# Causal Representation Learning for Medical Image Analysis

Changjie Lu

## Summary of the Proposal

Statistical machine learning algorithms have achieved state-of-the-art results on benchmark datasets, outperforming humans in many tasks. However, the out-of-distribution data and confounder, which have an unpredictable causal relationship, significantly degrade the performance of the existing models. Causal Representation Learning (CRL) has recently been a promising direction to address the causal relationship problem in vision understanding. This research proposal summarized recent advances in CRL in vision and analyzed the future research objectives. Firstly, we introduced the basic concept of causal inference. Secondly, we analyzed the CRL theoretical work, especially in invariant risk minimization, and the practical work in feature understanding and transfer learning. Finally, we proposed future research objectives in medical image analysis and expected outcomes.

## Background

Correlation does not imply causation [21]. One famous example is the Simpson's paradox[86] (see Fig. 1). Even if the series of events do have causality, it is hard to distinguish that relationship. One effective way of learning causality is to conduct a randomized controlled trial (RCT), randomly assigning participants into a treatment group or a control group so that people can observe the effect via the outcome variable. However, RCT is inflexible because it targets the sample average, which makes the mechanism unclear. Another widely used information type is observational data, which records every event that could be observed. Nowadays, machine learning algorithms attempt to learn patterns by fitting the observational data, losing sight of the causality. This results in poor performance when generalizing the model to an unseen distribution or learning the wrong causality.

To escape the dilemma mentioned above, Pearl firstly introduced a causality system with the three-layer causal hierarchy, called Pearl Causal Hierarchy (PCH), which contains Association, Intervention, and Counterfactuals[63][76][64]. To support this theory, Pearl developed structural causal model (SCM), which combines structural equation models (SEM), potential outcome framework, and the directed acyclic graphs (DAG)[79][74][61][78][65] for probabilistic reasoning and causal analysis, typically using the do-calculus[62]. With these tools, causal analysis infers probabilities under not only statistical conditions but also the dynamics of probabilities under changing conditions [64]. Currently, causal inference is a popular research direction with comprehensive literature [99][35][20][5][107], and is widely applied in decision evaluation (e.g. healthcare), counterfactual estimation (e.g. representation learning methods), and dealing with selection bias (e.g. advertising, recommendation) [99].

Although most of the causal inference could be applied in low-dimensional data (e.g., tabular data, describable events), research in high-dimensional data is still a struggle. In computer vision, for example, which often suffers from confounder that the model tends to classify a cow in dessert as a camel. The invention of advanced network architecture (i.e. resnet[27],transformer[85], etc.) may even enhance this misunderstanding. Recently, much research introduced the task-driven solution, attempting to discover the fundamental mechanism. (e.g. in image deraining, [71][109][36] introduced the progressive algorithm, in low-light image enhancement, [14] imitate the principle of camera imaging, in point cloud analysis, [68][69] introduced point abstraction.) However, these models still struggle in o.o.d prediction.

Causal representation learning (CRL) is a useful tool to unscramble mechanisms. CRL assumes the data are latent causal variables that are causal related and satisfy conditional SCM, using nonlinear mapping. With this assumption, CRL could discover the causal relationship via estimating the distribution after intervention if the causal latent variable and SCM are learnable. Moreover, CRL could even imagine the unseen data according to the counterfactual results, making models

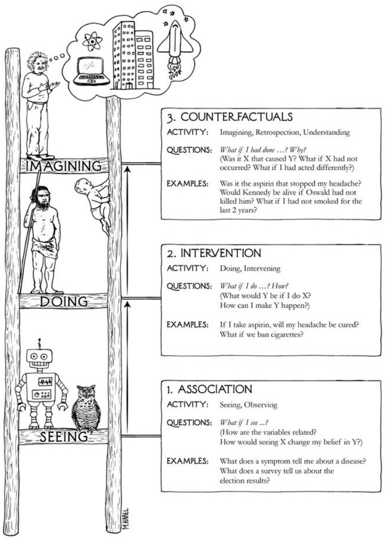

IMAGINING
DOING
SEEING
3. COUNTERFACTUALS
ACTIVITY: Imagining, Retrospection, Understanding
QUESTIONS: What if I had done …? Why?
(Was it X that caused Y? What if X had not occurred? What if I had acted differently?)
EXAMPLES: Was it the aspirin that stopped my headache? Would Kennedy be alive if Oswald had not killed him? What if I had not smoked for the last 2 years?
2. INTERVENTION
ACTIVITY: Doing, Intervening
QUESTIONS: What if I do …? How?
(What would Y be if I do X? How can I make Y happen?)
EXAMPLES: If I take aspirin, will my headache be cured? What if we ban cigarettes?
1. ASSOCIATION
ACTIVITY: Seeing, Observing
QUESTIONS: What if I see …?
(How are the variables related? How would seeing X change my belief in Y?)
EXAMPLES: What does a symptom tell me about a disease? What does a survey tell us about the election results?
MAAYAN

Figure 1: Three levels of Pearl Causal Hierarchy (PCH). (Drawing by [26]) The first level is association $P(y|x)$,(e.g. supervised or unsupervised learning). The second level is intervention $P(y|do(x), c)$, e.g.(feature learning, few-show learning). The third level is counterfactual $P(y_x|x', y')$, (e.g. zero-shot learning, long-tailed classification.)

robust in the o.o.d prediction. However, distinguishing the confounder and discovering the SCM is very challenging. Therefore, some assumptions like sparsity and independent causal mechanism are introduced as an inductive bias in CRL[75]. Recently, theoretical works with CRL have been developed (e.g. Low-rank[70][6], Generalized Independent Noise condition[95][96], invariant risk minimization[3, 73, 37, 53, 54, 110, 106, 55]) and has shown promising performance in feature understanding (e.g. scene graph generation, pretraining, long-tailed data)[87][84][34, 51, 88, 98, 82, 89, 100, 30, 83, 101, 102] and transfer learning problem (e.g. adversarial methods, generalization, adaptation)[104, 97, 32, 80, 33, 31].

In this proposal, we first introduced the basic concept of causal inference (section 1), theoretical work, practical work (section 2), future research directions and objectives in medical image analysis (section 3), and expected outcomes (section 4).

