

# 000 001 002 003 004 005 LEARNING CAUSAL ALIGNMENT FOR RELIABLE 006 DISEASE DIAGNOSIS 007 008 009

010 **Anonymous authors**  
 011 Paper under double-blind review  
 012  
 013  
 014  
 015  
 016  
 017  
 018  
 019  
 020  
 021  
 022  
 023  
 024

## ABSTRACT

025 Aligning the decision-making process of machine learning algorithms with that of  
 026 experienced radiologists is crucial for reliable diagnosis. While existing methods  
 027 have attempted to align their diagnosis behaviors to those of radiologists reflected  
 028 in the training data, this alignment is primarily associational rather than causal, re-  
 029 sulting in pseudo-correlations that may not transfer well. In this paper, we propose  
 030 a causality-based alignment framework towards aligning the model’s decision pro-  
 031 cess with that of experts. Specifically, we first employ counterfactual generation  
 032 to identify the causal chain of model decisions. To align this causal chain with  
 033 that of experts, we propose a causal alignment loss that enforces the model to fo-  
 034 cus on causal factors underlying each decision step in the whole causal chain. To  
 035 optimize this loss that involves the counterfactual generator as an implicit func-  
 036 tion of the model’s parameters, we employ the implicit function theorem equipped  
 037 with the conjugate gradient method for efficient estimation. We demonstrate the  
 038 effectiveness of our method on two medical diagnosis applications, showcasing  
 039 faithful alignment to radiologists.  
 040  
 041

## 1 INTRODUCTION

042 Alignment is essential for developing reliable medical diagnosis systems Zhuang & Hadfield-Menell  
 043 (2020). For instance, in lung cancer diagnosis, using models that are misaligned with clinical pro-  
 044 tocols can result in reliance on contextual features or instrument markers (Fig. 1 (c)) for diagnosis,  
 045 leading to misdiagnosis and loss of timely treatment.

046 Despite the importance, alignment in medical imaging systems is largely understudied. Existing  
 047 studies that are mostly related to us primarily focused on visual alignment, including Zhang et al.  
 048 (2018); Chen et al. (2019); Brady et al. (2023) that proposed to learn object-centric representations,  
 049 and Hind et al. (2019); Rieger et al. (2020) that adopted multi-task learning schemes to predict labels  
 050 and expert decision bases simultaneously. Particularly, recent works Ross et al. (2017); Gao et al.  
 051 (2022); Zhang et al. (2023) have proposed to regularize the model’s input gradient to be within  
 052 expert-annotated areas. However, their alignment with expert behaviors is only associational, rather  
 053 than causal, making their models still biased towards spurious correlated features. This limitation is  
 054 further explained in Fig. 1 (a), where two decision chains with different causal structures can exhibit  
 055 similar correlation patterns.

056 In this paper, we propose a causal alignment approach that focuses on the alignment in the underly-  
 057 ing causal mechanism of the decision-making process. Specifically, we first identify causal factors  
 058 behind each decision step using counterfactual generation. We then propose a causal alignment loss  
 059 to enforce these identified causal factors to be aligned within those annotated by the radiologists.  
 060 To optimize this loss that involves the counterfactual generator as an implicit function of the model  
 061 parameters, we employ the implicit function theorem equipped with the conjugate gradient algo-  
 062 rithm for efficient estimation. To illustrate, we consider the lung cancer diagnosis as shown in Fig. 1  
 063 (b). Guided by the alignment loss, our model can mimic the clinical decision pipeline, which first  
 064 identifies the imaging area that describes attributes of the lesion, and then diagnoses based on these  
 065 attributes Xie et al. (2020). Such training is facilitated by employing causal attribution Zhao et al.  
 066 (2023) for inferring attributes that are causally related to the diagnosis. Returning to the lung cancer  
 067 diagnosis example, Fig. 1 (c) shows that our method can learn causally aligned representations, in  
 068 contrast to the features adopted by baseline methods, which are challenging to interpret.

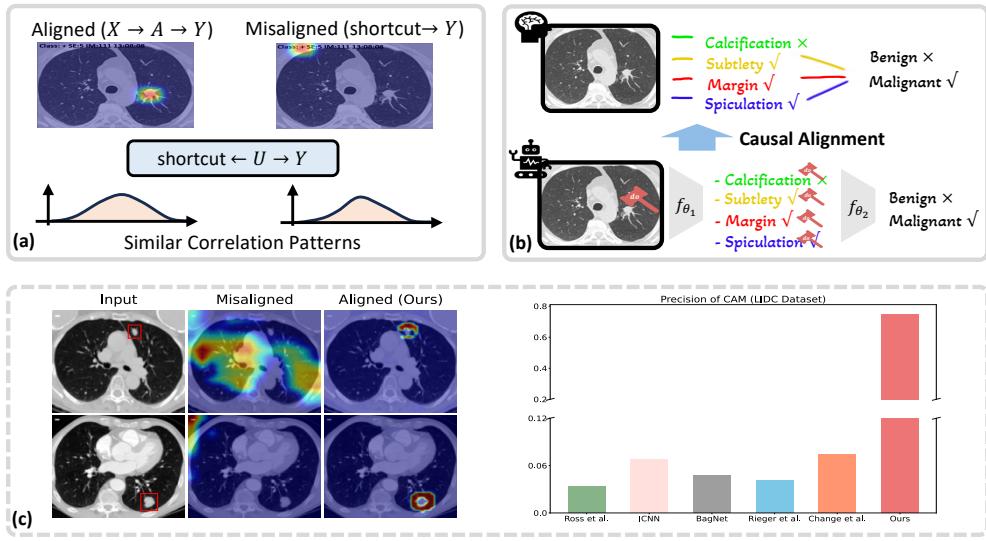


Figure 1: (a) Two decision chains with different causal structures but present similar correlation patterns. The left chain “mass ( $X$ )  $\rightarrow$  attributes ( $A$ )  $\rightarrow$  label ( $Y$ )” aligns with radiologists, while the right chain “shortcut  $\rightarrow$  label ( $Y$ )” is misaligned. However, both ( $X, Y$ ) and (shortcut,  $Y$ ) are correlated, due to the confounding bias between the shortcut and  $Y$ . (b) Our approach for learning causally aligned models. We first identify features and attributes that causally influence the model’s decision, then align them to that of radiologists in a hierarchical fashion. Here, “Calcification”, “Subtlety”, “Margin”, etc., refer to attributes listed in Armato III et al. (2011). The check (resp. cross) mark denotes the presence (resp. absence) of an attribute (c) Class Activation Mapping (CAM) visualization and comparison of CAM precision on lung cancer diagnosis.

**Contributions.** To summarize, our contributions are:

1. **(Causal alignment)** We propose a novel causal alignment approach to achieve alignment of causal mechanisms underlying the decision process of experienced radiologists.
2. **(Optimization)** We propose an efficient optimization algorithm by employing the implicit function theorem along with the conjugate gradient method.
3. **(Experiment)** We demonstrate the utility of our approach through significant improvements in alignment and diagnosis, on lung cancer and breast cancer diagnosis tasks.

## 2 RELATED WORKS

**Learning Visual Alignment.** Alignment is more broadly studied, e.g., in natural language processing Ouyang et al. (2022) and reinforcement learning Ibarz et al. (2018). In the realm of visual alignment, Hind et al. (2019); Rieger et al. (2020) proposed to align deep learning models with humans by simultaneously predicting the class label and the decision area. Zhang et al. (2018); Liu et al. (2021); Müller et al. (2023) aligned the decision-making process of neural networks by incorporating expert knowledge into architecture design. Of particular relevance to our work are Ross et al. (2017); Gao et al. (2022); Zhang et al. (2023), which suggested constraining the input image gradient to be significant in areas annotated by experts. However, the input gradient can be biased by pseudo-correlations that exist between expert features and shortcut features Geirhos et al. (2020), leading to a misaligned model. In contrast, we adopt counterfactual generation to identify causal areas that determine the model’s prediction. By ensuring these factors are confined to expert-annotated areas, our model can be effectively aligned with the expert’s decision process.

**Explaining Medical AI.** Explainability is essential for physicians to trust and utilize medical diagnosis models Lipton (2017). To achieve this, attribution-based methods explained model predictions by assessing the importance of different features Suryani et al. (2022); Yuen (2024). Example-based methods utilized similar images Barnett et al. (2021) or prototypes Gallée et al. (2024) to interpret the underlying decision rules. However, these approaches focused on interpreting models that have

108 been trained. If misalignment occurs during the training process, their utility could be limited. In  
 109 contrast, we propose an alignment loss to learn an **intrinsically explainable model**.  
 110

### 111 3 PROBLEM SETUP & BACKGROUND

113 In this section, we **formulate our problem** and introduce the background knowledge.  
 114

115 **Problem Setup.** We consider the classification scenario, where the system contains an image  $x \in \mathcal{X}$   
 116 and a label  $y \in \mathcal{Y}$  from an **expert annotator**. In addition to  $y$ , we assume the expert also provides  
 117 an explanation  $e$  to explain his decision of labeling  $x$  as  $y$ . Commonly, the explanation could refer  
 118 to **region of interest** annotations or **attribute descriptions**. For example, radiologists often write  
 119 an **annotation** section and an **observation** section, which respectively describe which body part is  
 120 abnormal and what phenomena are observed, in their reports to explain their diagnosis Xie et al.  
 121 (2020). Motivated by this, we assume for each sample, the explanation can be formulated as a binary  
 122 mask  $m$  indicating the **abnormal area**, along with a binary attribute description  $a = [a_1, \dots, a_p] \in \mathcal{A}$   
 123 of the abnormality. In this regard, our data can be denoted as  $\mathcal{D} = \{(x_i, y_i, e_i = (m_i, a_i))\}_{i=1}^n$ .  
 124 With this data, **our goal** is then to **learn** a classifier  $f_\theta : \mathcal{X} \mapsto \mathcal{Y}$  that **i**) predicts  $y$  accurately **ii**) has  
 125 a decision mechanism that is aligned with the radiologists. **Note that our procedure does not depend**  
 126 **on specific model families.**

127 **Structural Counterfactuals Pearl (2013).** To measure the likelihood that one event caused another,  
 128 Pearl (2009) defines the following counterfactual quantity known as the **probability of causation**,

$$P(Y_x = y | X = x_0, Y = y_0), \quad (1)$$

129 which reads as “the probability of  $Y$  would be  $y$  had  $X$  been  $x$  if we factually observed that  $X = x_0$   
 130 and  $Y = y_0$ ”<sup>1</sup>. Here,  $Y_x$  denotes the unit-counterfactual Pearl (2009) or potential outcome. In our  
 131 scenario, rather than considering the whole image  $x$ , we are interested in specific regions within  
 132 the image that causally determine the model’s decision. To identify these regions, we adopt the  
 133 following counterfactual generation scheme.

