MELBA Journal Machine Learning for Biomedical Imaging

Attri-Net: A Globally and Locally Inherently Interpretable Model for Multi-Label Classification Using Class-Specific Counterfactuals

Susu **Sun** [1], Stefano **Woerner** [1], Andreas **Maier** [2], Lisa M. **Koch** [3,4], Christian F. **Baumgartner** [1,5]

1 Cluster of Excellence: Machine Learning - New Perspectives for Science, University of Tübingen, Tübingen, Germany
2 Department of Computer Science, Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen, Germany
3 Hertie Institute for Artificial Intelligence in Brain Health, University of Tübingen, Tübingen, Germany
4 Department of Diabetes, Endocrinology, Nutritional Medicine and Metabolism, Inselspital, Bern University Hospital, University of Bern, Bern, Switzerland
5 Faculty of Health Sciences and Medicine, University of Lucerne, Lucerne, Switzerland

Abstract

Interpretability is crucial for machine learning algorithms in high-stakes medical applications. However, high-performing neural networks typically cannot explain their predictions. Post-hoc explanation methods provide a way to understand neural networks but have been shown to suffer from conceptual problems. Moreover, current research largely focuses on providing local explanations for individual samples rather than global explanations for the model itself. In this paper, we propose Attri-Net, an inherently interpretable model for multi-label classification that provides both local and global explanations. Attri-Net first counterfactually generates class-specific attribution maps to highlight the disease evidence, then performs classification with logistic regression classifiers based solely on the attribution maps. Local explanations for each prediction can be obtained by interpreting the attribution maps weighted by the classifiers' weights. Global explanation of whole model can be obtained by jointly considering learned average representations of the attribution maps for each class (called the class centers) and the weights of the linear classifiers. To ensure the model is "right for the right reason", we introduce a mechanism to guide the model's explanations to align with human knowledge. Our comprehensive evaluations show that Attri-Net can generate high-quality explanations consistent with clinical knowledge while not sacrificing classification performance. Our code is available at `https://github.com/ss-sun/Attri-Net-V2`.

Keywords

Explainable machine learning, Inherently interpretable model, Multi-label classification, Model guidance

Article informations

`https://doi.org/10.59275/j.melba.2025-gb33`
Volume 3, Received: 2025-01-06, Published 2025-10-20
Corresponding author: susu.sun@uni-tuebingen.de

1. Introduction

Deep neural networks have significantly improved the performance of various medical image analysis tasks in experimental settings (Litjens et al., 2017). However, the black-box nature of deep learning models may lead to a lack of trust (Dietvorst et al., 2015), or more concerningly, blind trust among clinicians (Tschandl et al., 2020; Gaube et al., 2021). Using black-box models may potentially also result in ethical and legal problems (Grote and Berens, 2020), thus hindering their clinical adoption.

Interpretability has been identified as a crucial property for deploying machine learning technology in high-stakes applications such as medicine (Rudin, 2019). The majority of explanation techniques fall into the category of *post-hoc* methods, which are model-agnostic and can generate explanations for any pre-trained model. However, post-hoc methods are not guaranteed to be faithful to the model's true decision mechanism (Adebayo et al., 2018; Arun et al., 2021). Furthermore, as we show in this paper, current post-hoc explanation techniques that were developed for multi-class natural image classification do not perform adequately in the multi-label scenario, in which multiple medical findings may co-occur in a single image.

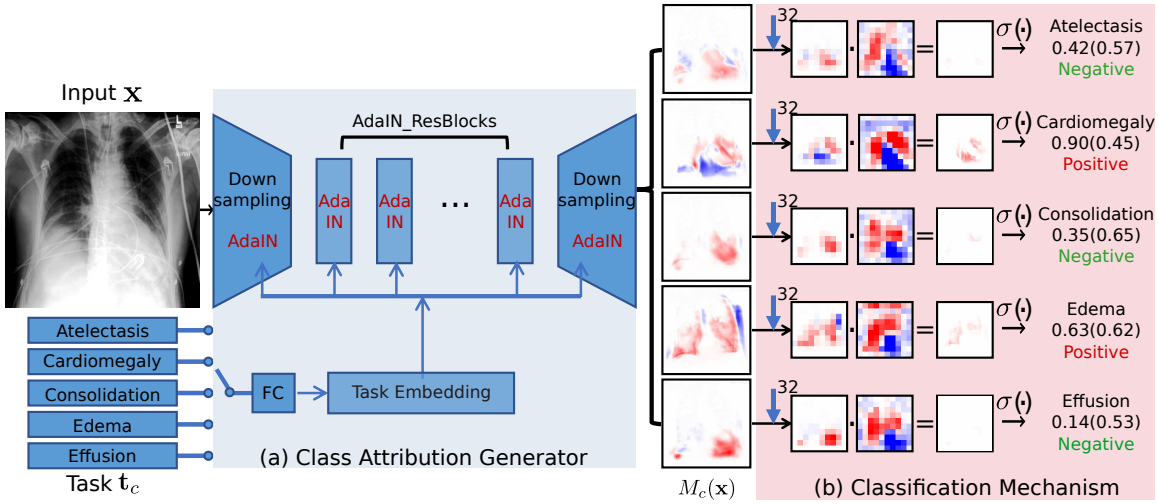


Figure 1: Overview of the Attri-Net framework. Given an input image **x** and a diagnostic task $\mathbf{t_c}$, the visual feature attribution generator **(a)** produces counterfactual attribution maps $M_c(\mathbf{x})$ that highlight specific disease effects. Logistic regression classifiers in **(b)** produce the final prediction for each class based on downsampled versions of these attribution maps.

An alternative way for approaching interpretability is to design inherently interpretable models, where the explanations are directly built into the model architecture. Recent works such as prototype-based neural networks (Chen et al., 2019), concept bottleneck models (Koh et al., 2020) or B-cos Networks (Böhle et al., 2024) can provide explanations for their predictions by revealing the actual decision process to the user. Thus, these explanations are considered to be faithful to the model's internal mechanism.

Based on the scope, the explanations can also be categorized as local or global explanations. Local explanations focus on understanding specific predictions for individual samples, while the global explanations aim to understand the overall decision mechanism of the entire model on the dataset level (Christoph, 2020; Bassan et al., 2024). Although prior research has primarily focused on local explanations (Bassan et al., 2024), for medical applications, it is also important to understand the behavior of machine learning models on a global level. Global explanations offer a valuable tool for verifying learned features and identifying potential spurious correlations, such as the model's reliance on task-irrelevant information. This is especially important since deep neural networks have a tendency to exploit shortcuts rather than focusing on task-relevant features (Geirhos et al., 2020). This shortcut learning behavior, as highlighted in (Sagawa et al., 2019), can significantly undermine the model's robustness and generalization capabilities.

To encourage deep neural networks to focus on task-relevant features for making predictions and increase the generalization of the model, recent research (Erion et al., 2021; Pillai et al., 2022; Rao et al., 2023; Li et al., 2018) has explored diverse methods of guiding models, such as enforcing desirable properties on the attributions or aligning explanations with human annotation. Given that many explanation methods are differentiable (Selvaraju et al., 2017; Böhle et al., 2024; Shrikumar et al., 2017), model guidance can be directly integrated with these explanations, allowing optimization for both classification performance and feature localization.

In this work, we propose Attri-Net, an inherently interpretable model designed for the multi-label classification scenario. The key contributions of this work are as follows: Attri-Net provides faithful local explanations for individual predictions and global explanations that reveal the entire model's behavior at the dataset level. We incorporate a model guidance mechanism into Attri-Net to encourage the model to be "right for the right reason", relying on minimal pixel-wise disease annotations, which is typical given the scarcity of expert annotations in the medical domain. Quantitative and qualitative evaluations show that Attri-Net generates high-quality local explanations while retaining classification performance comparable to state-of-the-art models. Furthermore, we demonstrate that Attri-Net's global explanation can effectively identify the spurious correlation learned by the model in short-cut learning settings, and the proposed model guidance mechanism can successfully mitigate this undesired behavior.

This work extends our previous conference paper (Sun et al., 2023b) by introducing the global explanation that captures the model's overall behaviors at the dataset level. We design new experiments to demonstrate its effectiveness in identifying shortcut learning tendencies. Furthermore, we incorporate a model guidance mechanism that leverages human knowledge to guide the model during the training process, encouraging the model to make decisions that are "right for the right reason". New experiments show that

this model guidance mechanism can effectively mitigate shortcut learning behavior. In addition, we expand the literature review and provide a more in-depth discussion of the limitations of our current approach and directions for future work.