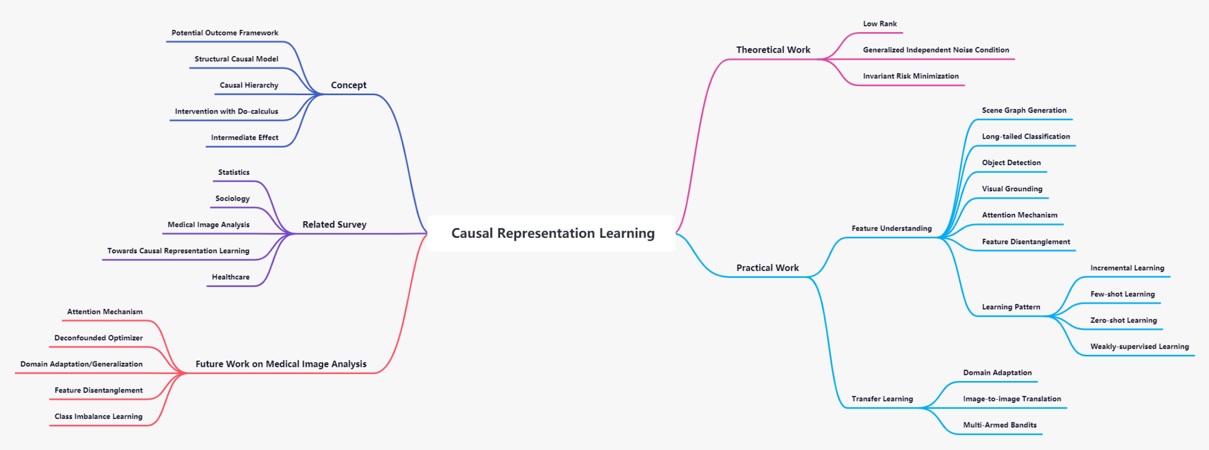


Figure 2: The mindmap of causal representation learning in vision

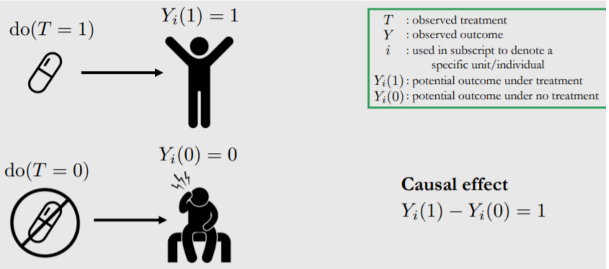
# 1 Concept of Causal Inference

In this section, several concepts of causal reasoning are introduced, including potential outcome framework, structural causal model(SCM), and causal graphs via do-calculus.

| | Condition | | | |
|---|---|---|---|---|
| Treatment | Mild | Severe | Total | Causal |
| A | 15%<br>(210/1400) | 30%<br>(30/100) | **16%**<br>**(240/1500)** | 19.4% |
| B | 10%<br>(5/50) | 20%<br>(100/500) | 19%<br>(105/550) | **12.9%** |

Table 1: The example of Simpson's paradox. Although the total deaths of treatment A is less than deaths of treatment B, the treatment B is a worthy choice in terms of causal analysis. In naive method, for example, the treatment effect of A (16%), is given by $\frac{1400}{1500}(0.15) + \frac{100}{1500}(0.30)$. In causal analysis, the treatment effect of A (19.4%) is given by $\frac{1450}{2050}(0.15) + \frac{600}{2050}(0.30)$.

## 1.1 Potential outcome framework



$\mathrm{do}(T=1)$
$Y_i(1)=1$
$\mathrm{do}(T=0)$
$Y_i(0)=0$
$T$ : observed treatment
$Y$ : observed outcome
$i$ : used in subscript to denote a specific unit/individual
$Y_i(1)$ : potential outcome under treatment
$Y_i(0)$ : potential outcome under no treatment
Causal effect
$Y_i(1) - Y_i(0) = 1$

Figure 3: Example of the potential outcomes (Drawn by [60]). The treatment effect are measured on whether to take the drug or not.

An example in Fig.3 shows how the potential outcomes framework works. However, we are in a dilemma that we cannot observe the causal effect on the same person. If one person takes the drug, we will lose the information of the person not taking the drug. This dilemma is known as the "fundamental problem of causal inference"[107].

## 1.2 Structural Causal Model

Firstly, we define the symbols.

- $X$, random variable
- $P(Y_x) := P(Y \mid do(X = x))$
- path $(X, Y)$: any path from X to Y.
- collider $Z$, $X \to Z \leftarrow Y, X \perp Y, X \not\perp Y \mid Z$.

The structural causal model[63][64] is a 4-tuple $\langle \mathbf{U}, \mathbf{V}, \mathcal{F}, P(\mathbf{U}) \rangle$, where:

- $\mathbf{U}$ is a set of background variables (exogenous variables), determined by factors outside the model.
- $\mathbf{V}$ is a set $\{V_1, V_2, \ldots, V_n\}$ of variables, called endogenous, determined by other variables within the model
- $\mathcal{F}$ is a set of functions $\{f_1, f_2, \ldots, f_n\}$, each $f_i$ is a mapping from $U_i \cup PA_i$(PA: parents) to $V_i$, where $U_i \subseteq U$ and $PA_i$ is a set of causes of $V_i$. The entire set $\mathcal{F}$ forms a mapping from $\mathbf{U}$ to $\mathbf{V}$. That is, for $i$ = 1,...,n, $v_i \leftarrow f_i\,(pa_i, v_i)$.
- $P(\mathbf{U})$ is a probability function defined over the domain of $\mathbf{U}$.

For example, suppose that there exists a causal relationship between treatment solution $X$ and lung function $Y$ of an asthma patient. Simultaneously, suppose that $Y$ also relies on the level of air pollution $Z$. Under this circumstance, $X$ and $Y$ are endogenous variables, $Z$ is exogenous variables. Therefore, the SCM can be instantiated as,

$$\begin{aligned} & U=\{Z, U_x, U_y\}, V=\{X, Y\}, F=\{f_X, f_Y\} \\ & f_X: X \leftarrow f_X(U_x) \\ & f_Y: Y \leftarrow f_Y(X, Z, U_y) \end{aligned} \tag{1}$$

## 1.3 Causal Hierarchy

### 1.3.1 Level1 seeing

For any SCM, the formula,

$$P^{\mathcal{M}}(\mathbf{Y}=\mathbf{y})=\sum_{\{\mathbf{u} \mid \mathbf{Y}(\mathbf{u})=\mathbf{y}\}} P(\mathbf{u}) \tag{2}$$

could estimate any joint distribution of $\mathbf{Y} \subset \mathbf{V}$ given by $\mathbf{Y}(U=u)$. Take the image classification task as an example. $\mathbf{V}=\mathbf{X} \bigcup \mathbf{Y}$, $X$ represents the images, $Y$ represents the labels. The aim is to model $P(\mathbf{Y}|\mathbf{X})$. At this level, we could only build the model by fitting the distribution of observational data.

### 1.3.2 Level2 doing

In the level of doing, a hypothesis is proposed and then verified. In this condition, a new SCM is built: $\mathcal{M}_{\mathbf{x}}=\langle\mathbf{U}, \mathbf{V}, \mathcal{F}_{\mathbf{x}}, P(\mathbf{U})\rangle$, where $\mathcal{F}_{\mathbf{X}}=\{f_i: V_i \notin \mathbf{X}\} \cup(\mathbf{X} \leftarrow \mathbf{x})$. Therefore, $P^{\mathcal{M}}$ can be estimated by,

$$P^{\mathcal{M}}\left(\mathbf{Y}_{\mathbf{x}}=\mathbf{y}_{\mathbf{x}}\right)=\sum_{\{\mathbf{u} \mid \mathbf{Y}_{\mathbf{x}}(\mathbf{u})=\mathbf{y}_{\mathbf{x}}\}} P(\mathbf{u}) \tag{3}$$

where $\mathbf{Y}_{\mathbf{x}}(\mathbf{u})$ represents $\mathcal{F}_{\mathrm{x}}(U=u)$.

### 1.3.3 Level3 imaging

In the level of imaging, the target is to know the effect whether another decision had been made, which can be formulated by,

$$P\left(Y_{x'}^{\prime} \mid X=x, Y=y\right) \tag{4}$$

Namely, imaging $do\left(X=x'\right)$ given $X=x, Y=y$. Based on this, the joint distribution $P^{\mathcal{M}}$ can be estimated by,

$$P^{\mathcal{M}}\left(\mathbf{y}_{\mathbf{x}}, \cdots, \mathbf{z}_{\mathbf{w}}\right)=\sum_{\{\mathbf{u} \mid \mathbf{Y}_{\mathbf{x}}(\mathbf{u})=\mathbf{y}_{\mathbf{x}}, \cdots, \mathbf{Z}_{\mathbf{w}}(\mathbf{u})=\mathbf{z}\}} P(\mathbf{u}) \tag{5}$$

for any $\mathbf{Y}, \mathbf{Z}, \cdots, \mathbf{X}, \mathbf{W} \subset \mathbf{V}$.