134 **Counterfactual (CF) Generation.** Given the classifier  $f_\theta$  and any sample pair  $(x_0, y_0)$ , CF generates  
 135 the counterfactual image  $x^*$  with respect to the counterfactual class  $y^* \neq y$  via Dhurandhar  
 136 et al. (2018); Verma et al. (2020); Guyomard et al. (2023); Augustin et al. (2024):  
 137

$$x^* = \arg \min_x \mathcal{L}_{ce}(f_\theta(x), y^*) + \alpha d(x, x_0), \quad (2)$$

138 where  $\mathcal{L}_{ce}$  is the cross-entropy loss for classification,  $d(\cdot, \cdot)$  is a distance metric that constrains the  
 139 modification to be sparse, and  $\alpha$  is the regularization hyperparameter. In this regard, the modified  
 140 area  $supp(x^* - x_0)$  is responsible for the classification of  $x_0$  as  $y_0$ , in that if we modified  $x_0$  to  
 141  $x^*$ , the model would have made a different decision  $y^*$ . Indeed, in Prop. A.5, we can show that  $x^*$   
 142 maximizes the probability of causation  $P_\theta(Y_x = y^* | X = x_0, Y = y_0)$ <sup>2</sup> induced by the classifier  $f_\theta$ ,  
 143 subject to  $d(x, x_0) \leq d_\alpha$  for some  $d_\alpha$ .

144 To ensure the realism of the generated image, we can implement Eq. (2) using gradient descent in  
 145 the image’s latent space. Notably, this approach has proven effective for generating realistic images  
 146 Goyal et al. (2019); Balasubramanian et al. (2020); Zemni et al. (2023), as also verified by the  
 147 visualization of generated counterfactual images in Fig. 9.

148 Indeed, Eq. (2) is similar to but different from the optimization in **Adversarial Attack (AA)** Szegedy  
 149 et al. (2013) concerning **perceptibility** Verma et al. (2020). Although both methods share the same  
 150 objective framework, CF aims at highlighting significant areas that explain the classifier’s decision  
 151 process, whereas AA favors making small and imperceptible changes to alter the prediction outcome  
 152 Wachter et al. (2017). This often leads to different choices of the distance function  $d(\cdot, \cdot)$  and the  
 153 hyperparameter  $\alpha$  Freiesleben (2022); Guidotti (2024).  
 154

### 156 4 METHODOLOGY

158 In this section, we introduce our framework for medical decision alignment. This section is com-  
 159 posed of three parts. First, in Sect. 4.1, we introduce a **causal alignment loss** based on counterfactual

161 <sup>1</sup>Under the *exogeneity* and *monotonicity* conditions for binary  $X, Y$ , this quantity is identifiable.

<sup>2</sup>This term is identifiable since  $f_\theta$  is known (see Prop. A.4 for details).

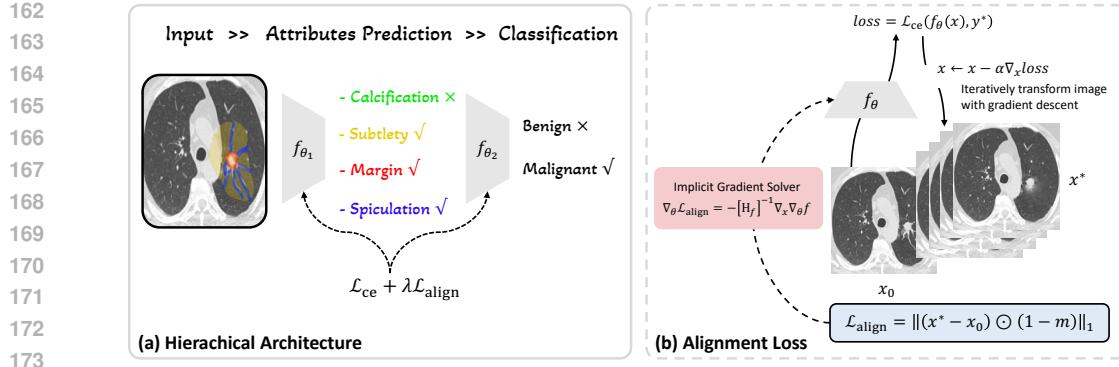


Figure 2: The schematic overview of our method. (a) We adopt a hierarchical structure that first provides attribute descriptions for the image, and then shows the diagnosis result. (b) Training with the proposed alignment loss. In the forward pass, a counterfactual image  $x^*$  is generated and used to compute the alignment loss  $\mathcal{L}_{\text{align}}$  relative to the expert’s annotation  $m$ . In the backward pass, we use an implicit gradient solver to obtain the gradient  $\nabla_{\theta} \mathcal{L}_{\text{align}}$  and use it to update the parameter  $\theta$ .

generation, to align the model’s decision bases with those of experts. Then, in Sect. 4.2, we propose to use the implicit function theorem equipped with conjugate gradient estimation to compute the gradient of our loss for optimization. Finally, in Sect. 4.3, we enhance our method with hierarchical alignment for cases where attribute annotations are available, using a hierarchical pipeline based on causal attribution. We summarize our framework in Fig. 2.

#### 4.1 CASUAL ALIGNMENT LOSS

In this section, we propose a causal alignment loss to align the model with the experts. For illustration, we first consider the case where the attribute annotations are unavailable. The idea of our loss is to penalize the model once its counterfactual image contains modifications beyond radiologist-annotated areas. Specifically, we optimize a loss  $\mathcal{L}_{\text{align}}$  of the following form:

$$\mathcal{L}_{\text{align}} := \frac{1}{n} \sum_{i=1}^n \|(x_i^* - x_i) \odot (1 - m_i)\|_{\ell_1}, \quad (3)$$

where  $\odot$  denotes the element-wise matrix product,  $x_i^*$  is the counterfactual image of  $x_i$  obtained by Eq. (2), and  $m \in \{0, 1\}^{\dim(x)}$  is the binary mask provided by radiologists. Then, by combining  $\mathcal{L}_{\text{align}}$  with the cross-entropy loss for classification, we have our overall training objective:

$$\mathcal{L} = \mathcal{L}_{\text{ce}} + \lambda \mathcal{L}_{\text{align}},$$

where  $\lambda$  is a tuning hyperparameter. To understand how the objective works towards alignment, note that  $x^*$  maximizes the counterfactual likelihood  $P_{\theta}(Y_x = y^* | x_0, y_0)$ , indicating that  $\text{supp}(x^* - x_0)$  represents the causal factors that influence the decision of the model  $f_{\theta}$ . Therefore, minimizing the distance between  $\text{supp}(x^* - x_0)$  and  $m$  encourages the model’s causal factors to align more closely to those of the experts.

Our loss enjoys several advantages over alternative methods in visual alignment. Compared to Liu et al. (2021); Müller et al. (2023) that incorporated prior knowledge into network architectures, our loss is more flexible and can be easily adapted to other scenarios and backbones. In contrast to Ross et al. (2017); Zhang et al. (2023) that constrained the input gradient, our approach can effectively avoid pseudo-features, benefiting from the identification of causal factors.

#### 4.2 OPTIMIZATION

In this section, we introduce the optimization process for the proposed alignment loss. For optimization, we need to compute the gradient  $\nabla_{\theta} \mathcal{L}_{\text{align}}$ , which involves the Jacobian matrix  $\nabla_{\theta} x^*$ . The main challenge here is that  $x^*$  is an *implicit function* of  $\theta$ , defined by the argmin operator in Eq. (2), which makes it hard to compute  $\nabla_{\theta} x^*$  explicitly.

To address this challenge, we resort to the Implicit Function Theorem (IFT), which allows us to compute the gradient in an implicit manner. Specifically, note that if  $x^*$  is the minimum point of the function  $T(x, \theta) := \mathcal{L}_{\text{ce}}(f_\theta(x), y^*) + \alpha d(x, x_0)$ , it should satisfy that:

$$\nabla_x T \Big|_{x^*} = 0.$$

According to the law of total derivation, this implies that:

$$\nabla_\theta \nabla_x T \Big|_{x^*} = \{\nabla_x(\nabla_x T) \cdot \nabla_\theta x^* + \nabla_\theta(\nabla_x T)\} \Big|_{x^*} = 0.$$

Therefore, computing  $\nabla_\theta x^*$  boils down to the problem of solving the following linear equation:

$$Hz^* = b, \quad (4)$$

where we denote  $H := \nabla_x(\nabla_x T)$  as the Hessian matrix,  $z^* := \nabla_\theta x^*$  as the Jacobian matrix, and  $b := -\nabla_\theta(\nabla_x T)$  as the negative mixed derivative for brevity.

Formally speaking, we have the following theorem:

**Theorem 4.1** (Implicit Function Theorem (IFT) Krantz & Parks (2002)). *Consider two vectors  $x, \theta$ , and a differentiable function  $T(x, \theta)$ . Let  $x^* := \arg \min_x T(x, \theta)$ . Suppose that: i) the argmin is unique for each  $\theta$ , and ii) the Hessian matrix  $H$  is invertible. Then  $x^*(\theta)$  is a continuous function of  $\theta$ . Further, the Jacobian matrix  $\nabla_\theta x^*$  satisfies the linear Eq. (4).*

Thm. 4.1 suggests that we can compute the Jacobian matrix using  $\nabla_\theta x^* = -H^{-1}b$ , which then gives  $\nabla_\theta \mathcal{L}_{\text{align}}$  with the chain-rule. Nonetheless, for imaging tasks, typically  $\theta$  is the parameter of high-dimensional neural networks, making it intractable to compute the Hessian matrix and its inverse. To address this issue, we employ the conjugate gradient algorithm Vishnoi et al. (2013) to estimate the solution of Eq. (4), without explicitly computing or storing the Hessian matrix. Notably, the conjugate gradient method has been successfully deployed in Hessian-free methods for deep learning Martens et al. (2010) and meta learning Sitzmann et al. (2020).

We briefly introduce the idea of conjugate gradient below, with a detailed discussion left to Vishnoi et al. (2013) (Chap. 6). To begin with, note that solving Eq. (4) is equivalent to solving:

$$z^* = \arg \min_z g(z), \text{ where } g(z) := \frac{1}{2} z^\top H z - b^\top z,$$

in that the minimum point  $z^*$  satisfies  $\nabla_z g \Big|_{z^*} = Hz^* - b = 0$ .

In this regard, we can implement gradient descent to minimize  $g(\cdot)$ , where the minimum point gives the solution of Eq. (4). During the minimization, the direction of the gradient updating is set to be conjugate (*i.e.*, orthogonal) to the residual  $b - Hz^{(i)}$ , where  $z^{(i)}$  is the estimate of  $z^*$  in the  $i$ -th iteration, in order to achieve optimal convergence rate. To achieve this without explicitly forming  $H$ , we can leverage the *Hessian vector product* Song & Vicente (2022). Specifically, for  $\epsilon$  that is a small perturbation around  $z$ , we have:

$$\nabla g(z + \epsilon z^{(i)}) \approx \nabla g(z) + H \epsilon z^{(i)}.$$

It then follows that:

$$Hz^{(i)} \approx \frac{\nabla g(z + \epsilon z^{(i)}) - \nabla g(z)}{\epsilon},$$

which means we can estimate  $Hz^{(i)}$  with the finite difference of  $\nabla g$  on the right-hand side.