2. Related Works

2.1 Post-hoc explanation methods

Post-hoc explanation methods are model agnostic and can be used to explain any trained model. A widely used group of techniques in this category are gradient-based methods such as Guided Backpropagation (Springenberg et al., 2014) or Integrated Gradients (Sundararajan et al., 2017). Those techniques provide insights into black-box neural networks by visualizing the gradient with respect to the input pixels. Recent studies (Adebayo et al., 2018; Arun et al., 2021) have demonstrated that these explanations do not change significantly when the target model changes, raising concerns about the ability of such post-hoc methods to faithfully reflect the model's behavior.

Perturbation-based methods such as LIME (Ribeiro et al., 2016) and SHAP (Lundberg and Lee, 2017) approximate the decision function by perturbing the input and observing the changes in the output. Also for these methods the explanation's faithfulness to the model's decision mechanism is not guaranteed. Due to their reliance on input perturbations, these methods can exhibit high variability across different runs. Factors such as how the input is segmented into parts (Pihlgren and Främling, 2025), the sampling procedures (Zhang et al., 2019), and the methods used to assign importance to the parts (Chen et al., 2018) all contribute to this variance. As a result, the explanations produced by these methods are often not robust.

Another line of work, including Class Activation Mappings (CAM) (Zhou et al., 2016) and GradCAM (Selvaraju et al., 2017) are based on the networks' final activation map and highlight the regions that are important for a specific class. These techniques are limited by the spatial resolution of their explanations. Moreover, Grad-CAM may highlight regions of an image that a model did not actually use for prediction (Draelos and Carin, 2020).

2.2 Inherently interpretable models

In contrast to post-hoc explanations, some recent work has focused on designing models that are inherently interpretable. These models are constructed such that the explanations are a built-in part of the decision mechanism. Therefore, the explanations can faithfully reveal the true decision mechanism.

Examples of inherently interpretable methods include rule-based models (Lakkaraju et al., 2016; Angelino et al., 2018) which break the decision mechanism into a set of independent if-then rules that are easily interpretable by humans. Inherently interpretable concept bottleneck models (Alvarez Melis and Jaakkola, 2018; Chen et al., 2020; Koh et al., 2020) first make predictions on human-interpretable concepts, and then use these concept activations to generate final predictions. This structure provides human-understandable explanations and enables human intervention. Prototype-based inherently interpretable models (Chen et al., 2019; Barnett et al., 2021) learn a set of prototypes from training images and make predictions by comparing regions of the input image to these prototypes. Both the final prediction and the explanation are derived from the similarity to the learned prototypes. Concept-based methods and prototype-based methods both provide human-friendly explanations. However, these methods do not provide spatial explanations, typically involve many hyperparameters and complex training regimes, and are challenging to train.

A number of works aim to generate inherently interpretable spatial explanations. For instance, BagNets (Brendel and Bethge, 2019; Donteu et al., 2023) produce spatial explanation based on small local image patches that contain class evidence. Recently proposed models like CoDA-Net (Bohle et al., 2021) and its more generalized version, B-cos Networks (Böhle et al., 2024) are currently the state-of-the-art models for providing inherently interpretable visual explanations at the pixel level. These methods employ a dynamic alignment mechanism and formulate the networks' prediction as a weighted sum of the input images.

2.3 Counterfactual-based explanations

Counterfactual-based explanations are a category of methods highly relevant to our work. These methods attempt to answer questions like "What would the image look like if it belonged to a different class?" (Schutte et al., 2021; Joshi et al., 2018; Boreiko et al., 2022; Jeanneret et al., 2023), or they aim to exaggerate features pertinent to the predicted class (Cohen et al., 2021; Singla et al., 2019).

Generative adversarial networks (GAN) (Goodfellow et al., 2020) are widely used for generating counterfactual explanations. For example, Atad et al. (2022) generate counterfactual explanations for Chest X-rays by manipulating the latent style space of StyleGAN. Mindlin et al. (2023) employ CycleGAN to produce attention-based counterfactuals for X-ray image classification. Garg et al. (2024) propose a GAN-based ante-hoc explainable classifier. And Qi et al. (2025) leverage the style space of StyleGAN to generate counterfactual explanations for classifier decisions on prostate MRI scans.

Apart from GANs, Autoencoders (Bank et al., 2023) have also been explored for generating counterfactual ex-

planations. The most relevant one with our work is the Gifsplanation (Cohen et al., 2021, 2025), which generates counterfactual explanations for chest x-rays by shifting the latent space of an autoencoder.

Recently, the Denoising Diffusion Probabilistic Models (DDPM) (Ho et al., 2020) and its variants have achieved remarkable success in the generative tasks and have been explored for generating counterfactual explanations (Augustin et al., 2022). For example, Jeanneret et al. (2023) generate post-hoc counterfactual explanations using DDPM, Bedel and Çukur (2024) propose DreamMR, which is the first diffusion-driven counterfactual explanation method for functional MRI. And Fathi et al. (2024) propose DeCoDEx to improve Diffusion-based Counterfactual Explanations in confounder detection.

Although there are various counterfactual-based explanation approaches, all models that we are aware of generate explanations in a post-hoc manner, which may lead to unfaithful explanations. To our knowledge, we present the first inherently interpretable model that leverages counterfactuals.

2.4 Global interpretability

One important goal of interpretability is to detect and avoid bias of the ML models (Ghassemi et al., 2021). In contrast to local explanations, which reveal the decision mechanism for individual samples, global explanations provide insights into the model behavior for an entire dataset (Reyes et al., 2020) and are therefore particularly helpful for detecting unwanted model behavior. The vast majority of prior research has focused on local explanations, with very few techniques tackling the global explanation problem. Post-hoc explanation SHAP can provide global explanations by running it on every sample and aggregating the SHAP values across the entire dataset, thereby offering insights into the model as a whole. The prototypes in prototypical networks (Snell et al., 2017; Chen et al., 2019) represent the cluster centers of each class and can serve as a global explanation for the classifier. Kim et al. (2016) proposed a technique capable of global explanations by learning both representative prototypical examples of the dataset and criticism examples that do not quite fit the model. Concept activation vectors are another strategy that allows obtaining explanations of entire classes or sets of examples (Kim et al., 2018). The system proposed in (Pereira et al., 2018) achieved both local and global interpretability by jointly considering existing correlations between imaging data, features, and target variables. To our knowledge, our work is the first inherently interpretable model that provides both local and global faithful explanations.

2.5 Visual feature attribution

Visual feature attribution is a task closely related to building explainable deep learning models. Rather than obtaining insights into a model's decision mechanism, visual feature attribution methods aim to visualize evidence of a particular class in an image. We emphasize that visual feature attribution is distinct from interpretable machine learning because its aim is not to predict an outcome. Nevertheless, the most frequently used approach to address this problem is training a neural network for the classification task and then employing spatial post-hoc explanation techniques described earlier (Baumgartner et al., 2017; Jamaludin et al., 2016; Zhu et al., 2017; Pinheiro and Collobert, 2015). Baumgartner et al. (2018) pointed out that these algorithms may lead to only a subset of class-relevant features being detected since not all class-relevant pixels are necessarily used by a classifier for its prediction. The authors proposed an alternative strategy for visual feature attribution based on counterfactuals produced using a Wasserstein GAN. Specifically, the residual between the original image and the generated counterfactual image of the opposite class was used to identify class relevant features. However, the technique requires knowledge of the ground-truth label *a priori* and can therefore not be used for classification. Moreover, the technique is limited to the binary scenario with a healthy and a pathological class. In our work, we build on this work to develop an interpretable multi-label classifier based on counterfactuals.

2.6 Model guidance

Deep neural networks make predictions by recognizing the discriminative features learned from images in the training set. However, deep neural networks have a tendency to exploit shortcuts, and the learned features may not necessarily transfer to the unseen images (Geirhos et al., 2020). To enhance model generalization and reduce potential bias, recent research has focused on incorporating task-relevant information into model guidance. For instance, Fathi et al. (2024) introduced a framework that employs a pre-trained spurious correlation detector to improve the accuracy of diffusion-based counterfactual explanations. Erion et al. (2021) introduced attribution priors such as smoothness and sparsity to the model during training to optimize for higher-level properties of explanations. Pillai et al. (2022) proposed Contrastive Grad-CAM Consistency to regularize the model to produce more consistent explanations. Prior studies (Li et al., 2018; Rao et al., 2023) have demonstrated the effectiveness of bounding box annotations in guiding the model. And Rao et al. (2023) has extensively evaluated various aspects, including loss functions, attribution methods, and depth of model guidance, concluding that incorporating human knowledge guidance into the model

can enhance the interpretability of the explanations and mitigate potential shortcut learning behavior. In this work, we investigate guiding our model through pixel-wise human annotations. We use the energy loss proposed in (Rao et al., 2023) to encourage our model to be "right for the right reason".