## 1.4 Intervention with do-calculus

### 1.4.1 D-separation

Two sets of nodes $X$ and $Y$ are d-separated by a set of nodes $Z$ if all of the paths between any node in $X$ and any node in $Y$ are blocked by $Z$. In Fig. 4, we provide an example to illustrate d-separation. If $X$ is the cause, $Y$ is the effect. The other node is confounding. If W1/W2/W3 and M1/M2 are conditioned on, $X$ and $Y$ are d-separated. If T2 is conditioned on, $X$ and $Y$ are not d-separated because the relationship between $X$ and $Y$ could be found by intervening T2. If both T1 and T2 are conditioned on, $X$ and $Y$ are d-separated because T1 blocks the information path at the bottom of the figure. This concept explains why $X \rightarrow M1 \rightarrow M2 \rightarrow Y$ is causal association and $X \rightarrow W1 \rightarrow W2 \rightarrow W3 \rightarrow Y$, $X \rightarrow T1 \rightarrow T2 \rightarrow T3 \rightarrow Y$ are non-causal associations.

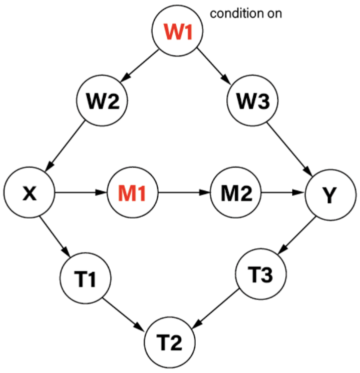


Figure 4: The demonstration for the d-separation. W1 and M1 are conditioned, which block all possible way from $X \to Y$. Therefore, X,Y are d-separated.

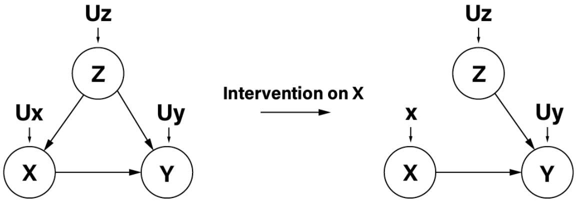


Figure 5: After the intervention on $X = x$, the edge from $Z \to X$ should be deleted.

### 1.4.2 Intervention

We use $do(X = x)$ to represent intervention, $P(Y = y|do(X = x))$ represents the probability of $Y = y$ when making $X = x$. In the graphical model, one edge is deleted to represent the intervention on a particular node. From Fig.5, two invariant equations can be formulated by,

$$\begin{aligned} &P_m(Y = y|Z = z, X = x) = P(Y = y|Z = z, X = x) \\ &P_m(Z = z) = P(Z = z) \end{aligned} \tag{6}$$

$Z$ and $X$ are d-separated after the modification, indicating that,

$$P_m(Z = z|X = x) = P_m(Z = z) = P(Z = z) \tag{7}$$

Therefore,

$$\begin{aligned} &P(Y = y|do(X = x)) = P_m(Y = y|X = x) \\ &= \sum_z P_m(Y = y|X = x, Z = z)P_m(Z = z|X = x) \\ &= \sum_z P_m(Y = y|X = x, Z = z)P_m(Z = z) \end{aligned} \tag{8}$$

Finally, the causal effect formula before intervention can be obtained using the relationship of invariance,

$$P(Y = y|do(X = x)) = \sum_z P(Y = y|X = x, Z = z)P(Z = z) \tag{9}$$

This formula is called the adjustment formula, calculating the relationship between $X$ and $Y$ For every value of $Z$.

Consider the example in Table.1 and the causal graph in Fig.5. $X = A/B$ donates to patients who take the drug A/B. $Z = Mild/Severe$ donates the level of illness. $Y$ donates the death rate. The effect of taking the drug A:

$$E[Y|do(X = A)] = \frac{1450}{2050}(0.15) + \frac{600}{2050}(0.30) \approx 0.194 \tag{10}$$

The effect of taking the drug B:

$$E[Y|do(X = B)] = \frac{1450}{2050}(0.10) + \frac{600}{2050}(0.20) \approx 0.129 \tag{11}$$

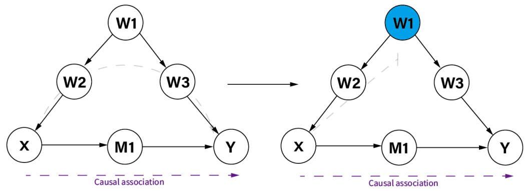


Figure 6: The example of the backdoor adjustment. The second line in Equ.12: W is a sufficient adjustment set, blocking all backdoor paths, only reserving the causation $X \rightarrow Y$. Third line in Equ.12: $do(X)$ blocks all T's parents.

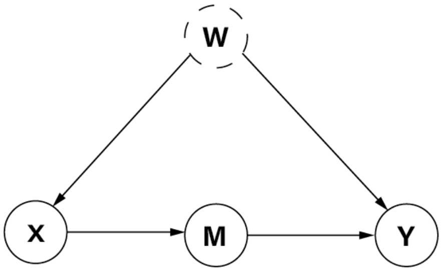


Figure 7: The example of the frontdoor adjustment. Typically, we do twice backdoor adjustments. Firstly, from $X \rightarrow M$, there is no backdoor path. Secondly, from $M \rightarrow Y$, T block the backdoor path $M \leftarrow X \leftarrow W \rightarrow Y$. Therefore, the final equation is defined in step 3.

The result indicates that treatment B has a better effect which is contradictory to the statistical results.

### 1.4.3 Backdoor criterion

$W$ is said to satisfy the backdoor criterion about $(X, Y)$, if $W$:

- blocks all paths between $X$ and $Y$.
- keeps the same directed paths from $X \rightarrow Y$.
- does not produce a new path.

**Backdoor adjustment**: if $W$ satisfies the backdoor criterion, then ATE (Average Treatment Effect) is identified.

$$\begin{aligned} P(y \mid do(X)) &= \sum_w P(y \mid do(X), w) P(w \mid do(X)) \\ &= \sum_w P(y \mid X, w) P(w \mid do(X)) \\ &= \sum_w P(y \mid X, w) P(w) \end{aligned} \tag{12}$$

### 1.4.4 Frontdoor criterion

$W$ is said to satisfy the backdoor criterion about $(X, Y)$, if:

- $W$ blocks all possible directed path from $X \rightarrow Y$.
- There is no backdoor path from $X \rightarrow W$
- All possible paths from $W \rightarrow Y$ are blocked by $X$.

**Frontdoor adjustment**:

- X on M: $P(m \mid do(x)) = P(m \mid x)$

- M on Y: $P(y \mid do(m)) = \sum_x P(y \mid m, x)P(x)$
- X on Y: $P(y \mid do(x)) = \sum_m P(m \mid do(x))P(y \mid do(m)) = \sum_m P(m \mid x) \sum_{x'} P\left(y \mid m, x'\right) P\left(x'\right)$

### 1.4.5 Intermediate Effect

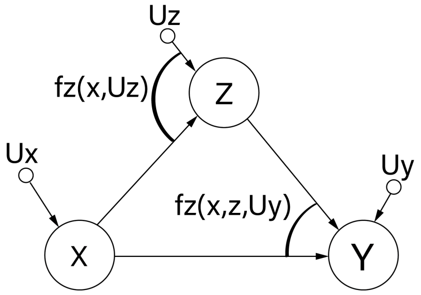


Figure 8: The mediation model without confounding.