Equipped with Thm. 4.1 especially the conjugate gradient method for estimation, we now summarize the optimization process for our loss in Alg. 1.

#### 4.3 HIERARCHICAL ALIGNMENT

In this section, we extend our method to the scenario where attribute annotations are available. We introduce a hierarchical alignment framework to mimic the clinical diagnostic procedure.

---

**270 Algorithm 1 Causal alignment training**


---

**271 Input:** Data  $\mathcal{D}$ ,  
**272 Output:** Decision model  $f_\theta$ ,  
**273 Hyperparameters:** Sparsity regularization  $\alpha$ , weight of alignment loss  $\lambda$ , learning rate  $\eta$ .  
**274** 1: **while** not converged **do**  
**275** 2: **\*\*Forward pass**  
**276** 3: Compute  $\mathcal{L}_{\text{ce}}$ .  
**277** 4: Optimize Eq. (2) to obtain  $x^*$  and compute  $\mathcal{L}_{\text{align}}$  using Eq. (3).  
**278** 5: Compute  $\mathcal{L} \leftarrow \mathcal{L}_{\text{ce}} + \lambda \mathcal{L}_{\text{align}}$ .  
**279** 6: **\*\*Back propagation**  
**280** 7: Estimate  $\nabla_\theta \mathcal{L}_{\text{align}}$  with conjugate gradient.  
**281** 8: Update  $\theta$ :  $\theta \leftarrow \theta - \eta \nabla_\theta \mathcal{L}$ . // or Adam  
**282** 9: **end while**

---

**283**  
**284 Causal diagram and assumptions.** We characterize this diagnostic process with the causal graph  
**285** in Fig. 3. According to McNitt-Gray et al. (2007); Lee et al. (2017), the first step in the diagnosis is  
**286** annotating each mass attribute from the image Xie et al. (2020). Therefore, we assume causal edges  
**287** from the image  $X$  to the attributes  $A$ . Since these attributes are directly annotated from the image,  
**288** we assume no additional dependencies among them, implying their conditional independence given  
**289**  $X$ . Building on these attributes, we further assume a causal relationship  $A \rightarrow Y$ , representing the  
**290** decision-making process from the attributes to the final decision label.

**291** Specifically, our classifier  $f_\theta$  consists of an  $f_{\theta_1} : \mathcal{X} \mapsto \mathcal{A}$  that  
**292** predicts the attributes from the image  $x$ , and an  $f_{\theta_2} : \mathcal{A} \mapsto \mathcal{Y}$   
**293** that classifies the label based on the predicted attributes, where  
**294**  $\theta_1$  and  $\theta_2$  are optimized in an end-to-end manner. For *counterfactual generation*, we first find attributes responsible for  
**295** predicting  $y$  by altering the predicted attributes  $\hat{a} := f_{\theta_1}(x)$  to the  
**296** counterfactual ones  $a^*$ . Then, we locate image features that  
**297** account for the modification of  $|a^* - \hat{a}|$  via another counterfactual  
**298** optimization over  $x$  and obtain the counterfactual image  
**299**  $x^*$ . For hierarchical alignment, we require both  $|a^* - \hat{a}|$  and  
**300**  $|x^* - x|$  to be aligned with the expert’s annotations of causal  
**301** attributes and image regions, respectively.

**302 Causal Attribution for Annotations.** Although the attribute  
**303** annotations can be available for many cases Armato III et al.  
**304** (2011); Lee et al. (2017), it is hard to know which ones of these  
**305** attributes causally determined the labeling of radiologists for each specific patient. To identify the  
**306** causal attributes for alignment, we employ causal attribution based on counterfactual causal effect  
**307** Zhao et al. (2023), which extends Eq. (1) to enable the quantification of the probability of causation  
**308** for any subsets of attributes while conditioning on the entire attribute vector. Specifically, given  
**309** evidence of the attributes  $A = a$  and the label  $Y = y$ , we calculate the Conditional Counterfactual  
**310** Causal Effect (CCCE) score for each attribute subset  $S \subseteq \{1, \dots, \dim(A)\}$ :

$$\text{CCCE}(S) := \mathbb{E}(Y_{A_S=1} - Y_{A_S=0} | A = a, Y = y),$$

which is the difference between the conditional expectations of the potential outcomes  $Y_{A_S=1}$  and  $Y_{A_S=0}$  given the evidence. Recall that each attribute  $A_i$  is binary. Then, according to Zhao et al. (2023) (Thm. 2),  $\text{CCCE}(S)$  is identifiable and equals to

$$\text{CCCE}(S) \stackrel{(1)}{=} 1 - \frac{P(Y_{A_S=1} = y | A = a)}{P(Y = y | A = a)} \stackrel{(2)}{=} 1 - \frac{P(Y = y | A_S = 1, A_{-S} = a_{-S})}{P(Y = y | A = a)},$$

where  $A_{-S}$  denotes attributes beyond the subset  $S$ . Here, “(1)” arises from the exogeneity condition that there is no confounding between  $A$  and  $Y$  (*i.e.*,  $Y_a \perp\!\!\!\perp A$ ), and “(2)” is based on the monotonicity condition<sup>3</sup> that  $Y_a \leq Y_{a'}$  if  $a \preceq a'$ <sup>4</sup>. Both conditions are natural to hold in our scenario. Specifically, the exogeneity condition holds since the radiologist’s decision  $Y$  is based only on attributes

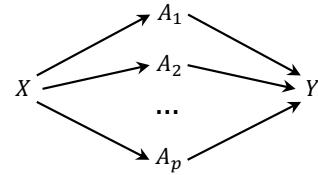


Figure 3: Causal diagram of radiologists’ decision process.  $A$  and  $Y$  denote the expert’s annotations of the attributes and the decision label, respectively.

<sup>3</sup>Zhao et al. (2023) also assumed exogeneity condition and the monotonicity conditions among  $A$ , if there exist causal relations among  $A$ . Since there are no causal relations among  $A$ , we do not need these conditions.

<sup>4</sup>Here,  $a$  and  $a'$  are both vectors.  $a \preceq a'$  denotes  $a_k \leq a'_k$  for each  $k$ .

(Fig. 3). For the monotonicity condition, it is easy to see that each intervention on any attribute from 0 to 1 (*e.g.*, from no speculation to speculation) will raise the probability of malignancy.

After computing the CCCE score for each attribute subset, we select the subset  $S$  with the highest CCCE as the set of attributes causally related to the label. Accordingly, we set the annotation vector  $r \in \{0, 1\}^{\dim(A)}$  such that  $r_S = (1, \dots, 1)^\top$ .

**Hierarchical Alignment.** With such annotations, we introduce our hierarchical alignment process. Specifically, our objective function over  $\theta = (\theta_1, \theta_2)$  is:

$$\mathcal{L}(\theta) := \mathcal{L}_{\text{ce}}(f_{\theta_2}(f_{\theta_1}(x)), y) + \mathcal{L}_{\text{ce}}(f_{\theta_1}(x), a) + \lambda_2 \mathcal{L}_{\text{align}}(\theta_2) + \lambda_1 \mathcal{L}_{\text{align}}(\theta_1), \quad (5)$$

where  $\lambda_1 > 0, \lambda_2 > 0$  are tuning hyperparameters. Here,  $\mathcal{L}_{\text{ce}}(f_{\theta_2}(f_{\theta_1}(x)), y)$  and  $\mathcal{L}_{\text{ce}}(f_{\theta_1}(x), a)$  denote the cross-entropy losses for predicting  $y$  and  $a$ , respectively.

The alignment loss  $\mathcal{L}_{\text{align}}(\theta_2)$  over  $\theta_2$  is defined as:

$$\mathcal{L}_{\text{align}}(\theta_2) := \frac{1}{n} \sum_{i=1}^n \| (a_i^*(\theta_2) - \hat{a}_i) \odot (1 - r_i) \|_{\ell_1},$$

where  $\hat{a}_i := f_{\theta_1}(x_i)$  and the counterfactual attributes  $a^*(\theta_2)$  is generated via:

$$a^*(\theta_2) = \arg \min_{a'} \mathcal{L}_{\text{ce}}(f_{\theta_2}(a'), y^*) + \alpha_2 d(a', \hat{a}). \quad (6)$$

Similarly, the alignment loss  $\mathcal{L}_{\text{align}}(\theta_1)$  over  $\theta_1$  is defined by Eq. (3), where the counterfactual image  $x^*(\theta_1)$  that explains the change of  $\hat{a}$  to  $a^*$  is generated by:

$$x^*(\theta_1) = \arg \min_{x'} \mathcal{L}_{\text{ce}}(f_{\theta_1}(x'), a^*) + \alpha_1 d(x', x). \quad (7)$$

With the objective Eq. (5), we optimize  $\theta$  by applying Alg. 1 to alignment terms. After the optimization, our decision process  $x \rightarrow f_{\theta_2}(f_{\theta_1}(x))$  aligns well with that of the experts, with  $f_{\theta_1}$  employing causal imaging factors to predict attributes, and  $f_{\theta_2}$  using the causal attributes to predict  $y$ .

## 5 EXPERIMENT

In this section, we evaluate our method on two medical diagnosis tasks: the benign/malignant classification of lung nodules and breast masses<sup>5</sup>.

### 5.1 EXPERIMENTAL SETUPS

**Datasets & Preprocessing.** We consider the LIDC-IDRI dataset Armato III et al. (2011) for lung nodule classification and the CBIS-DDSM dataset Lee et al. (2017) for breast mass classification.

The LIDC-IDRI dataset contains thoracic CT images, each associated with bounding boxes indicating the nodule areas, six radiologist-annotated attributes (subtlety, calcification, margin, speculation, lobulation, and texture) and a malignancy score ranging from 1 to 5. Before analysis, we preprocess the images by resampling the pixel space and normalizing the intensity. We label those images with malignancy scores of 1-3 as benign ( $y = 0$ ) and those with scores of 4-5 as malignant ( $y = 1$ ). We split the dataset into training ( $n = 731$ ), validation ( $n = 238$ ), and test ( $n = 244$ ) sets. The CBIS-DDSM dataset contains breast mammography images with fine-grained annotations (mass bounding boxes, attributes, and malignancy). We preprocess the images by removing the background and normalizing the intensity. We use the provided binary malignancy label and six annotated attributes (subtlety, shape, circumscription, obscuration, ill-definiteness, and spiculation). We follow the official dataset split, with 691 masses in the training set and 200 masses in the test set.

To test the ability of our method to learn expert-aligned features, we add a “+”/“-” symbol on the top-left corner of each image as a spuriously correlated feature. This symbol coincides with the malignancy label in the training set, where images with  $y = 1$  are labeled with “+” and those with  $y = 0$  are labeled with “-”; but are assigned randomly in the validation and test sets. A well-aligned model should focus on the radiologist-annotated areas rather than the symbol.

<sup>5</sup>We provide results on additional diagnosis tasks and data modalities in Appx. C.1.