3. Methods

3.1 Framework Overview

In this section, we introduce our proposed inherently interpretable "Attri-Net" model. Attri-Net is a multi-label classifier that makes predictions for C classes, where each class c with label $y_c \in \{0, 1\}$ corresponds to the presence or absence of a specific medical finding in an image.

The core idea of our approach is to first counterfactually generate class attribution maps containing the evidence for each diagnostic task $\mathbf{t}_c$ (Fig. 1(a)), and then perform classification solely based on the attribution maps using logistic regression classifiers (Fig. 1(b)). The attribution maps weighted by the logistic regression weights directly show how specific predictions are calculated and provide local explanations. A global explanation can be obtained by jointly visualising learned average representations of the attribution maps for each class (class centers) and the classfiers' linear weights.

3.2 Class Attribution Generator

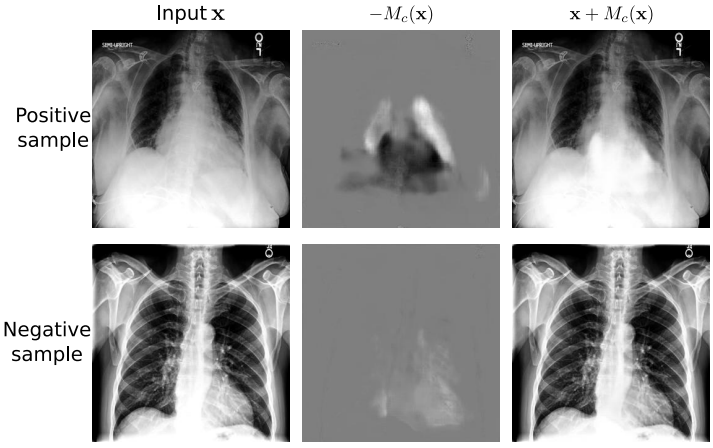


Figure 2: Examples of counterfactual images. The top row shows an input image $\mathbf{x}$ that is positive in cardiomegaly, the attribution map $M_c(\mathbf{x})$ (with a flipped sign for better visualization effect), and the counterfactual image $\hat{\mathbf{x}}$. The bottom row shows images for a negative sample. As expected the residual changes $M_c(\mathbf{x})$ are large for the positive sample and small for the negative sample.

The central component of Attri-Net is the class attribution generator (see Fig. 1(a)). We define this generator as an image-to-image translation network, denoted as $M_c(\mathbf{x}) : \mathbb{R}^{H\times W} \mapsto \mathbb{R}^{H\times W}$, following the approach in (Baumgartner et al., 2018). It learns an additive mapping M_c to produce a counterfactual image $\hat{\mathbf{x}}$, defined as

$$\hat{\mathbf{x}} = \mathbf{x} + M_c(\mathbf{x}) \,.$$

The objective is to make the generated counterfactual image $\hat{\mathbf{x}}$ indistinguishable from the real images sampled from the distribution $p(\mathbf{x}|y_c = 0)$ that do *not* contain class c.

Intuitively, M_c captures the changes required for each pixel in the input to remove the positive effect of class c from the image. Consequently, M_c will contain larger changes for images from the positive class $p(\mathbf{x}|y_c = 1)$ compared with images from the negative class $p(\mathbf{x}|y_c = 0)$. An illustrative example of generated counterfactuals $\hat{\mathbf{x}}$ for the disease "cardiomegaly" is presented in Fig. 2, with additional examples available in the Appendix A.

In order to enable M_c to learn the difference between the class-positive and class-negative distributions of class c, we employ a class-specific discriminator network D_c. The network D_c is trained alongside M_c to distinguish fake images $(\mathbf{x} + M_c(\mathbf{x}))$ from real negative class images from $p(\mathbf{x}|y_c = 0)$. Specifically, we adopt the adversarial Wasserstein GAN loss (Arjovsky et al., 2017; Baumgartner et al., 2018). The discriminator loss can be written as

$$\begin{aligned}\mathcal{L}^{(c)}_{\text{disc}} = &\mathop{\mathbb{E}}_{p(\mathbf{x}|y_c=0)}[-D_c(\mathbf{x})] \\ &+ \mathop{\mathbb{E}}_{p(\mathbf{x}|y_c=1)}[D_c(\mathbf{x} + M_c(\mathbf{x}))] \,. \end{aligned} \tag{1}$$

The adversarial class attribution generator loss maximises the second term of the equation above:

$$\mathcal{L}^{(c)}_{\text{adv}} = \mathop{\mathbb{E}}_{p(\mathbf{x}|y_c=1)}[-D_c(\mathbf{x} + M_c(\mathbf{x}))] \,. \tag{2}$$

The class attribution generator loss ensures that $\hat{\mathbf{x}}$ is a realistic counterfactual not containing class c and, by extension, that M_c is a realistic residual attribution map containing all positive evidence of class c.

To encourage the class attribution maps to focus on class-relevant information and avoid learning superfluous pixels not belonging to a given class, we incorporate an additional L_1 regularization term on M_c to encourage it to be sparse. The regularization term is defined as follows:

$$\begin{aligned}\mathcal{L}^{(c)}_{\text{reg}} = &\alpha_0 \mathop{\mathbb{E}}_{p(\mathbf{x}|y_c=0)}[\|M_c(\mathbf{x})\|_1] \\ &+ \alpha_1 \mathop{\mathbb{E}}_{p(\mathbf{x}|y_c=1)}[\|M_c(\mathbf{x})\|_1] \,. \end{aligned} \tag{3}$$

We assign a higher weight, denoted as α_0, to samples from the class-negative category and a lower weight, denoted as α_1, to class-positive examples. This weighting

reflects the intuition that minimal adjustments are required for samples from the negative class compared to those from the positive class. We choose $\alpha_0 = 2, \alpha_1 = 1$ based on preliminary parameter tuning experiments and use them for all experiments in this paper.

In the context of the multi-label classification task, our objective is to generate individual explanations for each medical finding. Although it is feasible to design a network M to produce class attribution maps for all classes as multiple output channels in a single forward pass, preliminary experiments on such an architecture revealed inadequate class attribution in the multi-label scenario. Instead, we introduce a task switch mechanism based on recent work (Sun et al., 2021) to enable the class attribution generator to switch between various diagnostic tasks. As shown in Fig. 1(a), the task code $\mathbf{t}_c$ is injected into the class attribution generator through adaptive instance normalization (AdaIN) layers to switch the network M to a specific mode M_c that focuses on diagnosing class c. Each task code $\mathbf{t}_c$ is a one-hot encoding spatially upsampled by a factor of 20 as in (Sun et al., 2021). Then the one-hot vector task code is converted into a task embedding via a small fully connected network and fed to AdaIN layers which are placed throughout the network. We apply the same mechanism also to the discriminator network D such that it can provide correct feedback to the respective class.

With task switching mechanism, the class attribution generator and discriminator can now be expressed as $M_c(\mathbf{x}) = M(\mathbf{x}, \mathbf{t}_c)$, and $D_c(\mathbf{x}) = D(\mathbf{x}, \mathbf{t}_c)$. The class attribution maps for all labels can be obtained by repeated forward passes through M while iterating through the $\mathbf{t}_c$ vectors of all classes. The specific architecture of M and D is discussed in Sec.3.7, and in more detail in Appendix F.

3.3 Classification Mechanism

As $M(\mathbf{x}, \mathbf{t}_c)$ learns the changes required to convert an image $\mathbf{x}$ into a sample from the negative distribution $p(\mathbf{x}|y_c = 0)$, it inherently encodes class-specific information and can be directly used for predicting class c. Test images that contain the disease will require large changes to make them appear healthy, while images that are already healthy are characterised by very small changes $M(\mathbf{x}, \mathbf{t}_c)$ (see Fig. 2). Since the maps $M(\mathbf{x}, \mathbf{t}_c)$ allow easy differentiation of positive and negative samples for a given disease, we can employ a simple linear classifier for the final classification step of each class. As shown in Fig. 1(b), the respective attribution map is downsampled and then used as input to the classifier. That is,

$$p(y_c|\mathbf{x}) = \sigma\Big(\sum_{i,j} w_{ij}^{(c)} \cdot S_\gamma(M(\mathbf{x}, \mathbf{t}_c))_{ij}\Big)\,, \tag{4}$$

where S_γ is a 2D average pooling operator that downsamples by a factor of γ, $w_{ij}^{(c)}$ denotes the weights associated with each pixel of the down-sampled attribution map for class c, and σ is the sigmoid function. In preliminary experiments, we evaluated with various values of γ and found $\gamma = 32$ to perform robustly and we used this value for all experiments.