In a causal model, a classical intermediate problem can be defined as:

$$x = f_X\left(U_X\right), z = f_Z\left(x, U_Z\right), y = f_Y\left(x, z, U_Y\right) \tag{13}$$

where $X$ donates treatment, $Z$ donates mediator, $Y$ donates outcome. $f_X, f_Z, f_Y$ are any function. $U_X, U_Z, U_Y$ are background variables (see. Fig.8). Based on this, the effect is analyzed by the intervention as follows,
Average Treatment Effect:

$$\begin{aligned} ATE &= E[Y_1] - E[Y_0] \\ &= E[Y \mid do(X=1)] - E[Y \mid do(X=0)]] \end{aligned} \tag{14}$$

Controlled Direct Effect:

$$\begin{aligned} CDE(z) &= E\left[Y_{1,Z} - Y_{0,Z}\right] \\ &= E[Y \mid do(X=1, Z=z)] - E[Y \mid do(X=0, Z=z)] \end{aligned} \tag{15}$$

Natural Direct Effect:

$$NDE = E\left[Y_{1,Z_0} - Y_{0,Z_0}\right] \tag{16}$$

Natural Indirect Effect:

$$NIE = E\left[Y_{0,Z_1} - Y_{0,Z_0}\right] \tag{17}$$

Total Direct Effect:

$$TDE = E\left[Y_{1,Z_1} - Y_{0,Z_1}\right] \tag{18}$$

Total Indirect Effect:

$$TIE = E\left[Y_{1,Z_1} - Y_{1,Z_0}\right] \tag{19}$$

## 2 Causal Representation Learning in Vision

In this section, we introduce recent advances of CRL in theoretical works and practical works.

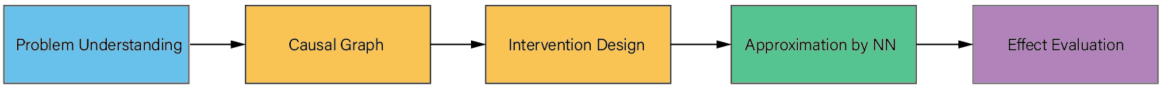


Figure 9: The general workflow for causal representation learning in vision.

## 2.1 Theoretical Work

Traditional machine learning methods attempt to minimize the empirical risk (ERM). However, these methods can not be generalized to the unseen domain. Moreover, ERM is often over-parameterized, resulting in the discovery of spurious correlations. Based on the concept of causal inference, invariant risk minimization[3] is introduced to solve the problems above. Consider a data representation $\Phi : \mathcal{X} \to \mathcal{H}$. we desire an invariant predictor across the environment $w \circ \Phi$. If there exists a predictor $w : \mathcal{H} \to \mathcal{Y}$ achieve optimal performance in every environment $\mathcal{E}_{\text{tr}}$ such that $w \in \underset{\bar{w}:\mathcal{H}\to\mathcal{Y}}{\arg\min} R^e(\bar{w} \circ \Phi)$. Therefore, the constrained optimization problem is defined as:

$$\begin{array}{cl} \min_{\substack{\Phi:\mathcal{X}\to\mathcal{H} \\ w:\mathcal{H}\to\mathcal{Y}}} & \sum_{e\in\mathcal{E}_{\text{tr}}} R^e(w \circ \Phi) \\ \text{subject to} & w \in \underset{\bar{w}:\mathcal{H}\to\mathcal{Y}}{\arg\min} R^e(\bar{w} \circ \Phi), \text{ for all } e \in \mathcal{E}_{\text{tr}} \end{array} \tag{20}$$

To solve it via gradient descent method, an alternative penalty term is proposed,

$$\min_{\Phi:\mathcal{X}\to\mathcal{Y}} \sum_{e\in\mathcal{E}_{\text{tr}}} R^e(\Phi) + \lambda \cdot \left\| \nabla_{w|w=1.0} R^e(w \cdot \Phi) \right\|^2 \tag{21}$$

where D measures the risk when changing the environment, $\lambda \in [0, \infty)$ is a hyper-parameter balancing the ERM and the IRM. This paper only discussed the linear condition of $w$. In the colored MNIST synthetic task, the IRM method achieved 66.9% accuracy, whereas the ERM method could only achieve 17.1% accuracy, even lower than random guessing.

Despite the excellent performance on linear conditions, [73] proved that IRM can not find the optimal invariant predictor on most occasions and even suffer from a catastrophic failure in a particular condition. This paper first introduced and analyzed the non-linear scene. [37] deeply discussed the limitation of the IRM and proved that the IRM prefers an invariant predictor with worse o.o.d generalization. Moreover, the invariant loses its effect when using IRM on empirical samples rather than the population distributions. To address the problem that the IRM fails in the non-linear condition, [53] utilized the Bayesian inference to diminish the over-fitting problem and tricks like ELBO and reparameterization to accelerate convergence speed. [110] proposed sparse IRM to prevent the spurious correlation leaking to sub-models. On the dataset side, [54] proposed a Heterogeneous Risk Minimization (HRM) structure to address the mixture of multi-source data without a source label. Based on this work, [55] extended to kernel space, enhancing the ability to deal with more complex data and invariant relationships. [106] thought that the IRM method could not deal with the relationship between input and output in different domains. Therefore, [106] tried to modify the parameters to make the model robust when domain shifting based on mate-learning structure.

IRM method also mutually benefited from another theory. [46] combined information bottleneck with IRM for domain adaptation. The loss is designed in a mutual information expression, whose structure is similar to the IRM.

## 2.2 Practical Work

Feature understanding and transfer learning are two specific research applications in CRL. Confounders are common in vision datasets, which mislead the machine model into catching the bad relationship. Research in feature understanding in CRL attempts to build the SCM and intervene in the node to discover the causal relationship. For transfer learning, statistical methods suffer from the o.o.d data. Discovering the causal or avoiding the adaptation risk reduce the complexity of training a new transfer learning algorithm and improve performance. This section will present the recent CRL advances in feature understanding and transfer learning.

### 2.2.1 Feature understanding

In object detection tasks, the highly correlated objects tend to occur in the same image (e.g. the chair and human, because people could sit on the chair instead of commensalism). VC R-CNN[87] thought that observational bias made the model ignore the common causal relationship. They introduced an intervention (confounder dictionary) to measure the true causal effect.

In scene graph generation, [84] compared the counterfactual scene and factual scene, using a Total Direct Effect (TDE) analysis framework to remove the bias in training. [82] proposed Align-RCNN to discover the feature relationship and concatenate those features dynamically.

