y_{DFS} (Bottom) achieved by the best synthesized program.

Y_0	Train Accuracy	Val Accuracy	Test Accuracy	FID(Train, Val)	FID(Train, Test)
mlp · mlp((cnn(**visible data**), **invisible data**))	2.10 ± 0.48	2.34 ± 0.11	2.46 ± 0.09	3.62 ± 0.15	4.01 ± 0.17
mlp((cnn(**visible data**), **invisible data**))	2.27 ± 0.51	2.39 ± 0.13	2.44 ± 0.07	4.72 ± 0.30	6.50 ± 0.44
Y_{DFS}					
mlp((cnn(**visible data**), **invisible data**))	52.10 ± 5.54 %	50.05 ± 2.78 %	54.23 ± 1.94%	0.073 ± 0.032	0.064 ± 0.021
mlp · mlp((cnn(**visible data**), **invisible data**))	49.10 ± 4.18 %	46.55 ± 5.89 %	45.46 ± 2.13%	0.088 ± 0.025	0.074 ± 0.037

Figure 8: The performance achieved by the second and third best programs discovered by the built-in program synthesis method.

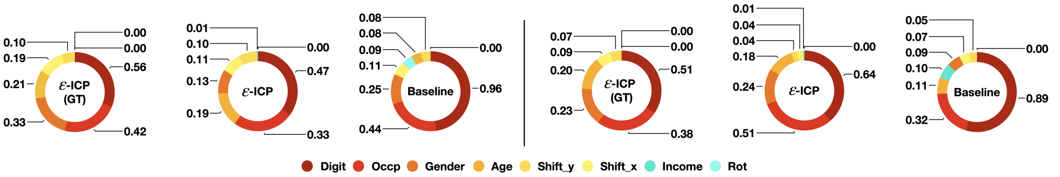


Figure 9: The average gradient norms w.r.t all the variables for Y_1 (left) and Y_2 (Right) identified by ε-ICP (GT), ε-ICP and baseline.