The classifiers for each class are trained using a standard binary classification loss $\mathcal{L}_{\text{cls}}^{(c)}$, i.e. a binary cross entropy loss. Note that, since our framework is trained end-to-end, M also receives gradients from that loss and is thereby encouraged to create class attribution maps that are linearly classifiable.

3.4 Center Loss

To further encourage the attribution maps to be discriminative, we apply the center loss proposed by (Wen et al., 2016), which has demonstrated its efficacy in fostering more distinctive feature representations. Extending this idea, we define center loss $\mathcal{L}_{\text{ctr}}^{(c)}$ as follows:

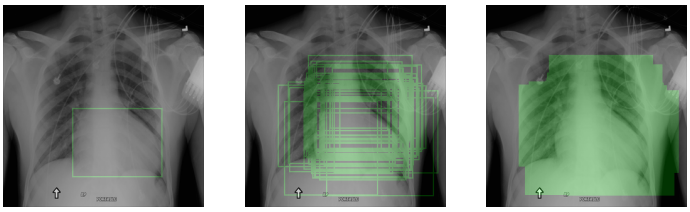
$$\begin{aligned}\mathcal{L}_{\text{ctr}}^{(c)} = &\frac{1}{2} \mathop{\mathbb{E}}_{p(\mathbf{x}|y_c=0)} \Big[\|M(\mathbf{x}, \mathbf{t}_c) - \mathbf{v}_{y_c=0}\|_2^2\Big] \\ &+ \frac{1}{2} \mathop{\mathbb{E}}_{p(\mathbf{x}|y_c=1)} \Big[\|M(\mathbf{x}, \mathbf{t}_c) - \mathbf{v}_{y_c=1}\|_2^2\Big]\,,\end{aligned} \tag{5}$$

where $\mathbf{v}_{y_c=0}, \mathbf{v}_{y_c=1} \in \mathbb{R}^{H \times W}$ are the negative and positive class centers of attribution map for class c.

The class centers are learnable and updated on mini-batch along with the update step of network M with separate optimizers as described by (Wen et al., 2016). In the forward pass, the center loss calculates the L_2 distance between the attribution map $M(\mathbf{x}, \mathbf{t}_c)$ and the corresponding class center. In the backward pass, both the attribution map $M(\mathbf{x}, \mathbf{t}_c)$ and the associated class center receive gradients from the loss and are updated accordingly. Since the center loss penalizes the distance between an attribution map and its corresponding class center, it encourages the attribution map to move closer to the class center. This reduces the intra-class distance while increasing the inter-class separation, making the attribution map more distinctive between positive and negative classes. Furthermore, the class centers aggregate the mean attribution maps of each class c across the dataset, offering insight into the model's average behavior. At the end of the training stage, these class centers can serve as a global explanation for the entire model as described in Sec.3.6.2.

3.5 Model Guidance

The inherently interpretable construction of Attri-Net allows to constrain the explanations using human guidance



(a)ground truth bbox **(b)**all ground truth bbox **(c)**binary pseudo-mask

Figure 3: Generation of pseudo guidance masks. An example for the disease cardiomegaly from the ChestX-ray8 dataset is shown. **(a)** A chest X-ray image with its ground truth bounding box annotation for cardiomegaly. **(b)** The same image with multiple cardiomegaly bounding box annotations from other cases in the ChestX-ray8 dataset. **(c)** The same image with the binary pseudo-mask generated from multiple cardiomegaly bounding box annotations.

in the form of disease segmentations or bounding boxes. Since the explanations are part of the decision mechanism, constraining the explanations directly affects the model's decision mechanism.

To incorporate guidance into our model we propose to train it with a guidance loss based on an energy-based formulation (Rao et al., 2023):

$$\mathcal{L}_{\text{guid}}^{(c)} = 1 - \frac{\sum_{h=1}^{H}\sum_{w=1}^{W} G_{c,hw}|M_c|_{,hw}}{\sum_{h=1}^{H}\sum_{w=1}^{W}|M_c|_{,hw}}\,. \tag{6}$$

Here, the guidance mask G_c is a binary image matching the input X-ray's dimensions, where pixels inside lesion regions are set to 1 and others to 0. It is derived from ground-truth lesion annotations, such as bounding boxes or disease segmentations. $|M_c|$ is the absolute value of the attribution map. The guidance loss encourages the model to focus on the regions within the guidance mask that contain task-relevant features while ignoring the regions outside.

3.5.1 Full guidance

In the ideal case, annotations are available for every training image and we can perform full guidance by directly incorporating the guidance loss with the other loss terms to jointly optimize for classification performance and localization of task-relevant features. This is, for example, the case for the VinDr-CXR dataset Nguyen et al. (2022), where expert-labeled bounding box annotations are available for every sample.

3.5.2 Pseudo-guidance

In most cases, expert-labeled bounding boxes or segmentations are costly to obtain and are available only for a small portion of the training data. For example, the large-scale chest X-ray dataset ChestX-ray8 (Wang et al., 2017) contains 108,948 images while only 880 images are provided with disease bounding box annotations. Most of the samples in this dataset only have class labels at the image level. As chest X-ray images primarily capture organs with relatively fixed positions, there is a noticeable overlap of disease incidence regions among individual patients. This intrinsic property of chest X-rays provides an opportunity to incorporate a pseudo guidance mechanism for samples that lack pixel-wise ground truth annotations. We create pseudo-guidance masks for a class by calculating the union of all ground truth bounding box annotations for that class. An example pseudo-mask generation for the class cardiomegaly is shown in Fig. 3.

We investigated different methods to incorporate pseudo-guidance alongside the limited ground truth annotations and determined that a mixed guidance strategy yielded the best performance. Specifically, during training, when a sample has a ground truth annotation, we guide the model using the actual ground truth. In cases where there is no ground truth, we use the pseudo masks as guidance to prompt the model to emphasize regions with a higher prevalence of diseases. We observed that oversampling cases with ground truth annotations during training such that they appear with a frequency of $\frac{1}{10}$ enhances the localization performance. We provide a more detailed analysis in the Appendix C.

3.6 Obtaining Explanations

During inference time, our model is capable of producing a prediction as well as a local per-sample explanation and a global explanation on the dataset level. In the following we describe how those two explanations types can be obtained.

3.6.1 Obtaining Local Explanations

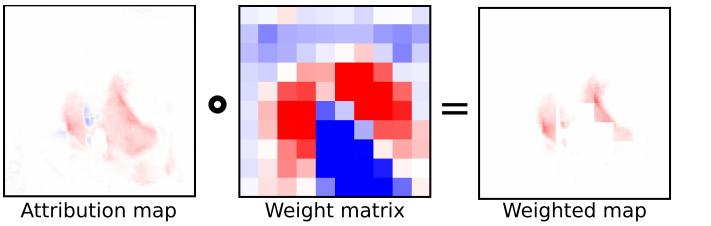


Figure 4: Local explanation of Attri-Net for an example from the CheXpert dataset with cardiomegaly. The weighted attribution map serves as local explanation for a specific prediction. It is defined as the element-wise product of the attribution map from class attribution generator and the weight matrix from corresponding logistic regression classifier.

Attri-Net directly allows us to obtain local explanations by considering the weighted attribution maps. As shown

in Fig. 4, the weighted attribution map is calculated by the element-wise product of the attribution map and the upsampled weight matrix of the corresponding logistic regression classifier. Note that this equivalent to the weights and feature multiplication inside the logistic regression classifiers, i.e. the term inside the sum of Eq. 4. The weighted attribution map therefore directly reveals the classifier's decision mechanism. We keep the positive values in the weighted map as evidence of class c. This attribution map, together with the final prediction, is provided to the users for inspection of a particular diagnosis.