| Task | Performance Improvement over 2nd Model | | |
|---|---|---|---|
| Scene Graph Generation[84] | Predicate Classification | Scene Graph Classification | Scene Graph Detection |
| | 51.1% | 56.4% | 31.3% |
| | Zero-Shot Relationship Retrieval | Sentence-to-Graph Retrieval | |
| | 25.0% | 33.7% | |
| Image Captioning[98] | Karpathy Split[38] 5 Captions | Karpathy Split Whole Set | MS-COCO[52] |
| | 0.250% | 1.55% | 1.02% |
| Attention Mechanism[89] | CNN-Based NICO[28] | CNN-Based ImageNet-9[15] | CNN-Based ImageNet-A |
| | 8.72% | 1.15% | 8.33% |
| | ViT-Based NICO | ViT-Based ImageNet-9 | ViT-Based ImageNet-A |
| | 8.08% | 2.78% | 12.7% |
| Few-shot Learning[101] | miniImageNet | tieredImageNet | |
| | 2.40% | 0.94% | |
| Long-tailed Classification[83] | LVIS V1.0[25] val set | ImageNet-LT | LVIS V0.5 val set |
| | 29.2% | 17.3% | 19.1% |
| | LVIS V0.5 eval test server | | |
| | 18.8% | | |
| Incremental Learning[30] | CIFAR-100[42] | ImageNet-Sub | ImageNet-Full |
| | 6.17% | 4.76% | 3.49% |
| Image Recognition[87] | MS-COCO | Open Images[44] | VQA2.0 test[2] |
| | 1.41% | 1.09% | 0.560% |
| Visual Grounding[34] | RefCOCO[39] | RefCOCO+ | RefCOCOg |
| | 2.09% | 2.36% | 1.77% |
| | ReferIt Game | Flickr30K Entities[67] | |
| | 1.14% | 0.36% | |
| Weakly-Supervised Learning[102] | PASCAL VOC 2012[19] | MSCOCO | |
| | 1.84% | 2.45% | |

Table 2: The performance improvement table for causal representation learning in practical works. Most of the works are plug-and-play and robust to other downstream tasks with only a few costs in total parameters.

In visual grounding (visual language tasks), the location of the target bounding box highly depends on the query instead of causal reasoning. [34] proposed a plug-and-play framework Referring Expression Deconfounder (RED), to make the backdoor adjustment to find the causal relationship between images and sentences.

In image captioning (visual language tasks), [98] thought that the pre-training model contains the confounder. The authors introduced the backdoor adjustment to deconfound the bias. Few-shot learning[91] also requires the pre-training model. Similarly, [101] proposed three solutions, including feature-wise adjustment, class-wise adjustment, and class-wise adjustment. These operations do not need to modify the backbone and could be applied easily in zero-shot learning, meta-learning[29] etc.

In the attention mechanism, most models worked well due to the i.i.d of the data. However, the o.o.d data degrades the performance when using attention. [89] proposed Casual Attention Module (Caam) on original CBAM-based CNN[92] and ViT[16]. This method utilized the IRM and adversarial training[23] with a partitioned dataset to discover confounding factor characteristics.

In feature (disentanglement) representation learning, the classical work simCLR[12] proposed a Self-Supervised Learning (SSL), using a contrastive objective method to recognize similar images. However, the generalization performance and the interpretability are poor. [88] introduced Iterative Partition-based Invariant Risk Minimization (IP-IRM), combining feature disentanglement (data partition) and IRM. This method could discover the critical causal representation for classification tasks. [100] proposed a counterfactual generation framework for zero-shot learning tasks[45] based on counterfactual faithfulness theorem. It designed a two-element classifier, disentangling the feature in class and sample.

In incremental learning[59], [30] used causal-effect theory to explain the forget and anti-forget. It designed models for distilling the causal effect of data, colliding effect, and removing SGD momentum[81] (optimizer) effect of guaranteeing the effectiveness of introducing new data to learn. The interference of the SGD momentum also occurs in long-tailed classification because the update direction of SGD contains the information on data distribution. [83] designed a multi-head normalized classifier in training and made counterfactual TDE inference in testing to remove the excessive tendency for head class.

In weakly-supervised learning[111] for semantic segmentation, the problems often occur in pseudo-masks, for example, object ambiguity, incomplete background, and incomplete foreground. [102] introduced a causal intervention——blocking the connection between context prior and images to remove the fake association between label and images. [102] proposed a context adjustment for the unknown context prior, that is, using a class-specific average mask to approximately constructing a confounder set. [13] proposed a category causality chain and an anatomy-causality chain to solve the ambiguous boundary and co-occurrence problems in medical image segmentation.

### 2.2.2 Transfer learning

One of the popular research directions in transfer learning is domain adaptation. [80] defined invariance specification: a 2-tuple $\langle \mathcal{P}, \mathbf{M} \rangle$, where $\mathcal{P}$ donates graphical representation, $\mathbf{M}$ donates a group of variable (control the data from different environments). If an invariance spec is found, we could get a stable representation such that the model can be applied to the target domain. Similarly, [104] considered domain adaptation as an inference problem, constructing a DAG and solving it by Bayesian inference.

In image-to-image translation, [97] concluded a DAG and developed an important reweighting-based learning method. This method can automatically select the images and perform translation simultaneously.

In general, transfer learning, [43] introduced the new problem of transfer learning for estimating heterogeneous treatment effects and developed several methods (e.g. Y-Learner). [72] proposed invariant models for transfer learning. [103] utilized a causal approach to Multi-Armed Bandits in reinforcement learning.[58] exploited causal inference to predict invariant conditional distribution in domain adaptation without prior knowledge of the causal graph and the type of interventions.

## 3 Research Objectives on Medical Image Analysis

Medical image analysis is a high-risk task, significantly requiring an explainable framework so that doctors and patients can rely on the diagnosis. The current works still lack interpretability and are treated as a black box. Causal representation learning is a promising learning paradigm for medical image analysis. In this section, we mention some prospective research directions.

### 3.1 Attention mechanism

Attention mechanism is widely applied in medical image analysis, and has shown promising results in lots of datasets and tasks[112][77][90][1][4][47], especially for pure attention model (e.g. ViT[16], Swin-Transformer[57], Swin-UNet[8]). The attention mechanism will bring interpretability to the model. For example, in organ segmentation (e.g. cardiac, brain), attention could highlight the features of a region and suppress other noisy parts. However, the attention mechanism suffers from the data distribution shift (CNN-based attention) and the small scale of datasets (Transformer based). [22] shown that the results are noisy on some datasets, even for the cases of attention mechanisms. The causal attention module (Caam)[84] is a promising method to solve this problem without changing the original framework.

### 3.2 Deconfounded optimizer

Most of the works have various settings of the optimizer, and the performance will be damaged if we change to another optimizer. As [30][83] mentioned above, the optimizer (e.g. SGD momentum) can be a confounder in incremental learning and long-tailed classification. The design of a multi-head normalized classifier and counterfactual TDE inference could be a solution in the medical field.

### 3.3 Domain adaptation/generalization

The data distribution shift problem would decrease the performance of the original model due to the various sources from different hospitals. Recently, domain adaptation[93][11][17][18][94][24] and domain generalization[56][105][48][49][50] techniques increased a great deal of accuracy (from ≈ 10% to ≈ 70% in these tasks. However, the model must be trained again when adding a new source for adaptation, and the tricks for domain generalization heavily depend on data pre-processing. Additionally, the performance (≈ 70%) still can not be trusted and applied in clinical application. The series works of IRM [3] provide a new general solution for domain adaptation/generalization problems.

### 3.4 Feature disentanglement

Feature disentanglement attempts to desperate independent feature to make the model explainable. Specifically, classical feature disentanglement methods utilized the Variational Autoencoder[40] or GAN[23] with a restriction in different channels (e.g. minimize mutual information) to disentangle the high-level semantic representation. In domain adaptation, [66] aimed to find the domain-specific feature and the domain invariant feature to make the model robust. In multi-task learning, [10] proposed a dual-stream network to share the common feature in latent space. [9] utilized a VAE to learn a multi-channel spatial representation of the anatomy. However, the restriction for disentanglement is still loose and lacks interpretability in the current work. We could refer to the concept of IP-IRM[88] to discover the causal representation of medical images.