**Evaluation Metrics.** To assess the alignment of our model relative to radiologists, we compute the Class Activation Mapping (CAM) Selvaraju et al. (2017) and report its precision relative to the annotated areas, *i.e.*,  $\frac{\text{Area of } (\text{CAM} \cap \text{Anno})}{\text{Area of CAM}}$ . We also report the overall classification accuracy.

**Implementation Details.** We use the Adam optimizer and set the learning rate as 0.001. We parameterize the attributes prediction network  $f_{\theta_1}$  with a seven-layer Convolutional Neural Network (CNN), and train it for 100 epochs with a batch size of 128 for each iteration. For the classification network  $f_{\theta_2}$ , we parameterize it with a two-layer Multi-Layer Perceptron (MLP), and train it for 30 epochs with a batch size of 128. Please refer to Appx. B for details of the network architectures. For the hyperparameters  $\alpha_1$  in Eq. (7) and  $\alpha_2$  in Eq. (6), we set them to  $\alpha_1 = 0.01, \alpha_2 = 0.0005$  for LIDC-IDRI and  $\alpha_1 = 0.07, \alpha_2 = 0.0005$  for CBIS-DDSM, respectively. For both datasets, we set  $\lambda_1 = \lambda_2 = 1$  in Eq. (5). **For causal attribution, we calculate the CCCE scores of subsets containing no more than three attributes and select the subset with the highest score.** We adopt the TorchOpt Ren et al. (2022) package to implement the conjugate gradient estimator. We repeat 3 different seeds to remove the effect of randomness.

## 5.2 COMPARISON WITH BASELINES

**Compared Baselines.** We compare our method with the following baselines: **i) Ross et al. (2017)** that achieved interpretability by penalizing the input gradient to be small in object-irrelevant areas; **ii) ICNN** Zhang et al. (2018) that modified traditional CNN with an interpretable convolution layer to enforce object-centered representations; **iii) BagNet** Brendel & Bethge (2019) that approximated CNN with white-box bag-of-features models; **iv) Rieger et al. (2020)** that required the model to produce a classification as well as an explanation (*i.e.*, multi-tasks learning); **v) Chang et al. (2021)** that augmented the dataset with various factual and counterfactual images to alleviate the problem of learning spurious features; and **vi) the Oracle classifier** in which we manually restrict the input features to areas annotated by radiologists.

Table 1: Comparison with baseline methods on LIDC-IDRI and CBIS-DDSM datasets. The result of our method is **boldfaced** and the best result among baseline methods is underlined. For the Oracle classifier, the input features are manually restricted to the areas annotated by radiologists.

Methodology	Precision of CAM		Classification accuracy	
	LIDC	DDSM	LIDC	DDSM
Ross et al. (2017)	0.034 (0.06)	0.084 (0.11)	<u>0.656 (0.00)</u>	0.559 (0.05)
Zhang et al. (2018)	0.068 (0.11)	0.110 (0.13)	<u>0.381 (0.03)</u>	0.581 (0.00)
Brendel & Bethge (2019)	0.048 (0.04)	0.090 (0.04)	0.358 (0.00)	<u>0.592 (0.00)</u>
Rieger et al. (2020)	0.041 (0.05)	<u>0.232 (0.17)</u>	0.343 (0.00)	0.586 (0.01)
Chang et al. (2021)	<u>0.074 (0.03)</u>	0.119 (0.07)	0.503 (0.08)	0.496 (0.08)
Oracle classifier	1.000 (0.00)	1.000 (0.00)	0.789 (0.00)	0.726 (0.01)
Ours	<b>0.751 (0.03)</b>	<b>0.805 (0.06)</b>	<b>0.722 (0.00)</b>	<b>0.656 (0.00)</b>

**Results & Analysis.** Tab. 1 reports the alignment precision and classification accuracy. As shown, our method demonstrates strong alignment with radiologists in the diagnostic process. This result verifies the utility of our alignment loss and the optimization process. In contrast, other methods with no alignment may learn unreliable features that are beyond the annotated areas, which deteriorates their alignment accuracy. Specifically, one should note that ICNN Zhang et al. (2018), BagNet Brendel & Bethge (2019), and Rieger et al. (2020) imposed no explicit constraint for learning explainable features. As a result, these methods can be easily biased by background features or pseudo-correlations in the image. Meanwhile, although gradient methods such as Ross et al. (2017) and Chang et al. (2021) explicitly constrained the input gradient to human decision areas, **the attention mechanism in their approaches only learn features that are correlated with, rather than causally linked to the disease label Grimsley et al. (2020).** As a result, these methods may capture spurious features outside the causal regions.

Due to the capability of capturing causal features, our method also significantly surpasses baseline models in terms of classification accuracy. This is due to the fact that, unlike the “+”/“-” symbol that demonstrates only spurious correlation to the label, features within the annotated areas have a causal relationship with the label, and therefore are transferable to test data.

432 Additionally, it is worth noting from Tab. 1 that even the oracle classifier only reaches a classification  
 433 accuracy of 72% - 79%, which seems to contradict some previous results Wu et al. (2018); Wang  
 434 et al. (2022) that claimed an accuracy of more than 99% in lung nodule classification and 90% in  
 435 breast mass classification. To comprehend, this discrepancy is primarily due to the exclusion of  
 436 challenging samples (those with a malignancy score of 3) in Wu et al. (2018), and the usage of  
 437 custom training/test sets split in Wang et al. (2022).

### 439 5.3 ABLATION STUDY

440 In this section, we perform an ablation study on the causal alignment loss (Sect. 4.1) and the hierar-  
 441 chical alignment process (Sect.4.3). The results are shown in Tab. 2.

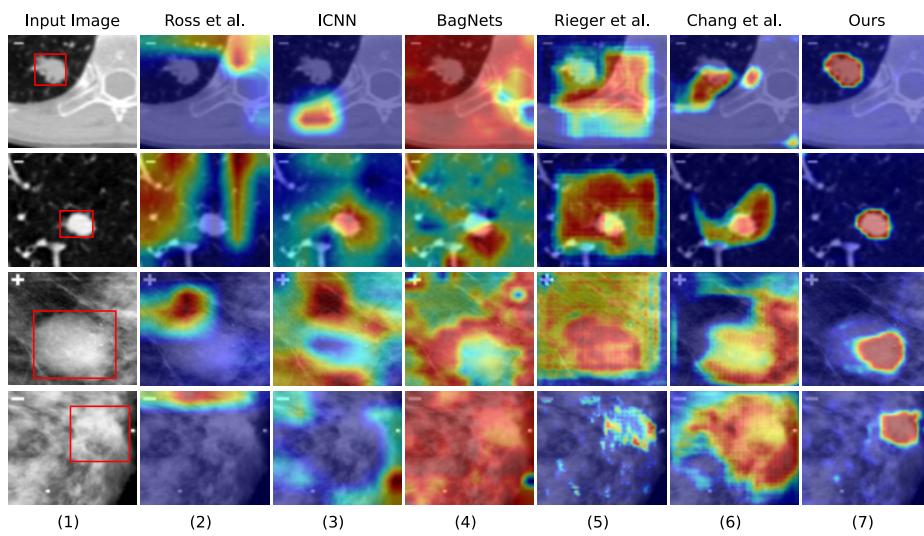
443 Table 2: Ablation study on LIDC-IDRI and CBIS-DDSM datasets.

$\mathcal{L}_{\text{align}}$	Hierarchical align	Precision of CAM		Classification accuracy	
		LIDC	DDSM	LIDC	DDSM
✗	✗	0.057 (0.07)	0.143 (0.20)	0.535 (0.08)	0.592 (0.00)
✓	✗	0.587 (0.08)	0.621 (0.03)	0.701 (0.02)	0.633 (0.03)
✓	✓	<b>0.751 (0.03)</b>	<b>0.805 (0.06)</b>	<b>0.722 (0.00)</b>	<b>0.656 (0.00)</b>

450 As we can see, both the alignment loss and the hierarchical procedure significantly improve the  
 451 performance. In detail, the alignment loss accounts for a substantial portion of the improvement,  
 452 yielding a 50% increase in CAM precision and a 15% boost in classification accuracy. Additionally,  
 453 the hierarchical training strategy contributes an extra 20% to alignment precision and a 2% increase  
 454 in classification performance. These results demonstrate the effectiveness of our alignment loss  
 455 in learning features that coincide with radiologist assessments, as well as the significance of the  
 456 hierarchical training strategy in mimicking the clinical diagnosis process.

### 458 5.4 VISUALIZATION

460 To further verify whether our method can learn radiologist-aligned features, we visualize the Class  
 461 Activation Mapping (CAM) and show the results in Fig. 4. The first two rows of Fig. 4 present cases  
 462 with lung nodules while the third and the fourth rows present cases with breast masses. The first  
 463 column shows the input image, with the nodule/mass areas marked by red bounding boxes, while  
 464 other columns present CAMs of various models.



483 Figure 4: CAM visualization. Each row denotes different cases. The first column is the input images,  
 484 where nodules and masses are marked by red bounding boxes. The second to seventh columns are  
 485 CAMs of compared baselines and our method, respectively. See Appx. D.2 for more results.

486 As shown, the activation of our method concentrates on the nodule/mass areas, especially on the  
 487 margins of the nodules/mass, which is a key feature for radiologists to evaluate the malignancy  
 488 Sandler et al. (2023). In contrast, the activation of baseline methods focuses on lesion-irrelevant  
 489 areas, such as the shortcut symbol “+”/“-” region in the top-left corner for Ross et al. (2017) and  
 490 Brendel & Bethge (2019), or the background areas for Zhang et al. (2018), Rieger et al. (2020), and  
 491 Chang et al. (2021). This visual analysis corroborates the quantitative results, demonstrating our  
 492 method’s ability to learn features that are well-aligned with the radiologist’s diagnostic process.  
 493

## 494 6 CONCLUSION AND DISCUSSION

495 In this paper, we present a causal alignment framework to bridge the gap between the decision-  
 496 making process of machine learning algorithms and experienced radiologists. By identifying the  
 497 causal features that influence the model’s decision, we can enforce the alignment of these causal  
 498 areas with those of the radiologists through a causal alignment loss. This further allows us to train a  
 499 hierarchical decision model that closely mirrors the expert’s decision pipeline. The effectiveness of  
 500 our approach is demonstrated by improved alignment in lung cancer and breast cancer diagnosis.  
 501

502 **Limitation and Future Works.** The optimization of our causal alignment loss can be computa-  
 503 tionally expensive due to the estimation of the implicit Jacobian matrix. We will investigate efficient  
 504 linear equation solving techniques Mou et al. (2016) to address this challenge. Additionally, we plan  
 505 to apply our loss to alignment learning in multi-modality models and robotic systems.