3.6.2 Obtaining Global Explanations

Prior to the deployment, model developers need to ensure that the classifier behaves reliably on a global level (e.g. does not use any spurious features). Attri-Net addresses this need by providing global explanations for the whole model's mechanism. We define the global explanation as the combination of class centers and classifier weights. After the training stage, the class centers $\mathbf{v}_{y_c=0}, \mathbf{v}_{y_c=1} \in \mathbb{R}^{H\times W}$, as defined in Sec.3.4, capture the average patterns of attribution maps across the entire training set, enabling users to assess whether the input features of the classifiers are clinically meaningful. Meanwhile, the logistic regression classifier weights reveal the areas of the images that the classifier is paying attention to. Thus, the class centers, together with the weight matrices, provide a transparent insight into Attri-Net's classification process, offering an explanation for the entire model. Figure 5 illustrates the global explanation for Attri-Net trained on the CheXpert dataset.

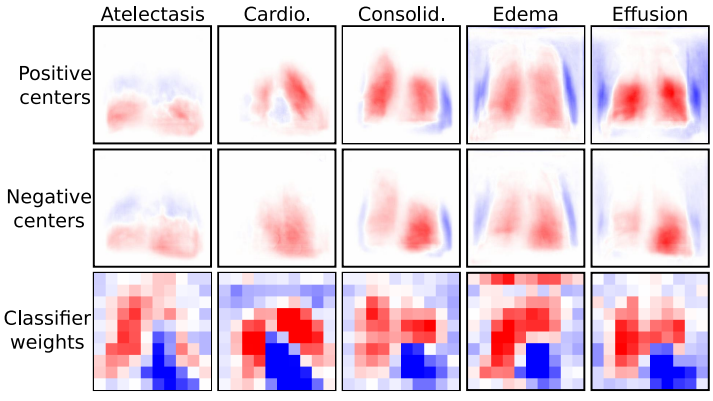


Figure 5: Global explanation of Attri-Net for a model trained on the CheXpert dataset. The positive and negative class centers of attribution maps and the corresponding classifiers' weight matrix together provide a global explanation for Attri-Net.

3.7 Model Architecture and Training

Attri-Net contains a total of 55.86 million parameters and consists of three key components: the class attribution generator M, the discriminator network D, and the logistic regression classifiers. We implemented M and D based on the StarGAN (Choi et al., 2018) architecture. To enable the task-switching functionality, we replaced instance normalization layers in the original StarGAN architecture with adaptive instance normalization (AdaIN) modules following (Sun et al., 2021). Each logistic regression classifier was implemented as a fully connected neural network layer with a 2D average pooling layer for downsampling. The architectures are described in more detail in the Appendix F.

The Attri-Net framework can be trained end-to-end with five loss terms enforcing our essential requirements: Firstly, we used the classification loss $\mathcal{L}_{\text{cls}}^{(c)}$, which apart from encouraging accurate classification, ensures that the attribution maps preserve sufficient class relevant information for a satisfactory classification result. Secondly, we adopted the adversarial loss $\mathcal{L}_{\text{adv}}^{(c)}$ and the regularization term $\mathcal{L}_{\text{reg}}^{(c)}$ to encourage discriminative and sparse attribution maps. Furthermore, the center loss term $\mathcal{L}_{\text{ctr}}^{(c)}$ moves the attribution maps toward the class centers and provides global explanations. Lastly, the guidance loss $\mathcal{L}_{\text{guid}}^{(c)}$ can be optionally used to inject human guidance during training and encourages the attribution maps to be consistent with human knowledge.

The overall training objective for the class attribution generator M with weight parameters φ was given by

$$\mathcal{L} = \sum_c \lambda_{\text{ad}}\mathcal{L}_{\text{adv}}^{(c)} + \lambda_{\text{cl}}\mathcal{L}_{\text{cls}}^{(c)} + \lambda_{\text{re}}\mathcal{L}_{\text{reg}}^{(c)} + \lambda_{\text{ct}}\mathcal{L}_{\text{ctr}}^{(c)} + \lambda_{\text{gd}}\mathcal{L}_{\text{guid}}^{(c)}\,, \tag{7}$$

where we used the hyperparameters λ_* to balance the losses. We chose $\lambda_{\text{ad}} = 1,\ \lambda_{\text{cl}} = 100,\ \lambda_{\text{re}} = 100,\ \lambda_{\text{ct}} = 0.01,\ \lambda_{\text{gd}} = 30$ for our experiments. An ablation study on the effect of the different losses can be found in Sec.4.6.

Throughout the training, we repeatedly iterate through the different classes c and, for each, draw two mini-batches, one containing positive samples of the current class and the other negative samples. We iteratively update M, D, and the classifiers. In one training step, only the classifier corresponding to the current target class is updated, while the classifiers for other diseases remain frozen. Following the original Wasserstein GAN (Arjovsky et al., 2017), the discriminator D undergoes more frequent updates to ensure it remains close to optimality throughout training. Specifically, we perform five discriminator update steps for each generator step. Additionally, for every 100 generation step and the first 25 generator steps, we perform an extra

100 discriminator update steps. Furthermore, with each generator step, we update the classifier five times.

We employed the ADAM optimizer (Kingma and Ba, 2014) with a learning rate of 10^{-4} and a batch size of 4 for the optimization of M, D, and the logistic regression classifiers. Additionally, following (Wen et al., 2016), we used stochastic gradient descent with a learning rate of 0.1 for updating the class centers in the center loss module. The model was trained for 100,000 generator steps to ensure both the classification performance and the attribution maps achieve a stable state. We trained Attri-Net on a single NVIDIA 2080Ti GPU for three days. The best model was chosen by evaluating the area under the ROC curve (AUC) on the validation set. After the training, the optimal decision threshold for each class was obtained by maximising the Youden index (sensitivity + specificity - 1) on the validation set. We also performed this step for the baseline methods. During inference, the average time for predicting all diseases on a single chest X-ray image is 0.194 seconds. Since the attention maps are the intermediate output during inference, generating explanations does not introduce additional computational cost.

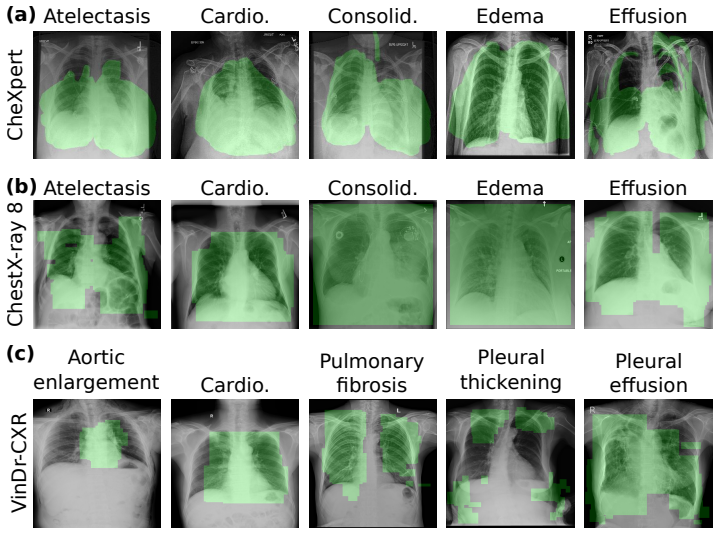


Figure 6: Binary pseudo masks on all three examined datasets.

Table 1: Number of samples for creating pseudo guidance.

Dataset	Atelet.	Cardio.	Consol.	Edema	Effusi.
CheXpert	75	66	32	42	64
ChestX-ray8	76	54	0	0	52

4. Experiments and Results

4.1 Experiment settings

4.1.1 Data

We performed experiments on three chest X-ray datasets: CheXpert (Irvin et al., 2019), ChestX-ray8 (Wang et al., 2017), and VinDr-CXR (Nguyen et al., 2022). Different from our prior work (Sun et al., 2023b), we now had access to the officially released test sets of CheXpert and VinDr-CXR datasets. In this paper, we present results for all three datasets based on the official data splits, ensuring the comparability of our findings with other published works that also used the official split. We used the official train, validation and test set for experiments on CheXpert dataset. Since ChestX-ray8 and VinDr-CXR only provide train, test split, we generated train and validation sets on the official train set with a split ratio of 0.8.