### 3.5 Class imbalance learning

The solution to the class imbalance problem traditionally relies on the data pre-processing (e.g. oversampling[41], re-weighting[108] etc.). But, we can not know the data distribution before training. Additionally, the trick, like re-weighting, will lead the head categories under-fitted. We could refer to [83] to invent a multi-head normalized classifier in training and make counterfactual TDE inference in testing to solve the long-tailed problem.

## 4 Expected Outcomes

The most intuitive result of causal representation learning is to improve the accuracy of tasks (e.g., detection and segmentation). More importantly, causal representation learning has a stronger interpretability of results. Here, we will give an example of outcomes using causal representation learning in self-supervised anomaly detection in medical image segmentation[7]. In the case of abnormal lung lesions, we tried to capture the invariant characteristics and trained an auto-encoder. This auto-encoder ensured that the reconstruction loss was minimal. If the detected image contained the characteristics of the exception, the reconstruction loss would be significant. Therefore, invariant risk minimization could be applied.

$$L_{\mathrm{IRM}}(\Phi, w) = \sum_{e \in \mathcal{E}_{\mathrm{tr}}} R^e(w \circ \Phi) + \lambda \cdot \mathrm{D}(w, \Phi, e) \tag{22}$$

For implementation:

```
scale = torch.Tensor([1.0]).cuda().require_grad_()
loss_rec = rec.err.mean()*scale.mean()
penalty_irm = torch.autograd.grad(loss_rec,scale,create_graph=True)[0]
penalty = torch.sum(penalty_irm**2)*0.2
```

|  | AE | AE+IRM |
|---|---|---|
| Rec | 0.669 | **0.673** |
| Intra | 0.694 | **0.715** |
| Inter | 0.815 | **0.816** |
| Test | 0.672 | **0.680** |

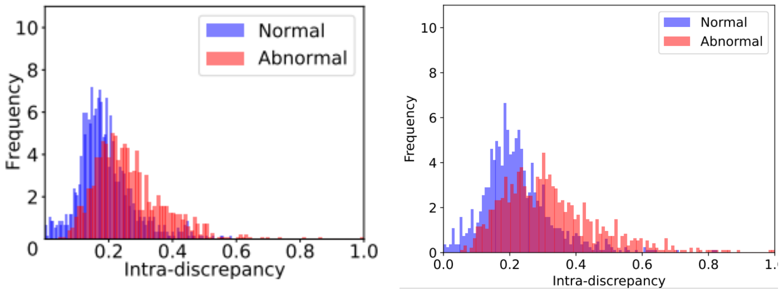


Figure 10: Classification accuracy table (unit: % and classification distribution figures (left:AE, right:AE+IRM).

The results showed IRM improved classification accuracy. At the same time, the reasons for the improvement can also be explained. As shown in the above analysis, IRM loss captures the domain invariant features, leading to a significant increase in abnormal image reconstruction loss. Therefore, the reconstructed fraction distribution of the abnormal image moves to the right, making it easier to distinguish from the normal image.

## 5 Conclusion and Prospect

This research proposal reviewed the development of causal representation learning from concept to application. Firstly, we introduced the basic knowledge of causal inference. Secondly, we analyze the theoretical works on IRM and practical works on feature understanding and transfer learning. Finally, we proposed research objectives for medical image analysis and showed an example of the research outcome.

## References

[1] Nabila Abraham and Naimul Mefraz Khan. A novel focal tversky loss function with improved attention u-net for lesion segmentation. In *2019 IEEE 16th international symposium on biomedical imaging (ISBI 2019)*, pages 683–687. IEEE, 2019.

[2] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015.

[3] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

[4] Reza Azad, Maryam Asadi-Aghbolaghi, Mahmood Fathy, and Sergio Escalera. Attention deeplabv3+: Multi-level context attention mechanism for skin lesion segmentation. In *European conference on computer vision*, pages 251–266. Springer, 2020.

[5] Elias Bareinboim, Juan D Correa, Duligur Ibeling, and Thomas Icard. On pearl's hierarchy and the foundations of causal inference. In *Probabilistic and Causal Inference: The Works of Judea Pearl*, pages 507–556. 2022.

[6] Sumanta Basu, Xianqi Li, and George Michailidis. Low rank and structured modeling of high-dimensional vector autoregressions. *IEEE Transactions on Signal Processing*, 67(5):1207–1222, 2019.

[7] Yu Cai, Hao Chen, Xin Yang, Yu Zhou, and Kwang-Ting Cheng. Dual-distribution discrepancy for anomaly detection in chest x-rays. *arXiv preprint arXiv:2206.03935*, 2022.

[8] Hu Cao, Yueyue Wang, Joy Chen, Dongsheng Jiang, Xiaopeng Zhang, Qi Tian, and Manning Wang. Swin-unet: Unet-like pure transformer for medical image segmentation. *arXiv preprint arXiv:2105.05537*, 2021.

[9] Agisilaos Chartsias, Thomas Joyce, Giorgos Papanastasiou, Scott Semple, Michelle Williams, David E Newby, Rohan Dharmakumar, and Sotirios A Tsaftaris. Disentangled representation learning in cardiac image analysis. *Medical image analysis*, 58:101535, 2019.

[10] Haoxuan Che, Haibo Jin, and Hao Chen. Learning robust representation for joint grading of ophthalmic diseases via adaptive curriculum and feature disentanglement. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 523–533. Springer, 2022.

[11] Cheng Chen, Qi Dou, Hao Chen, Jing Qin, and Pheng Ann Heng. Unsupervised bidirectional cross-modality adaptation via deeply synergistic image and feature alignment for medical image segmentation. *IEEE transactions on medical imaging*, 39(7):2494–2505, 2020.

[12] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

[13] Zhang Chen, Zhiqiang Tian, Jihua Zhu, Ce Li, and Shaoyi Du. C-cam: Causal cam for weakly supervised semantic segmentation on medical image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11676–11685, 2022.

[14] Ziteng Cui, Kunchang Li, Lin Gu, Shenghan Su, Peng Gao, Zhengkai Jiang, Yu Qiao, and Tatsuya Harada. You only need 90k parameters to adapt light: A light weight transformer for image enhancement and exposure correction.

[15] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[16] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[17] Qi Dou, Cheng Ouyang, Cheng Chen, Hao Chen, Ben Glocker, Xiahai Zhuang, and Pheng-Ann Heng. Pnp-adanet: Plug-and-play adversarial domain adaptation network at unpaired cross-modality cardiac segmentation. *IEEE Access*, 7:99065–99076, 2019.

[18] Qi Dou, Cheng Ouyang, Cheng Chen, Hao Chen, and Pheng-Ann Heng. Unsupervised cross-modality domain adaptation of convnets for biomedical image segmentations with adversarial loss. *arXiv preprint arXiv:1804.10916*, 2018.

[19] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results. http://www.pascal-network.org/challenges/VOC/voc2012.

[20] Markus Gangl. Causal inference in sociological research. *Annual review of sociology*, 36:21–47, 2010.

[21] Daniel E Geer Jr. Correlation is not causation. *IEEE Security & Privacy*, 9(2):93–94, 2011.

[22] Tiago Gonçalves, Isabel Rio-Torto, Luís F Teixeira, and Jaime S Cardoso. A survey on attention mechanisms for medical applications: are we moving towards better algorithms? 2022.

[23] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

[24] Mingxuan Gu, Sulaiman Vesal, Ronak Kosti, and Andreas Maier. Few-shot unsupervised domain adaptation for multi-modal cardiac image segmentation. *arXiv preprint arXiv:2201.12386*, 2022.

[25] Agrim Gupta, Piotr Dollar, and Ross Girshick. Lvis: A dataset for large vocabulary instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5356–5364, 2019.