## 506 507 ETHICAL STATEMENT

508 The datasets for this study were collected under ethical guidelines, including participant consent  
 509 and anonymization. This paper introduces a causal alignment loss to align AI decision-making with  
 510 that of experienced radiologists, enhancing the credibility of AI-assisted diagnosis. Additionally,  
 511 the counterfactual generation and causal attribution methods identify the causal factors behind ra-  
 512 diologists’ diagnoses, improving our understanding of the diagnostic process and facilitating the  
 513 development of trustworthy medical AI. Nonetheless, we emphasize that our method serves as an  
 514 AI assistance tool for radiologists and should only be used under their supervision.

## 516 517 REFERENCES

- 518 Samuel G Armato III, Geoffrey McLennan, Luc Bidaut, Michael F McNitt-Gray, Charles R Meyer,  
 519 Anthony P Reeves, Binsheng Zhao, Denise R Aberle, Claudia I Henschke, Eric A Hoffman,  
 520 et al. The lung image database consortium (lidc) and image database resource initiative (idri):  
 521 a completed reference database of lung nodules on ct scans (available under the tcia data usage  
 522 policy and restrictions). *Medical physics*, 38(2):915–931, 2011.
- 523 Maximilian Augustin, Yannic Neuhaus, and Matthias Hein. Dig-in: Diffusion guidance for inves-  
 524 tigating networks-uncovering classifier differences neuron visualisations and visual counterfac-  
 525 tual explanations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern*  
 526 *Recognition*, pp. 11093–11103, 2024.
- 527 Rachana Balasubramanian, Samuel Sharpe, Brian Barr, Jason Wittenbach, and C Bayan  
 528 Bruss. Latent-cf: a simple baseline for reverse counterfactual explanations. *arXiv preprint*  
 529 *arXiv:2012.09301*, 2020.
- 530 Alina Jade Barnett, Fides Regina Schwartz, Chaofan Tao, Chaofan Chen, Yiniao Ren, Joseph Y Lo,  
 531 and Cynthia Rudin. Interpretable mammographic image classification using case-based reasoning  
 532 and deep learning. *arXiv preprint arXiv:2107.05605*, 2021.
- 533 Jack Brady, Roland S Zimmermann, Yash Sharma, Bernhard Schölkopf, Julius Von Kügelgen, and  
 534 Wieland Brendel. Provably learning object-centric representations. In *International Conference*  
 535 *on Machine Learning*, pp. 3038–3062. PMLR, 2023.
- 536 Wieland Brendel and Matthias Bethge. Approximating cnns with bag-of-local-features models  
 537 works surprisingly well on imagenet. In *International Conference on Learning Representations*,  
 538 2019.

- 540 Chun-Hao Chang, George Alexandru Adam, and Anna Goldenberg. Towards robust classification  
 541 model by counterfactual and invariant data generation. In *Proceedings of the IEEE/CVF Conference*  
 542 on Computer Vision and Pattern Recognition, pp. 15212–15221, 2021.
- 543
- 544 Chaofan Chen, Oscar Li, Daniel Tao, Alina Barnett, Cynthia Rudin, and Jonathan K Su. This looks  
 545 like that: deep learning for interpretable image recognition. *Advances in neural information*  
 546 *processing systems*, 32, 2019.
- 547 Dominik Csiba and Peter Richtárik. Global convergence of arbitrary-block gradient methods for  
 548 generalized polyak- $\{\backslash L\}$  ojasiewicz functions. *arXiv preprint arXiv:1709.03014*, 2017.
- 549
- 550 Amit Dhurandhar, Pin-Yu Chen, Ronny Luss, Chun-Chen Tu, Paishun Ting, Karthikeyan Shan-  
 551 mugam, and Payel Das. Explanations based on the missing: Towards contrastive explanations  
 552 with pertinent negatives. *Advances in neural information processing systems*, 31, 2018.
- 553 Simon Du, Jason Lee, Haochuan Li, Liwei Wang, and Xiyu Zhai. Gradient descent finds global  
 554 minima of deep neural networks. In *International conference on machine learning*, pp. 1675–  
 555 1685. PMLR, 2019.
- 556
- 557 Timo Freiesleben. The intriguing relation between counterfactual explanations and adversarial ex-  
 558 amples. *Minds and Machines*, 32(1):77–109, 2022.
- 559 Luisa Gallée, Catharina Silvia Lisson, Christoph Gerhard Lisson, Daniela Drees, Felix Weig, Daniel  
 560 Vogelee, Meinrad Beer, and Michael Götz. Evaluating the explainability of attributes and proto-  
 561 types for a medical classification model. In *World Conference on Explainable Artificial Intelli-  
 562 gence*, pp. 43–56. Springer, 2024.
- 563
- 564 Yuyang Gao, Tong Steven Sun, Liang Zhao, and Sungsoo Ray Hong. Aligning eyes between humans  
 565 and deep neural network through interactive attention alignment. *Proceedings of the ACM on*  
*566 Human-Computer Interaction*, 6(CSCW2):1–28, 2022.
- 567
- 568 Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel,  
 569 Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nature*  
*Machine Intelligence*, 2(11):665–673, 2020.
- 570
- 571 Yash Goyal, Ziyan Wu, Jan Ernst, Dhruv Batra, Devi Parikh, and Stefan Lee. Counterfactual visual  
 572 explanations. In *International Conference on Machine Learning*, pp. 2376–2384. PMLR, 2019.
- 573
- 574 Christopher Grimsley, Elijah Mayfield, and Julia Bursten. Why attention is not explanation: Surgical  
 575 intervention and causal reasoning about neural models. 2020.
- 576
- 577 Riccardo Guidotti. Counterfactual explanations and how to find them: literature review and bench-  
 578 marking. *Data Mining and Knowledge Discovery*, 38(5):2770–2824, 2024.
- 579
- 580 Victor Guyomard, Françoise Fessant, Thomas Guyet, Tassadit Bouadi, and Alexandre Termier. Gen-  
 581 erating robust counterfactual explanations. In *Joint European Conference on Machine Learning*  
 582 and *Knowledge Discovery in Databases*, pp. 394–409. Springer, 2023.
- 583
- 584 Benjamin D Haeffele and René Vidal. Global optimality in neural network training. In *Proceedings*  
 585 of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7331–7339, 2017.
- 586
- 587 Michael Hind, Dennis Wei, Murray Campbell, Noel CF Codella, Amit Dhurandhar, Aleksandra  
 588 Mojsilović, Karthikeyan Natesan Ramamurthy, and Kush R Varshney. Ted: Teaching ai to explain  
 589 its decisions. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pp.  
 590 123–129, 2019.
- 591
- 592 Borja Ibarz, Jan Leike, Tobias Pohlen, Geoffrey Irving, Shane Legg, and Dario Amodei. Reward  
 593 learning from human preferences and demonstrations in atari. *Advances in neural information*  
 594 *processing systems*, 31, 2018.
- 595
- 596 Kenji Kawaguchi and Jiaoyang Huang. Gradient descent finds global minima for generalizable deep  
 597 neural networks of practical sizes. In *2019 57th Annual Allerton Conference on Communication,*  
*598 Control, and Computing (Allerton)*, pp. 92–99. IEEE, 2019.

- 594 Steven George Krantz and Harold R Parks. *The implicit function theorem: history, theory, and*  
 595 *applications*. Springer Science & Business Media, 2002.
- 596
- 597 Rebecca Sawyer Lee, Francisco Gimenez, Assaf Hoogi, Kanae Kawai Miyake, Mia Gorovoy, and  
 598 Daniel L Rubin. A curated mammography data set for use in computer-aided detection and  
 599 diagnosis research (available under the tcia data usage policy and restrictions). *Scientific data*,  
 600 4(1):1–9, 2017.
- 601 Zachary C Lipton. The doctor just won’t accept that! *arXiv preprint arXiv:1711.08037*, 2017.
- 602
- 603 Yuhang Liu, Fandong Zhang, Chaoqi Chen, Siwen Wang, Yizhou Wang, and Yizhou Yu. Act like  
 604 a radiologist: towards reliable multi-view correspondence reasoning for mammogram mass de-  
 605 tection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):5947–5961,  
 606 2021.
- 607
- 608 James Martens et al. Deep learning via hessian-free optimization. In *Icml*, volume 27, pp. 735–742,  
 609 2010.
- 610 Michael F McNitt-Gray, Samuel G Armato III, Charles R Meyer, Anthony P Reeves, Geoffrey  
 611 McLennan, Richie C Pais, John Freymann, Matthew S Brown, Roger M Engelmann, Peyton H  
 612 Bland, et al. The lung image database consortium (lidc) data collection process for nodule detec-  
 613 tion and annotation. *Academic radiology*, 14(12):1464–1474, 2007.
- 614 Shaoshuai Mou, Zhiyun Lin, Lili Wang, Daniel Fullmer, and A Stephen Morse. A distributed  
 615 algorithm for efficiently solving linear equations and its applications (special issue jcw). *Systems  
 & Control Letters*, 91:21–27, 2016.
- 616
- 617 Philip Müller, Felix Meissen, Johannes Brandt, Georgios Kaassis, and Daniel Rueckert. Anatomy-  
 618 driven pathology detection on chest x-rays. In *International Conference on Medical Image Com-  
 puting and Computer-Assisted Intervention*, pp. 57–66. Springer, 2023.
- 619
- 620 Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong  
 621 Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to fol-  
 622 low instructions with human feedback. *Advances in neural information processing systems*, 35:  
 623 27730–27744, 2022.
- 624
- 625 Judea Pearl. Causal inference in statistics: An overview. 2009.
- 626
- 627 Judea Pearl. Structural counterfactuals: A brief introduction. *Cognitive science*, 37(6):977–985,  
 628 2013.
- 629
- 630 Boris T Polyak. Gradient methods for solving equations and inequalities. *USSR Computational  
 Mathematics and Mathematical Physics*, 4(6):17–32, 1964.
- 631
- 632 Jie Ren, Xidong Feng, Bo Liu, Xuehai Pan, Yao Fu, Luo Mai, and Yaodong Yang. Torchopt: An  
 633 efficient library for differentiable optimization (available under the apache license, version 2.0).  
*arXiv preprint arXiv:2211.06934*, 2022.
- 634
- 635 Laura Rieger, Chandan Singh, William Murdoch, and Bin Yu. Interpretations are useful: penaliz-  
 636 ing explanations to align neural networks with prior knowledge. In *International conference on  
 machine learning*, pp. 8116–8126. PMLR, 2020.
- 637
- 638 Andrew Slavin Ross, Michael C Hughes, and Finale Doshi-Velez. Right for the right reasons: Train-  
 639 ing differentiable models by constraining their explanations. *arXiv preprint arXiv:1703.03717*,  
 2017.
- 640
- 641 Kim L Sandler, Travis S Henry, Arya Amini, Saeed Elojeimy, Aine Marie Kelly, Christopher T  
 642 Kuzniewski, Elizabeth Lee, Maria D Martin, Michael F Morris, Neeraja B Peterson, et al. Acr  
 643 appropriateness criteria® lung cancer screening: 2022 update. *Journal of the American College  
 of Radiology*, 20(5):S94–S101, 2023.
- 644
- 645 Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh,  
 646 and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based local-  
 647 ization. In *Proceedings of the IEEE international conference on computer vision*, pp. 618–626,  
 2017.