We scaled all images to a smaller size of 320×320 pixels for training and scaled the bounding box annotations and segmentation masks accordingly. Since most of the chest X-ray images in the three datasets are frontal, we excluded a small number of lateral images from the CheXpert dataset. On CheXpert and ChestX-ray8 datasets, following (Irvin et al., 2019), we focused on five findings based on their clinical importance and prevalence: (a) Atelectasis, (b) Cardiomegaly, (c) Consolidation, (d) Edema, and (e) Pleural effusion. The VinDr-CXR dataset is much smaller than the above two datasets and has a different label distribution. To make sure there are enough positive samples for the model to learn disease relevant features, we selected the following five pathologies: (a) Aortic enlargement, (b) Cardiomegaly, (c) Pulmonary fibrosis, (d) Pleural thickening, and (e) Pleural effusion.

4.1.2 Data-splits for Model Guidance

Though there are a large number of chest X-ray images in the CheXpert and ChestX-ray8 datasets, the number of images with pixel-level annotations of pathologies is much smaller. As explained in Sec.3.5.2, we addressed this limitation by generating pseudo masks. These masks were created using the limited ground truth pixel-level annotations and were then used as pseudo-guidance for samples lacking such annotations. In the following we describe how the existing data with pixel-level annotations was split for pseudo-guidance and evaluation.

The ChestX-ray8 dataset provides bounding box annotations for 880 images. We divided these annotations with a ratio of 40% for generating pseudo masks for training and 60% for evaluation. Note that in ChestX-ray8, there are no bounding box annotations available for consolidation or edema positive samples. Therefore, we created a very loose guidance in the form of a 300×300 square, aiming to guide the model to focus on the central regions of the image.

In recent work (Saporta et al., 2022), ground truth segmentations were released for 187 images in the CheXpert validation set and 499 images in the test set. We used the ground truth segmentations from the CheXpert validation

set to create pseudo masks, reserving the test set for evaluation. The distribution of samples used to create pseudo guidance can be found in Table 1, and the resulting pseudo masks are presented in Fig. 6.

The VinDr-CXR dataset includes bounding box annotations for all images, allowing us to assess the full guidance we described in Sec.3.5.1. As a comparison, we additionally performed an experiment by using the pseudo-guidance mechanism as introduced in Sec.3.5.2. We generated pseudo masks by using a small subset of 75 samples from each disease (See Fig. 6(c)) and only used them to guide the model during training as we did with ChestX-ray8 and CheXpert datasets. This allowed us to evaluate the performance gap between full guidance and pseudo-guidance.

4.1.3 Baseline methods

To assess our model's classification performance, we compared it with a standard black-box ResNet50 (He et al., 2016) model as well as the B-cos ResNet50 (Böhle et al., 2024), a state-of-the-art inherently interpretable classifier. We trained our Attri-Net using the settings described in Sec.3.7. For training the ResNet50 and the B-cos ResNet50, we employed the ADAM optimizer (Kingma and Ba, 2014) with a learning rate of 10^{-4} and a batch size of 4. The models were trained for 50 epochs to achieve convergence, and the best-performing models with the highest AUC score on the validation set were selected.

We further assessed the local explanations provided by our proposed AttriNet, the B-cos Network, as well as five post-hoc explanation techniques which we applied to the ResNet-50 black-box model. Specifically, we compared to Guided Backpropagation (Springenberg et al., 2014), GradCAM (Selvaraju et al., 2017), LIME (Ribeiro et al., 2016), SHAP (Lundberg and Lee, 2017) and the recently proposed Gifsplanation (Cohen et al., 2021).

Since our model and B-cos Networks are both inherently interpretable, they could both optionally be trained with our guidance loss proposed in Sec.3.5. We trained both models with and without model guidance to investigate its effect on explanation quality and classification performance.

4.2 Evaluation Metrics

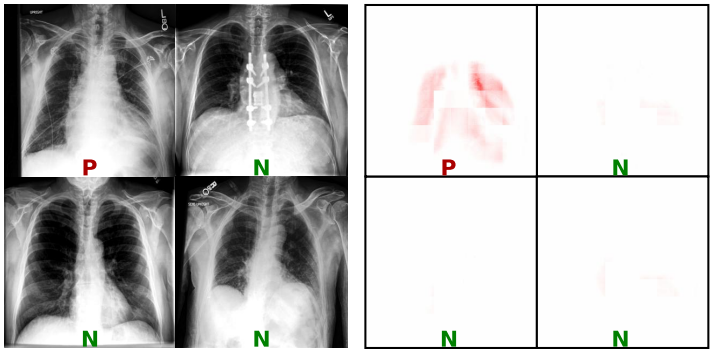
4.2.1 Classification Metric

The classification performance was evaluated using the area under the ROC curve (AUC).

4.2.2 Explainability Metrics

We evaluated the quality of the explanations using two metrics we defined: class sensitivity and disease sensitivity.

We defined class sensitivity following the approach described by Bohle et al. (2021). Class sensitivity measures



(a) Input image grid **(b)** Local explanation grid

Figure 7: Example of image grid used for computing the class sensitivity metric for the disease cardiomegaly. **(a)** Input image grid with one positive and three negative samples. **(b)** Corresponding local explanations generated by Attri-Net.

the intuition that explanations should be different for different image classes (Khakzar et al., 2022). In our scenario, class sensitivity implies that the explanations for the disease-positive class should be distinct from those for the disease-negative class. Following Bohle et al. (2021), we generated a series of 2×2 grids of explanations, each containing only one positive example of a given disease (see example in Fig. 7). The class sensitivity was then computed by dividing the sum of attributions in the positive example by the sum of all attributions in the grid. Ideally, all disease negative explanations should be blank because there is no disease effect in the negative samples, resulting in a class sensitivity score of 1. Conversely, in the worst-case scenario where positive and negative sample explanations are indistinguishable, the class sensitivity score would be $\frac{1}{4}$. Similar to Bohle et al. (2021), we created 200 grids using the most confident positive and negative samples and computed the average class sensitivity across class c. In cases where certain diseases lacked an adequate number of positive samples in the test set, we constructed fewer image grids, with the number of grids equal to the count of correctly predicted positive examples.

We defined the disease sensitivity following the energy-based pointing game metric proposed by Wang et al. (2020). Disease sensitivity measured how much of the attributions for a given disease are concentrated inside the ground truth bounding box or segmentation mask. It was computed by summing the attributions within the bounding box annotation and dividing by the sum of all attributions. The disease sensitivity scores were averaged across all classes c and samples with ground truth pixel-wise annotations. Specifically, this included the remaining 60% of bounding box annotated samples for the ChestX-ray8 dataset, as well as all samples in the test set for the CheXpert and VinDr-CXR datasets.

Table 2: Classification performance measured by area under the ROC curve (AUC).

Model	CheXpert	ChestX-ray8	VinDr-CXR
Stanford baseline (Irvin et al., 2019)	0.907	-	-
DeepAUC (Yuan et al., 2021)	0.930	-	-
LSE (Ye et al., 2020)	-	0.755	-
ChestNet (Ye et al., 2020)	-	0.790	-
ResNet50	**0.875**	0.778	0.764
B-cos ResNet50	0.866	0.757	**0.836**
B-cos ResNet50 (guided)	0.839	0.754	0.828
ours	0.873	**0.779**	0.789
ours (pseudo guidance)	0.848	0.774	0.773
ours (full guidance)	-	-	0.782

4.3 Evaluation of Classification Performance

The classification outcomes of our model, the black-box model ResNet50, the inherently interpretable model B-cos ResNet50, and the guided variants of our model and B-cos ResNet50 are presented in Table 2. The disease-wise classification performance is provided in Table 5 in Appendix B. Additionally, we report the top-performing result in the CheXpert competition (Yuan et al., 2021), the baseline outcome outlined in the dataset paper (Irvin et al., 2019), and published result from (Ye et al., 2020) on ChestX-ray8. Attri-Net overall performed comparable to the state-of-the-art, with an AUC that was similar to other methods on CheXpert, slightly lower on Vindr-CXR, and slightly better on ChestX-ray8. In comparison to our prior work (Sun et al., 2023b), both the ResNet50 model and our own model exhibited a decline in classification performance when assessed on the Vindr-CXR official test set. This is primarily due to the increased difficulty of the test set. Furthermore, we observed a slight decrease in classification performance with the guided versions of both our model and B-cos ResNet50. This suggested that the unguided versions may exploit class-irrelevant features to achieve optimal classification performance, whereas these features were restricted in the guided models. Additionally, on the Vindr-CXR dataset, the model trained with full guidance outperformed the model trained with pseudo guidance by a small amount, indicating that more precise guidance leads to improved performance.

4.4 Evaluation of the Local Explanations

We derived local explanation for Attri-Net using the weighted attribution maps as detailed in Sec.3.6.1 and compared our local explanations with those from B-cos ResNet50 and five post-hoc methods that explain the prediction for the black-box ResNet50 model.