[26] Maayan Harel. Lmu, cmsi 498: your window into the cromulent world of cognitive systems. 2019.

[27] 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*, pages 770–778, 2016.

[28] Yue He, Zheyan Shen, and Peng Cui. Nico: A dataset towards non-iid image classification. 2019.

[29] Timothy Hospedales, Antreas Antoniou, Paul Micaelli, and Amos Storkey. Meta-learning in neural networks: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 44(9):5149–5169, 2021.

[30] Xinting Hu, Kaihua Tang, Chunyan Miao, Xian-Sheng Hua, and Hanwang Zhang. Distilling causal effect of data in class-incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3957–3966, 2021.

[31] Biwei Huang, Fan Feng, Chaochao Lu, Sara Magliacane, and Kun Zhang. Adarl: What, where, and how to adapt in transfer reinforcement learning. *arXiv preprint arXiv:2107.02729*, 2021.

[32] Biwei Huang, Kun Zhang, Mingming Gong, and Clark Glymour. Causal discovery and forecasting in nonstationary environments with state-space models. In *International conference on machine learning*, pages 2901–2910. PMLR, 2019.

[33] Biwei Huang, Kun Zhang, Jiji Zhang, Joseph D Ramsey, Ruben Sanchez-Romero, Clark Glymour, and Bernhard Schölkopf. Causal discovery from heterogeneous/nonstationary data. *J. Mach. Learn. Res.*, 21(89):1–53, 2020.

[34] Jianqiang Huang, Yu Qin, Jiaxin Qi, Qianru Sun, and Hanwang Zhang. Deconfounded visual grounding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 998–1006, 2022.

[35] Guido W Imbens and Donald B Rubin. *Causal inference in statistics, social, and biomedical sciences.* Cambridge University Press, 2015.

[36] Kui Jiang, Zhongyuan Wang, Peng Yi, Chen Chen, Baojin Huang, Yimin Luo, Jiayi Ma, and Junjun Jiang. Multi-scale progressive fusion network for single image deraining. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8346–8355, 2020.

[37] Pritish Kamath, Akilesh Tangella, Danica Sutherland, and Nathan Srebro. Does invariant risk minimization capture invariance? In *International Conference on Artificial Intelligence and Statistics*, pages 4069–4077. PMLR, 2021.

[38] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3128–3137, 2015.

[39] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 787–798, 2014.

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

[41] Sotiris Kotsiantis, Dimitris Kanellopoulos, Panayiotis Pintelas, et al. Handling imbalanced datasets: A review. *GESTS international transactions on computer science and engineering*, 30(1):25–36, 2006.

[42] Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton. Cifar-100 (canadian institute for advanced research).

[43] Sören R Künzel, Bradly C Stadie, Nikita Vemuri, Varsha Ramakrishnan, Jasjeet S Sekhon, and Pieter Abbeel. Transfer learning for estimating causal effects using neural networks. *arXiv preprint arXiv:1808.07804*, 2018.

[44] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, et al. The open images dataset v4. *International Journal of Computer Vision*, 128(7):1956–1981, 2020.

[45] Hugo Larochelle, Dumitru Erhan, and Yoshua Bengio. Zero-data learning of new tasks. In *AAAI*, volume 1, page 3, 2008.

[46] Bo Li, Yifei Shen, Yezhen Wang, Wenzhen Zhu, Dongsheng Li, Kurt Keutzer, and Han Zhao. Invariant information bottleneck for domain generalization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 7399–7407, 2022.

[47] Chen Li, Yusong Tan, Wei Chen, Xin Luo, Yuanming Gao, Xiaogang Jia, and Zhiying Wang. Attention unet++: A nested attention-aware u-net for liver ct image segmentation. In *2020 IEEE International Conference on Image Processing (ICIP)*, pages 345–349. IEEE, 2020.

[48] Haoliang Li, YuFei Wang, Renjie Wan, Shiqi Wang, Tie-Qiang Li, and Alex Kot. Domain generalization for medical imaging classification with linear-dependency regularization. *Advances in Neural Information Processing Systems*, 33:3118–3129, 2020.

[49] Lei Li, Veronika A Zimmer, Wangbin Ding, Fuping Wu, Liqin Huang, Julia A Schnabel, and Xiahai Zhuang. Random style transfer based domain generalization networks integrating shape and spatial information. In *International Workshop on Statistical Atlases and Computational Models of the Heart*, pages 208–218. Springer, 2020.

[50] Lei Li, Veronika A Zimmer, Julia A Schnabel, and Xiahai Zhuang. Atrialgeneral: Domain generalization for left atrial segmentation of multi-center lge mris. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 557–566. Springer, 2021.

[51] Moxin Li, Fuli Feng, Hanwang Zhang, Xiangnan He, Fengbin Zhu, and Tat-Seng Chua. Learning to imagine: Integrating counterfactual thinking in neural discrete reasoning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 57–69, 2022.

[52] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[53] Yong Lin, Hanze Dong, Hao Wang, and Tong Zhang. Bayesian invariant risk minimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16021–16030, 2022.

[54] Jiashuo Liu, Zheyuan Hu, Peng Cui, Bo Li, and Zheyan Shen. Heterogeneous risk minimization. In *International Conference on Machine Learning*, pages 6804–6814. PMLR, 2021.

[55] Jiashuo Liu, Zheyuan Hu, Peng Cui, Bo Li, and Zheyan Shen. Kernelized heterogeneous risk minimization. *arXiv preprint arXiv:2110.12425*, 2021.

[56] Quande Liu, Cheng Chen, Jing Qin, Qi Dou, and Pheng-Ann Heng. Feddg: Federated domain generalization on medical image segmentation via episodic learning in continuous frequency space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1013–1023, 2021.

[57] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021.

[58] Sara Magliacane, Thijs Van Ommen, Tom Claassen, Stephan Bongers, Philip Versteeg, and Joris M Mooij. Domain adaptation by using causal inference to predict invariant conditional distributions. *Advances in neural information processing systems*, 31, 2018.

[59] Marc Masana, Xialei Liu, Bartlomiej Twardowski, Mikel Menta, Andrew D Bagdanov, and Joost van de Weijer. Class-incremental learning: survey and performance evaluation on image classification. *arXiv preprint arXiv:2010.15277*, 2020.

[60] Brady Neal. Introduction to causal inference from a machine learning perspective. *Course Lecture Notes (draft)*, 2020.

[61] Judea Pearl. *Probabilistic reasoning in intelligent systems: networks of plausible inference.* Morgan kaufmann, 1988.

[62] Judea Pearl. A probabilistic calculus of actions. In *Uncertainty Proceedings 1994*, pages 454–462. Elsevier, 1994.

[63] Judea Pearl. Causal inference in statistics: An overview. *Statistics surveys*, 3:96–146, 2009.

[64] Judea Pearl. *Causality.* Cambridge university press, 2009.

[65] Judea Pearl et al. Models, reasoning and inference. *Cambridge, UK: CambridgeUniversityPress*, 19(2), 2000.

[66] Chenhao Pei, Fuping Wu, Liqin Huang, and Xiahai Zhuang. Disentangle domain features for cross-modality cardiac image segmentation. *Medical Image Analysis*, 71:102078, 2021.

[67] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pages 2641–2649, 2015.

[68] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 652–660, 2017.

[69] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in neural information processing systems*, 30, 2017.

[70] Chuan-Xian Ren, Xiao-Lin Xu, and Hong Yan. Generalized conditional domain adaptation: A causal perspective with low-rank translators. *IEEE transactions on cybernetics*, 50(2):821–834, 2018.