- 648 Vincent Sitzmann, Eric Chan, Richard Tucker, Noah Snavely, and Gordon Wetzstein. Metasdf:  
 649 Meta-learning signed distance functions. *Advances in Neural Information Processing Systems*,  
 650 33:10136–10147, 2020.
- 651 Lili Song and Luís Nunes Vicente. Modeling hessian-vector products in nonlinear optimization:  
 652 new hessian-free methods. *IMA Journal of Numerical Analysis*, 42(2):1766–1788, 2022.
- 653 Ade Irma Suryani, Chuan-Wang Chang, Yu-Fan Feng, Tin-Kwang Lin, Chih-Wen Lin, Jen-Chieh  
 654 Cheng, and Chuan-Yu Chang. Lung tumor localization and visualization in chest x-ray images  
 655 using deep fusion network and class activation mapping. *IEEE Access*, 10:124448–124463, 2022.
- 656 Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow,  
 657 and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.
- 658 Sahil Verma, Varich Boonsanong, Minh Hoang, Keegan E Hines, John P Dickerson, and Chirag  
 659 Shah. Counterfactual explanations and algorithmic recourses for machine learning: A review.  
 660 *arXiv preprint arXiv:2010.10596*, 2020.
- 661 Nisheeth K Vishnoi et al. Lx= b. *Foundations and Trends® in Theoretical Computer Science*, 8  
 662 (1–2):1–141, 2013.
- 663 Sandra Wachter, Brent Mittelstadt, and Chris Russell. Counterfactual explanations without opening  
 664 the black box: Automated decisions and the gdpr. *Harv. JL & Tech.*, 31:841, 2017.
- 665 Chu-Ran Wang, Fei Gao, Fandong Zhang, Fangwei Zhong, Yizhou Yu, and Yizhou Wang. Dis-  
 666 entangling disease-related representation from obscure for disease prediction. In *International  
 667 Conference on Machine Learning*, pp. 22652–22664. PMLR, 2022.
- 668 Botong Wu, Zhen Zhou, Jianwei Wang, and Yizhou Wang. Joint learning for pulmonary nodule  
 669 segmentation, attributes and malignancy prediction. In *2018 IEEE 15th International Symposium  
 670 on Biomedical Imaging (ISBI 2018)*, pp. 1109–1113. IEEE, 2018.
- 671 Yao Xie, Melody Chen, David Kao, Ge Gao, and Xiang’Anthony’ Chen. Chexplain: enabling physi-  
 672 cians to explore and understand data-driven, ai-enabled medical imaging analysis. In *Proceedings  
 673 of the 2020 CHI Conference on Human Factors in Computing Systems*, pp. 1–13, 2020.
- 674 Kevin Kam Fung Yuen. A tutorial on explainable image classification for dementia stages using  
 675 convolutional neural network and gradient-weighted class activation mapping. *arXiv preprint  
 676 arXiv:2408.10572*, 2024.
- 677 Mehdi Zemni, Mickaël Chen, Éloi Zablocki, Hédi Ben-Younes, Patrick Pérez, and Matthieu Cord.  
 678 Octet: Object-aware counterfactual explanations. In *Proceedings of the IEEE/CVF conference on  
 679 computer vision and pattern recognition*, pp. 15062–15071, 2023.
- 680 Quanshi Zhang, Ying Nian Wu, and Song-Chun Zhu. Interpretable convolutional neural networks.  
 681 In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 8827–  
 682 8836, 2018.
- 683 Yifei Zhang, Siyi Gu, Yuyang Gao, Bo Pan, Xiaofeng Yang, and Liang Zhao. Magi: Multi-annotated  
 684 explanation-guided learning. In *Proceedings of the IEEE/CVF International Conference on Com-  
 685 puter Vision*, pp. 1977–1987, 2023.
- 686 Ruiqi Zhao, Lei Zhang, Shengyu Zhu, Zitong Lu, Zhenhua Dong, Chaoliang Zhang, Jun Xu, Zhi  
 687 Geng, and Yangbo He. Conditional counterfactual causal effect for individual attribution. In  
 688 *Uncertainty in Artificial Intelligence*, pp. 2519–2528. PMLR, 2023.
- 689 Simon Zhuang and Dylan Hadfield-Menell. Consequences of misaligned ai. *Advances in Neural  
 690 Information Processing Systems*, 33:15763–15773, 2020.
- 691

## APPENDIX

### A Causal Alignment Theory

702	<b>B Further Details on Implementation</b>	<b>15</b>
703		
704		
705	<b>C Extra Experimental Results</b>	<b>16</b>
706	C.1 Applicability to Different Data Modalities . . . . .	16
707	C.2 Insensitivity to Hyperparameters . . . . .	16
708	C.3 Results under Different Shortcut Symbols . . . . .	16
709	C.4 Impact of different distance functions . . . . .	17
710	C.5 Causal graph obtained via the PC algorithm . . . . .	18
711		
712		
713		
714	<b>D Visualization</b>	<b>19</b>
715	D.1 Visualization of Counterfactual Images . . . . .	19
716	D.2 Visualization of CAMs . . . . .	20
717		
718		
719	<b>A CAUSAL ALIGNMENT THEORY</b>	
720		
721	In this section, we discuss some theoretical aspects of causal alignment. We adopt the following	
722	notational convenience. Let $X$ and $Y$ denote the image and the predicted label, respectively. Let	
723	$Y_x$ denote the potential outcome of the predicted label under $X$ being $x$ . Let $\phi_{\theta,\zeta}$ be the generation	
724	process that maps from the original image $x_0$ and the label $y_0$ to the counterfactual image $x$ , which	
725	relies on the model parameter $\theta$ and random seed $\zeta$ .	
726	To entail the discussion, we require the following assumptions:	
727	<b>Assumption A.1</b> (Consistency). We assume that for each individual, the predicted label $Y$ when	
728	$X = x$ is exactly the potential outcome $Y_x$ .	
729	<b>Assumption A.2.</b> We assume Eq. (2) has a unique global minimum solution.	
730	<i>Remark A.3.</i> It can be shown that the global minimum of Eq. (2) can be attained via gradient descent	
731	under smoothness, and Polyak-Łojasiewicz conditions Csiba & Richtárik (2017); Polyak (1964).	
732	For deep learning optimization, the global minimum can be obtained if $f_\theta$ is over-parameterized Du	
733	et al. (2019) or has sufficient width Haeffele & Vidal (2017); Kawaguchi & Huang (2019).	
734		
735	We first show the probability of causation $P_\theta(Y_x = y X = x_0, Y = y_0)$ is identifiable.	
736	<b>Proposition A.4.</b> Assume Asms. A.2 and Asm. A.1, then the probability of causation is identifiable	
737	with	
738	$P_\theta(Y_x = y X = x_0, Y = y_0) = P_\theta(Y = y x)P_\theta(x x_0, y_0).$	
739		
740	<i>Proof.</i> Denote the counterfactual generator as $\phi_{\theta,\zeta}$ . If we fix the model parameter $\theta$ and the random	
741	seed $\zeta$ , $\phi_{\theta,\zeta}$ is a deterministic function, which means the conditional probability $P_\theta(x' x_0, y_0) =$	
742	$\mathbb{1}(x' = \phi_{\theta,\zeta}(x_0, y_0)) = \mathbb{1}(x' = x)$ for any $x'$ . In this regard, we have:	
743	$\begin{aligned} P_\theta(Y_x = y X = x_0, Y = y_0) &= \int P_\theta(Y_x = y x', x_0, y_0)P_\theta(x' x_0, y_0)dx' \\ &= P_\theta(Y_x = y x, x_0, y_0)P_\theta(x x_0, y_0). \end{aligned}$	
744		
745		
746	Further, under the fixed seed $\zeta$ , the potential outcome $Y_x$ is fully determined by the classifier $f_\theta$ and	
747	the counterfactual image $x$ :	
748	$Y_x = \text{sign}(f_\theta(x, u)),$	
749	where $u$ denotes the realization of the randomness $U$ in network prediction under the seed $\zeta$ . Therefore,	
750	we have $Y_x \perp\!\!\!\perp (X_0, Y_0) X = x$ and	
751	$\begin{aligned} P_\theta(Y_x = y X = x_0, Y = y_0) &= P_\theta(Y_x = y x, x_0, y_0)P_\theta(x x_0, y_0) \\ &= P_\theta(Y_x = y x)P_\theta(x x_0, y_0) \\ &\stackrel{(1)}{=} P_\theta(Y = y x)P_\theta(x x_0, y_0), \end{aligned}$	
752		
753		
754		
755		

where “(1)” is due to Asms. A.1. We then have the identifiability equation.  $\square$

Below, we show  $x^*$  in Eq. (2) maximizes the probability of causation, indicating that  $\text{supp}(x^* - x_0)$  represents the causal factors that determine the model’s decisions. Consequently, minimizing  $\mathcal{L}_{\text{align}}$  encourages the model’s causal factors to align more closely with those of the experts.

**Proposition A.5.** Assume Asm. A.2 and Asm. A.1, we then have:

$$x^* = \arg \max_{x:d(x,x_0) \leq d_\alpha} P_\theta(Y_x = y^* | X = x_0, Y = y_0)$$

for some  $d_\alpha$ .

We first show that Eq. (2) is equivalent to the following constrained optimization problem:

$$x^* = \arg \min_{x:d(x,x_0) \leq d_\alpha} \mathcal{L}_{\text{ce}}(f_\theta(x), y^*). \quad (8)$$

To this end, let  $d_\alpha := d(x^*, x_0)$  and let  $x^\circ := \arg \min_{x:d(x,x_0) \leq d_\alpha} \mathcal{L}_{\text{ce}}(f_\theta(x), y^*)$ , we show:

$$\mathcal{L}_{\text{ce}}(f_\theta(x^*), y^*) + \lambda d(x^*, x_0) = \mathcal{L}_{\text{ce}}(f_\theta(x^\circ), y^*) + \lambda d(x^\circ, x_0). \quad (9)$$

Since Asm. A.2 ensures the uniqueness of the minimum of Eq. (2), it then follows that  $x^* = x^\circ$  and Eq. (8) holds. Now, note that  $x^*$  satisfies  $d(x^*, x_0) \leq d_\alpha$ , which means:

$$\mathcal{L}_{\text{ce}}(f_\theta(x^*), y^*) \geq \mathcal{L}_{\text{ce}}(f_\theta(x^\circ), y^*).$$

Since  $x^\circ$  satisfies  $d(x^\circ, x_0) \leq d_\alpha = d(x^*, x_0)$ , we further have:

$$\mathcal{L}_{\text{ce}}(f_\theta(x^*), y^*) + \lambda d(x^*, x_0) \geq \mathcal{L}_{\text{ce}}(f_\theta(x^\circ), y^*) + \lambda d(x^\circ, x_0).$$

Since  $x^*$  minimizes Eq. (2), we also have:

$$\mathcal{L}_{\text{ce}}(f_\theta(x^*), y^*) + \lambda d(x^*, x_0) \leq \mathcal{L}_{\text{ce}}(f_\theta(x^\circ), y^*) + \lambda d(x^\circ, x_0).$$

Therefore, we have Eq. (9) holds.