4.4.1 Quantitative Analysis of Explanations

Table 3 presents the class sensitivity and disease sensitivity of all local explanations that we compared. Additional disease-wise results along with the 95% confidence interval are provided in Appendix E.

Our method consistently outperformed all other methods, demonstrating significantly higher class sensitivity across all datasets. The high class sensitivity underscores Attri-Net's ability to provide more distinguishable explanations for disease-positive and disease-negative samples.

Moreover, as depicted in Table 3, when trained with guidance, our Attri-Net achieved the highest disease sensitivity score across all datasets, indicating superior localization of disease-relevant regions compared to alternative methods. A paired t-test of our model trained with and without guidance on the CheXpert dataset reveals a statistically significant improvement when guidance was incorporated ($t = 28.284$, $p < 0.0001$). Additionally, the discrepancy in disease sensitivity between models trained with full guidance and those with pseudo guidance on VinDr-CXR emphasizes the importance of precise guidance in enhancing localization performance. Notably, similar to our Attri-Net, the disease sensitivity of the other inherently interpretable model, B-cos ResNet50, also improved after adding guidance on CheXpert and ChestX-ray8 datasets. However, the degree of improvement fell short of ours. This substantial performance gap highlighted that Attri-Net was particularly suitable for integrating guidance. This is because Attri-Net is explicitly designed for multi-label classification by distinguishing between positive and negative samples for each disease, resulting in more class-specific attribution maps. Therefore, the guidance effectively helps the model focuson the most relevant regions.

4.4.2 Qualitative Analysis of Explanations

The qualitative examination of example explanations supported the quantitative results. Attri-Net was capable of generating local explanations that effectively emphasize the anatomical regions associated with the respective classes (see Fig. 8 for a representative example from the ChestX-ray8 dataset). Besides, explanations for highly confident predictions, such as cardiomegaly, exhibited a more pronounced disease effect compared to negative predictions such as Edema. Furthermore, the attributions for different classes were clearly distinct, each highlighting different anatomical areas. In contrast, it was challenging to understand the explanations from other post-hoc methods and the inherently interpretable baseline. For instance, explanations derived from Guided Backpropagation and B-cos ResNet50 were very noisy and hard to interpret. The explanations from the counterfactual-based Gifsplanation approach were easier to interpret, yet they consistently emphasized simi-

Table 3: Comparison of class sensitivity and disease sensitivity.

Model	Class sensitivity			Disease Sensitivity		
	CheXpert	ChestX-ray8	VinDr-CXR	CheXpert	ChestX-ray8	VinDr-CXR
GB	0.303	0.263	0.267	0.175	0.176	0.047
GCam	0.191	0.225	0.176	0.192	0.125	0.061
LIME	0.254	0.229	0.252	0.103	0.122	0.031
SHAP	0.351	0.434	0.306	0.219	0.278	0.067
Gifsp.	0.300	0.688	0.317	0.191	0.178	0.052
B-cos ResNet50	0.266	0.276	0.240	0.259	0.235	0.089
B-cos ResNet50 (guided)	0.271	0.322	0.240	0.279	0.247	0.075
ours	**0.684**	0.951	0.862	0.207	0.158	0.075
ours (pseudo guidance)	0.622	**0.965**	**0.920**	**0.400**	**0.327**	0.156
ours (full guidance)	-	-	0.872	-	-	**0.204**

lar regions for different diseases. More examples from the CheXpert and VinDr-CXR datasets can be found in the Appendix D.

We further examined Attri-Net explanations on images with pixel-wise ground truth annotations (Fig. 10) and observed that after adding pseudo-guidance to the model, the local explanations produced by Attri-Net better matched the areas where the pathologies are located, which was consistent with the high disease sensitivity in our quantitative evaluation. This suggested that incorporating guidance assisted the model in being "right for the right reason."

4.5 Evaluation of the Global Explanations

We assessed the performance of the global explanation mechanism of Attri-Net by investigating whether it can be used to uncover synthetically induced spurious correlation.

Following our previously proposed evaluation framework (Sun et al., 2023a) we created a semi-synthetic dataset derived from the CheXpert dataset by contaminating 50% of cardiomegaly positive samples with spurious tag signal (see Fig. 9 (a), (b)). This produces a spurious correlation between the tag and the presence of the disease cardiomegaly. In (Sun et al., 2023a) we showed that models trained on such a biased dataset will exhibit shortcut learning behavior, wherein they rely on spurious signals rather than relevant features for prediction.

In this work, we trained the Attri-Net model on this contaminated dataset and qualitatively and quantitatively evaluated the model's reliance on the shortcut using the global explanation mechanism. Note that the other interpretable methods we assessed in Sec.4.4 lack the capability to offer global explanations. In fact Attri-Net is to our knowledge the first method to provide global spatial explanations. We thus limit our proof-of-concept to Attri-Net.

In a second step, we investigated using model guidance to alleviate the model's shortcut learning behavior. Since we know the regions of spurious signals, we created guidance to avoid the spurious signal's region and encouraged the model to use the features inside the center region of the image (Fig. 9 (c)). We evaluated the global explanations for the model trained with guidance to verify whether the inclusion of guidance could reduce the model's dependence on spurious signals.

4.5.1 Quantitative Results

Since we have introduce a known shortcut behaviour on the tag signal, we expect a global explanation to correctly highlight the reliance on the tag signal in contaminated test images. To measure this we have previously introduced the confounder sensitivity metric (Sun et al., 2023a). Confounder sensitivity measures how many of the top 10% most significant pixels identified by the explanation fall on the synthetically added tag signal. We computed the confounder sensitivity using the positive class center of cardiomegaly, resulting in a score of 0.747. This indicates that 74.7% of the spurious signal pixels were captured by the top 10% most significant pixels in the positive class center of cardiomegaly.

To further verify that adding guidance can direct the model towards using the correct features, we added the guidance shown in Fig. 9(c) to the model to encourage the model to avoid using spurious signals. After adding guidance, the confounder sensitivity of the new class center dropped to 0.002, which means that the most significant explanation pixels are not on the spurious signals anymore.

4.5.2 Qualitative Results

We visualize the global explanation of the model trained on the contaminated CheXpert dataset in Fig. 11(a) and assess if the spurious signal can be detected in the first place. From the positive class center of cardiomegaly, we can clearly see the text "CXR-ROOM1" which we added as the spurious signal. Meanwhile, we observed faintly discernible textual signals of edema within the class centers, indicating

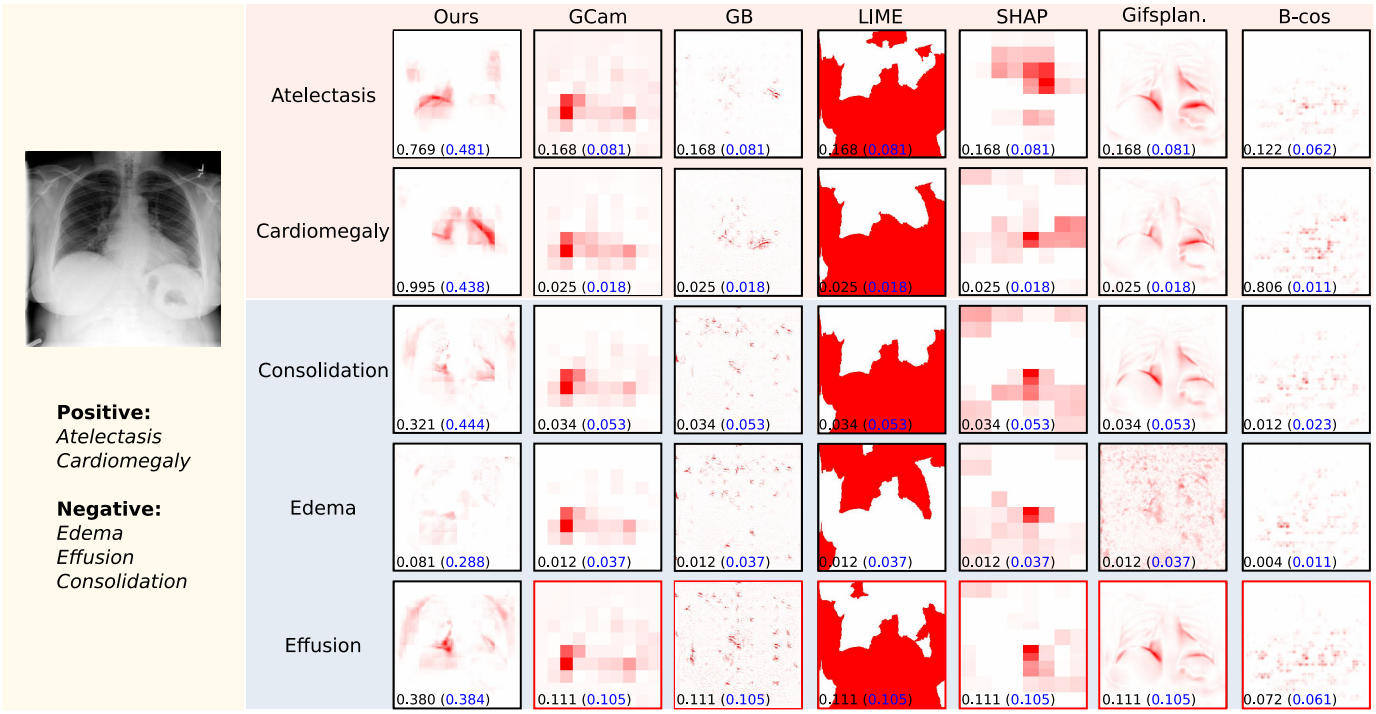
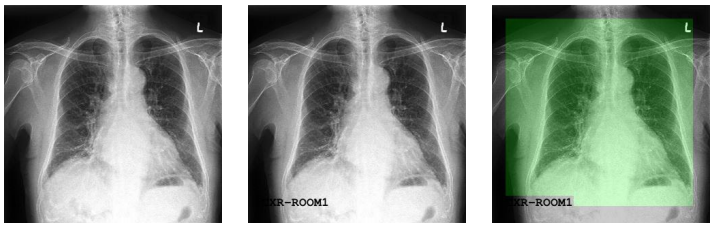