[71] Dongwei Ren, Wangmeng Zuo, Qinghua Hu, Pengfei Zhu, and Deyu Meng. Progressive image deraining networks: A better and simpler baseline. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3937–3946, 2019.

[72] 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.

[73] Elan Rosenfeld, Pradeep Ravikumar, and Andrej Risteski. The risks of invariant risk minimization. *arXiv preprint arXiv:2010.05761*, 2020.

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

[75] Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021.

[76] Ilya Shpitser and Judea Pearl. Complete identification methods for the causal hierarchy. *Journal of Machine Learning Research*, 9:1941–1979, 2008.

[77] Ashish Sinha and Jose Dolz. Multi-scale self-guided attention for medical image segmentation. *IEEE journal of biomedical and health informatics*, 25(1):121–130, 2020.

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

[79] Jerzy Splawa-Neyman, Dorota M Dabrowska, and TP Speed. On the application of probability theory to agricultural experiments. essay on principles. section 9. *Statistical Science*, pages 465–472, 1990.

[80] Adarsh Subbaswamy and Suchi Saria. I-spec: An end-to-end framework for learning transportable, shift-stable models. *arXiv preprint arXiv:2002.08948*, 2020.

[81] Ilya Sutskever, James Martens, George Dahl, and Geoffrey Hinton. On the importance of initialization and momentum in deep learning. In *International conference on machine learning*, pages 1139–1147. PMLR, 2013.

[82] Mitra Tajrobehkar, Kaihua Tang, Hanwang Zhang, and Joo-Hwee Lim. Align r-cnn: A pairwise head network for visual relationship detection. *IEEE Transactions on Multimedia*, 24:1266–1276, 2021.

[83] Kaihua Tang, Jianqiang Huang, and Hanwang Zhang. Long-tailed classification by keeping the good and removing the bad momentum causal effect. *Advances in Neural Information Processing Systems*, 33:1513–1524, 2020.

[84] 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*, pages 3716–3725, 2020.

[85] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[86] Clifford H Wagner. Simpson's paradox in real life. *The American Statistician*, 36(1):46–48, 1982.

[87] Tan Wang, Jianqiang Huang, Hanwang Zhang, and Qianru Sun. Visual commonsense r-cnn. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10760–10770, 2020.

[88] Tan Wang, Zhongqi Yue, Jianqiang Huang, Qianru Sun, and Hanwang Zhang. Self-supervised learning disentangled group representation as feature. *Advances in Neural Information Processing Systems*, 34:18225–18240, 2021.

[89] Tan Wang, Chang Zhou, Qianru Sun, and Hanwang Zhang. Causal attention for unbiased visual recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3091–3100, 2021.

[90] Xudong Wang, Shizhong Han, Yunqiang Chen, Dashan Gao, and Nuno Vasconcelos. Volumetric attention for 3d medical image segmentation and detection. In *International conference on medical image computing and computer-assisted intervention*, pages 175–184. Springer, 2019.

[91] Yaqing Wang, Quanming Yao, James T Kwok, and Lionel M Ni. Generalizing from a few examples: A survey on few-shot learning. *ACM computing surveys (csur)*, 53(3):1–34, 2020.

[92] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018.

[93] Fuping Wu and Xiahai Zhuang. Cf distance: A new domain discrepancy metric and application to explicit domain adaptation for cross-modality cardiac image segmentation. *IEEE Transactions on Medical Imaging*, 39(12):4274–4285, 2020.

[94] Fuping Wu and Xiahai Zhuang. Unsupervised domain adaptation with variational approximation for cardiac segmentation. *IEEE Transactions on Medical Imaging*, 40(12):3555–3567, 2021.

[95] Feng Xie, Ruichu Cai, Biwei Huang, Clark Glymour, Zhifeng Hao, and Kun Zhang. Generalized independent noise condition for estimating latent variable causal graphs. *Advances in Neural Information Processing Systems*, 33:14891–14902, 2020.

[96] Feng Xie, Ruichu Cai, Biwei Huang, Clark Glymour, Zhifeng Hao, and Kun Zhang. Generalized independent noise condition for estimating linear non-gaussian latent variable graphs. 2020.

[97] Shaoan Xie, Mingming Gong, Yanwu Xu, and Kun Zhang. Unaligned image-to-image translation by learning to reweight. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14174–14184, 2021.

[98] Xu Yang, Hanwang Zhang, and Jianfei Cai. Deconfounded image captioning: A causal retrospect. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[99] Liuyi Yao, Zhixuan Chu, Sheng Li, Yaliang Li, Jing Gao, and Aidong Zhang. A survey on causal inference. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 15(5):1–46, 2021.

[100] Zhongqi Yue, Tan Wang, Qianru Sun, Xian-Sheng Hua, and Hanwang Zhang. Counterfactual zero-shot and open-set visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15404–15414, 2021.

[101] Zhongqi Yue, Hanwang Zhang, Qianru Sun, and Xian-Sheng Hua. Interventional few-shot learning. *Advances in neural information processing systems*, 33:2734–2746, 2020.

[102] Dong Zhang, Hanwang Zhang, Jinhui Tang, Xian-Sheng Hua, and Qianru Sun. Causal intervention for weakly-supervised semantic segmentation. *Advances in Neural Information Processing Systems*, 33:655–666, 2020.

[103] Junzhe Zhang and Elias Bareinboim. Transfer learning in multi-armed bandit: a causal approach. In *Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems*, pages 1778–1780, 2017.

[104] Kun Zhang, Mingming Gong, Petar Stojanov, Biwei Huang, Qingsong Liu, and Clark Glymour. Domain adaptation as a problem of inference on graphical models. *Advances in Neural Information Processing Systems*, 33:4965–4976, 2020.

[105] Ling Zhang, Xiaosong Wang, Dong Yang, Thomas Sanford, Stephanie Harmon, Baris Turkbey, Bradford J Wood, Holger Roth, Andriy Myronenko, Daguang Xu, et al. Generalizing deep learning for medical image segmentation to unseen domains via deep stacked transformation. *IEEE transactions on medical imaging*, 39(7):2531–2540, 2020.

[106] Marvin Zhang, Henrik Marklund, Nikita Dhawan, Abhishek Gupta, Sergey Levine, and Chelsea Finn. Adaptive risk minimization: Learning to adapt to domain shift. *Advances in Neural Information Processing Systems*, 34:23664–23678, 2021.

[107] Wenhao Zhang, Ramin Ramezani, and Arash Naeim. Causal inference in medicine and in health policy, a summary. *arXiv preprint arXiv:2105.04655*, 2021.

[108] Zizhao Zhang and Tomas Pfister. Learning fast sample re-weighting without reward data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 725–734, 2021.

[109] Shen Zheng, Changjie Lu, Yuxiong Wu, and Gaurav Gupta. Sapnet: Segmentation-aware progressive network for perceptual contrastive deraining. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 52–62, 2022.

[110] Xiao Zhou, Yong Lin, Weizhong Zhang, and Tong Zhang. Sparse invariant risk minimization. In *International Conference on Machine Learning*, pages 27222–27244. PMLR, 2022.

[111] Zhi-Hua Zhou. A brief introduction to weakly supervised learning. *National science review*, 5(1):44–53, 2018.

[112] Wenyong Zhu, Liang Sun, Jiashuang Huang, Liangxiu Han, and Daoqiang Zhang. Dual attention multi-instance deep learning for alzheimer's disease diagnosis with structural mri. *IEEE Transactions on Medical Imaging*, 40(9):2354–2366, 2021.