We then show  $x^*$  maximize the probability of causation. From Eq. (8), we have:

$$x^* = \arg \max_{x:d(x,x_0) \leq d_\alpha} P_\theta(Y = y^* | x) P_\theta(x | x_0, y_0),$$

where the term  $P_\theta(x | x_0, y_0) = \mathbb{1}(x = \phi_{\theta, \zeta}(x_0, y_0))$  represents the generating process of  $x$ , and the term  $P_\theta(Y = y^* | x)$  represents maximizing the logarithm likelihood in the cross-entropy loss.

Then, according to the identification quantity of  $P_\theta(Y_x = y^* | X = x_0, Y = y_0)$  shown in Prop. A.4, we have:

$$x^* = \arg \max_{x:d(x,x_0) \leq d_\alpha} P_\theta(Y_x = y^* | X = x_0, Y = y_0).$$

This concludes the proof.  $\square$

## B FURTHER DETAILS ON IMPLEMENTATION

Below, we show the network architectures used in lung nodule classification (Fig. 5) and breast mass classification (Fig. 6).

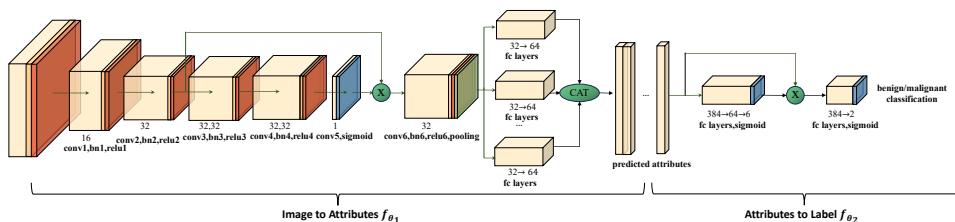


Figure 5: Network architecture used in lung nodule classification.

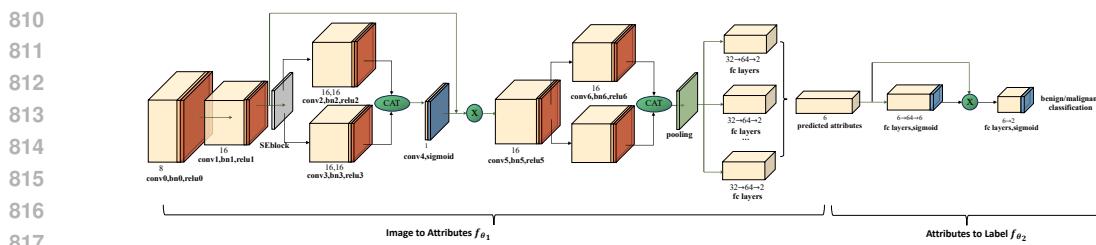


Figure 6: Network architecture used in breast mass classification.

## C EXTRA EXPERIMENTAL RESULTS

### C.1 APPLICABILITY TO DIFFERENT DATA MODALITIES

In this section, we demonstrate the applicability of our method to different data modalities. Specifically, we consider brain MRI data from the BraTS dataset, breast ultrasound data from the Aryashah2k dataset, lung CT data from the LIDC-IDRI dataset Armato III et al. (2011), and breast mammogram data from the CBIS-DDSM dataset Lee et al. (2017). The results are presented in Tab. 3, showing that our method is consistently accurate across various data types.

Table 3: Performance of our method and baselines on different data modalities. The result of our method is **boldfaced** and the best result among baselines is underlined.

Methodology	Precision of CAM				Classification Accuracy			
	MRI	Ultra.	CT	Mamm.	MRI	Ultra.	CT	Mamm.
Ross et al. (2017)	0.036	<b>0.197</b>	0.034	0.084	<b>0.730</b>	0.679	<b>0.656</b>	0.559
Zhang et al. (2018)	<b>0.168</b>	0.159	0.068	0.110	<b>0.698</b>	<b>0.764</b>	0.381	0.581
Brendel & Bethge (2019)	0.111	0.165	0.048	0.090	0.270	0.321	0.358	<u>0.592</u>
Rieger et al. (2020)	0.097	0.184	0.041	<u>0.232</u>	0.099	0.509	0.343	0.586
Chang et al. (2021)	0.147	0.127	<u>0.074</u>	0.119	0.410	0.270	0.503	0.496
Ours	<b>0.908</b>	<b>0.872</b>	<b>0.751</b>	<b>0.805</b>	<b>0.835</b>	<b>0.797</b>	<b>0.722</b>	<b>0.656</b>

### C.2 INSENSITIVITY TO HYPERPARAMETERS

In this section, we present the performance of our method across various hyperparameter configurations, as shown in Tab. 4 and 5. The results demonstrate that our method is robust to changes in hyperparameter settings, consistently achieving accurate alignment with the radiologists.

Table 4: Performance under different hyperparameters  $\alpha_1$ , which is the weight of the normalization term in counterfactual generation. The results are obtained from the CBIS-DDSM dataset.

$\alpha_1$	Precision of CAM	Classification Accuracy
0.05	0.819	0.650
0.06	0.801	0.648
0.07	0.805	0.656
0.08	0.833	0.642
0.09	0.796	0.655

### C.3 RESULTS UNDER DIFFERENT SHORTCUT SYMBOLS

Below, we show the performance of our method under various shortcut symbol settings. Specifically, we consider three cases: the +/- marker, intensity change, and the absence of a symbol. The results are presented in Tab. 6, showing that our method is effective across different shortcut symbols.

864  
 865 Table 5: Performance under different hyperparameters  $\lambda$ , which is the weight of the alignment loss  
 866 in the total loss. The results are obtained from the CBIS-DDSM dataset.

$\lambda$	Precision of CAM	Classification Accuracy
0.8	0.793	0.650
0.9	0.776	0.637
1.0	0.805	0.656
1.1	0.818	0.649
1.2	0.826	0.642

873  
 874 Table 6: Performance under different shortcut symbols.  
 875

Symbol	Precision of CAM		Classification Accuracy	
	LIDC	DDSM	LIDC	DDSM
None	0.783	0.882	0.707	0.652
Intensity	0.760	0.783	0.723	0.670
+/-	0.751	0.805	0.722	0.656

#### C.4 IMPACT OF DIFFERENT DISTANCE FUNCTIONS

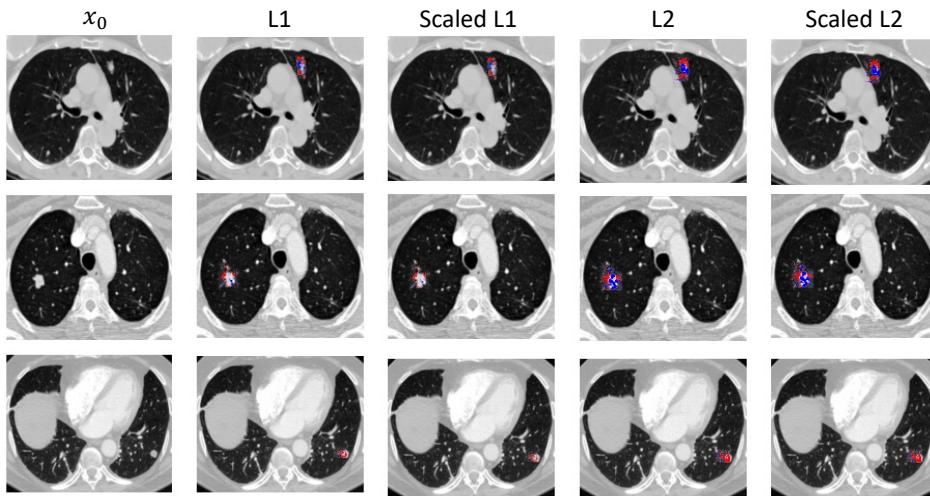
In the following, we conduct experiments to study the impact of different distance functions Wachter et al. (2017) on counterfactual generation, with results presented in Fig. 7. Here, the scaled  $\ell_1$  norm is defined as:

$$d(x_i, x_i^*) := \sum_{k=1}^{\dim(x_i)} \frac{|x_{i,k} - x_{i,k}^*|}{\text{MAD}_k}, \quad \text{where } \text{MAD}_k := \text{median}_i(|x_{i,k} - \text{median}_j(x_{j,k})|)$$

and the scaled  $\ell_2$  norm is defined as:

$$d(x_i, x_i^*) := \sum_{k=1}^{\dim(x_i)} \frac{|x_{i,k} - x_{i,k}^*|^2}{\text{std}_k}$$

where  $\text{std}_k$  is the standard deviation of the feature  $k$  among all samples.



914  
 915 Figure 7: Generated counterfactual images using different distance functions.  
 916

917 As shown, the  $\ell_1$  norm encourages sparse modifications in critical features for malignancy assessment, such as spiculation and margin. In contrast, the  $\ell_2$  norm tends to produce uniform changes

918 across the whole nodule. This observation is consistent with the findings of Wachter et al. (2017).  
 919 Moreover, we notice that the performance of  $\ell_1/\ell_2$  norms and their scaled versions are similar, which  
 920 can be attributed to the fact that we have already normalized the pixel values before training.  
 921

922 We then show the performance of our method under different distance functions in Tab. 8. The  
 923 results indicate that the  $\ell_1$  norm outperforms the  $\ell_2$  norm, which can be attributed to the sparser  
 924 modifications made by the  $\ell_1$  norm, facilitating more accurate localization of causal decision areas.  
 925

926 Table 7: Ablation study of different distance functions on alignment and classification accuracy.

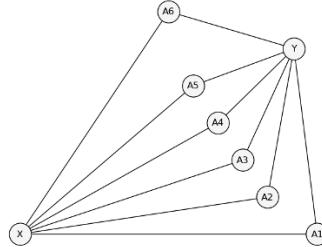
927 Distance functions	Precision of CAM	Classification accuracy
928 $\ell_1$	0.751	0.722
929 Scaled $\ell_1$	0.714	0.702
930 $\ell_2$	0.646	0.681
931 Scaled $\ell_2$	0.640	0.658

934 Table 8: Average number of pixels modified more than  $10^{-3}$  under the  $\ell_1$  and  $\ell_2$  norms

935	$\ell_1$ norm	$\ell_2$ norm
936	106.3	1146.5
937		
938		

### 939 C.5 CAUSAL GRAPH OBTAINED VIA THE PC ALGORITHM

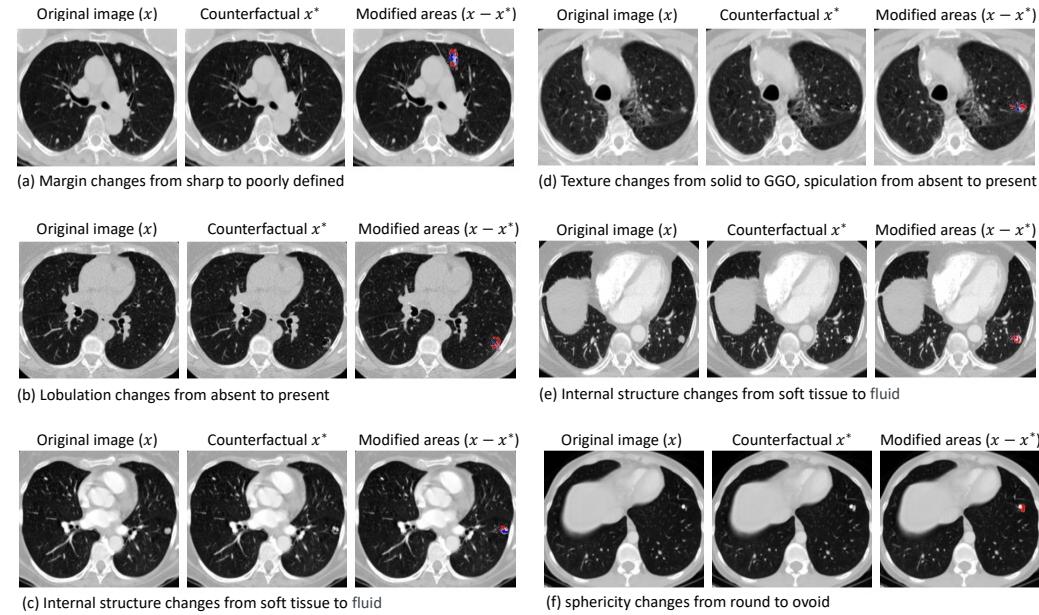
940 We also try the PC algorithm to recover the causal graph from data (see Fig. 8) under the Markov  
 941 and faithfulness assumptions. We find the skeleton of the recovered graph is consistent with that of  
 942 Fig. 3.

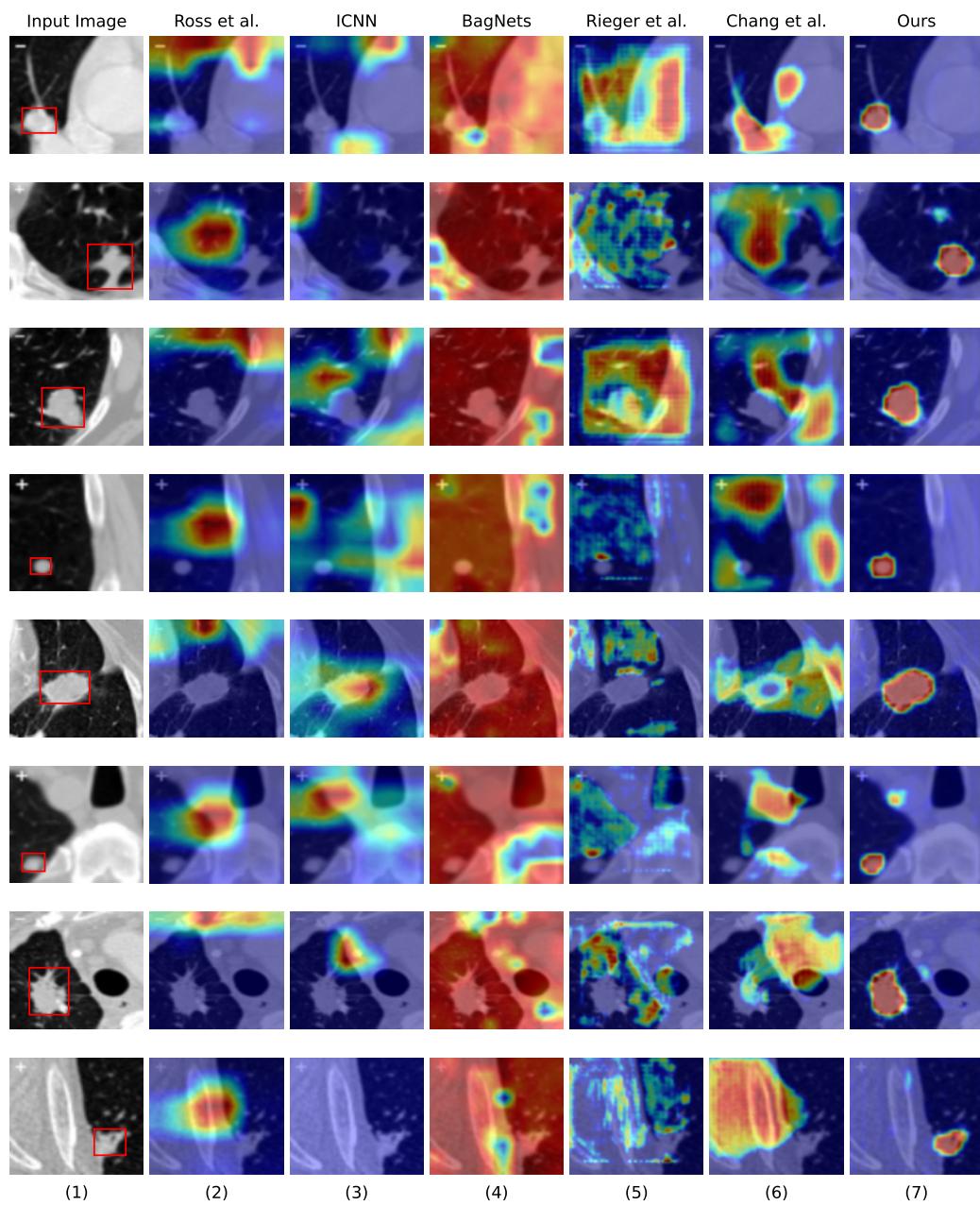


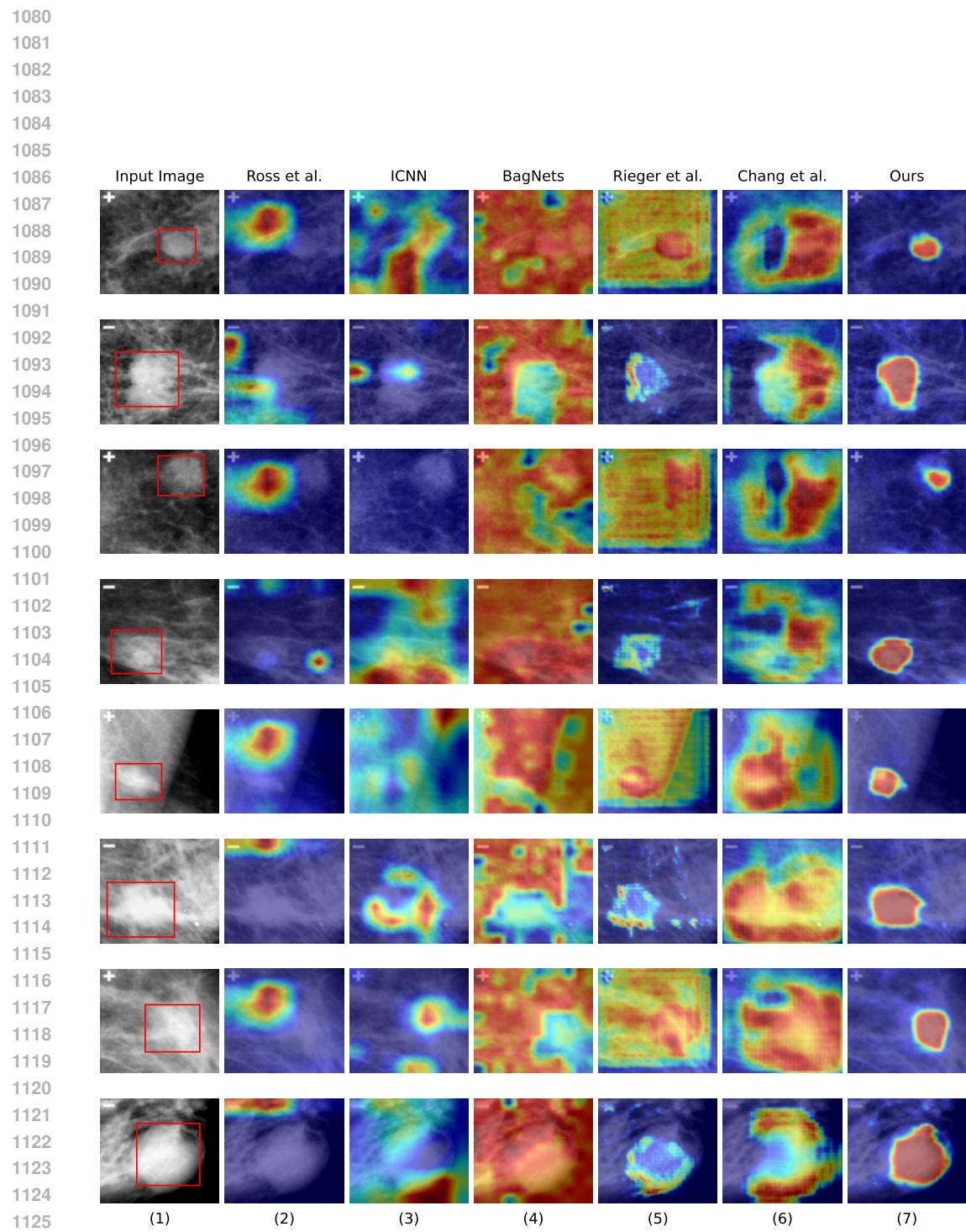
955 Figure 8: Recovered causal graph over nodule features ( $X$ ), attributes ( $A$ ), and label ( $Y$ ).  
 956

972 **D VISUALIZATION**  
973

974 **D.1 VISUALIZATION OF COUNTERFACTUAL IMAGES**  
975

976 In this section, we visualize the generated counterfactual images and show the result in Fig. 9. As  
977 we can see, the counterfactual modifications are clearly perceptible and align with specific clinical  
978 concepts, thereby validating the effectiveness of our counterfactual generation method.  
979

1001 Figure 9: Generated counterfactual images on the LIDC-IDRI dataset. For each sub-figure, the  
1002 left, middle, and right images denote the original image  $x$ , the counterfactual image  $x^*$ , and the  
1003 modified area  $\text{supp}(x - x^*)$ , respectively. Positive modifications are marked in red and negative  
1004 ones are marked in blue. We can observe that the counterfactual modifications all correspond to  
1005 certain clinical attributes of the nodule, for example, in (a), the margin attribute changes from sharp  
1006 to poorly defined when the label  $y$  changes from benign to malignant.  
1007  
1008  
1009  
1010  
1011  
1012  
1013  
1014  
1015  
1016  
1017  
1018  
1019  
1020  
1021  
1022  
1023  
1024  
1025

1026 D.2 VISUALIZATION OF CAMS  
10271028 In this section, we provide more visualizations of the CAMs.  
10291072     Figure 10: CAM visualization on the LIDC-IDRI dataset.  
1073  
1074  
1075  
1076  
1077  
1078  
1079

1128 Figure 11: CAM visualization on the CBID-DDSM dataset.  
1129  
1130  
1131  
1132  
1133