Figure 8: Visual comparison of explanations for an example image from the ChestX-ray8 dataset. Predicted class probabilities are indicated in the lower left corner of each attribution map with the respective decision threshold in parentheses. Wrong predictions are highlighted with red boxes.



(a) original **(b)** contaminated **(c)** guidance-mask

Figure 9: Synthetically induced short-cut learning for global interpretability experiment. **(a)** A cardiomegaly positive sample from the CheXpert dataset. **(b)** The contaminated sample after adding the text "CXR-ROOM1" as a spurious signal. **(c)** The guidance mask encourages the model to use the features inside it and avoid using the spurious signal.

that Attri-Net used features from the spurious tag signal when diagnosing edema. Given that we only introduced the spurious signal to the positive class for cardiomegaly, this observation suggested a potential correlation between cardiomegaly and edema, agreeing with clinical observation in (Dodek et al., 1972). Since our model was trained to capture all information related to diagnosing disease, it was particularly useful for detecting faulty model behavior and the potential bias in the datasets. After adding model guidance, we found that the text was removed from the new global explanation (See Fig. 11(b)), which indicated that the new model trained with guidance was less dependent on the spurious signal.

4.6 Ablation study

We performed an ablation study on the five loss terms for training the class attribution generator network M described in Sec.3. In Fig. 12, we show examples of attribution maps from our model trained with different losses, and in Table 4, we list the quantitative evaluation results of these models, respectively.

We found that models trained with different losses perform similarly in classification AUC. However, the attribution maps visually change greatly, and the quantitative evaluation results of class sensitivity and disease sensitivity vary a lot. As shown in Fig. 12, after adding the adversarial loss term $\mathcal{L}_{\text{adv}}$, the attribution maps focus on disease-relevant regions and become easy to understand. The regularization term $\mathcal{L}_{\text{reg}}$ encourages the attribution maps to be sparse. The addition of center loss $\mathcal{L}_{\text{ctr}}$ improves the classification slightly but does not substantially affect the attribution maps. More importantly, the class centers provide a possible way to interpret Attri-Net globally. From Table 4, the model guidance loss $\mathcal{L}_{\text{guid}}$ greatly improves disease sensitivity, which is clearly shown from the attribution maps (last column) in Fig. 12.

5. Discussion and Conclusion

We proposed Attri-Net, a novel inherently interpretable multi-label classifier that provides faithful local and global

Table 4: Ablation study on five losses. Evaluated on ChestX-ray8 dataset.

Model	Loss terms	Classification AUC	Class sensitivity	Disease sensitivity
Attri-Net$_{cls}$	$\mathcal{L}_{cls}$	0.756	0.436	0.200
Attri-Net$_{cls_adv}$	$\mathcal{L}_{cls} + \mathcal{L}_{adv}$	0.765	0.665	0.211
Attri-Net$_{cls_adv_reg}$	$\mathcal{L}_{cls} + \mathcal{L}_{adv} + \mathcal{L}_{reg}$	0.778	0.954	0.178
Attri-Net$_{cls_adv_reg_ctr}$	$\mathcal{L}_{cls} + \mathcal{L}_{adv} + \mathcal{L}_{reg} + \mathcal{L}_{ctr}$	**0.779**	0.951	0.158
Attri-Net$_{all}$	$\mathcal{L}_{cls} + \mathcal{L}_{adv} + \mathcal{L}_{reg} + \mathcal{L}_{ctr} + \mathcal{L}_{guid}$	0.774	**0.965**	**0.327**

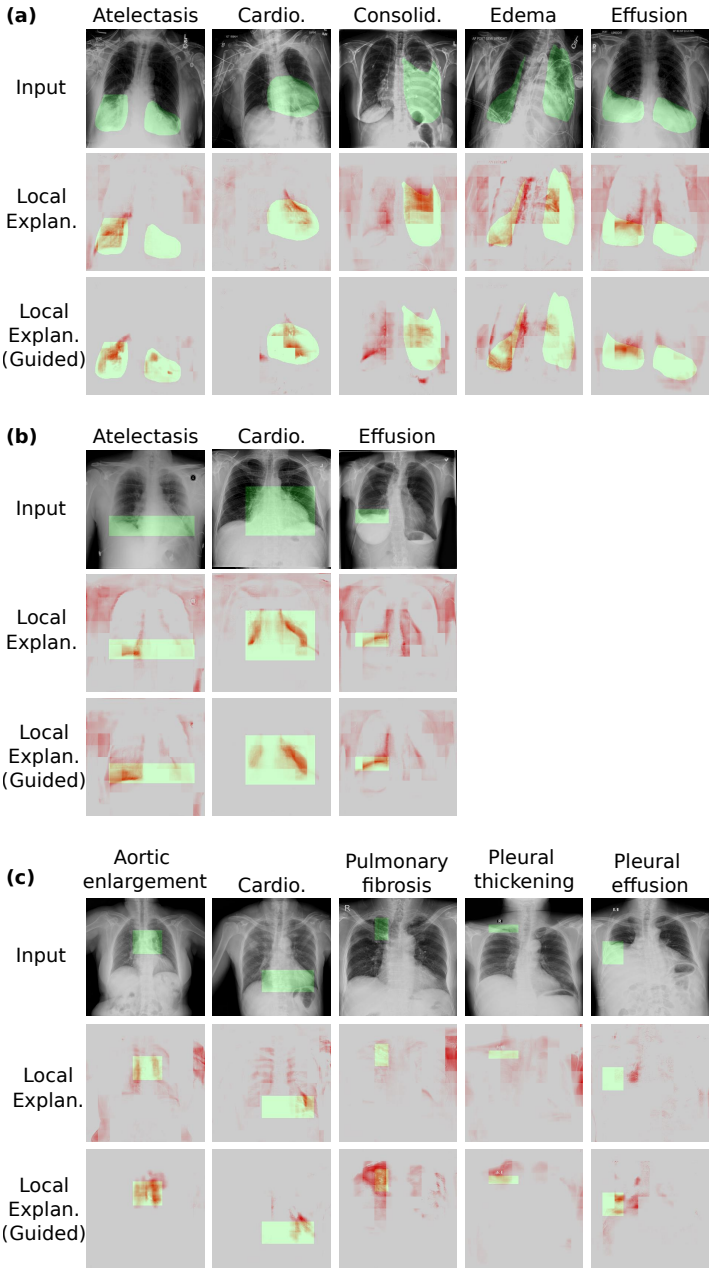


Figure 10: Local explanations produced by Attri-Net trained with and without guidance. Examples from three datasets: **(a)** CheXpert, **(b)** ChestX-ray8, **(c)** Vindr-CXR. Ground truth segmentation masks and bounding boxes are indicated in green regions.

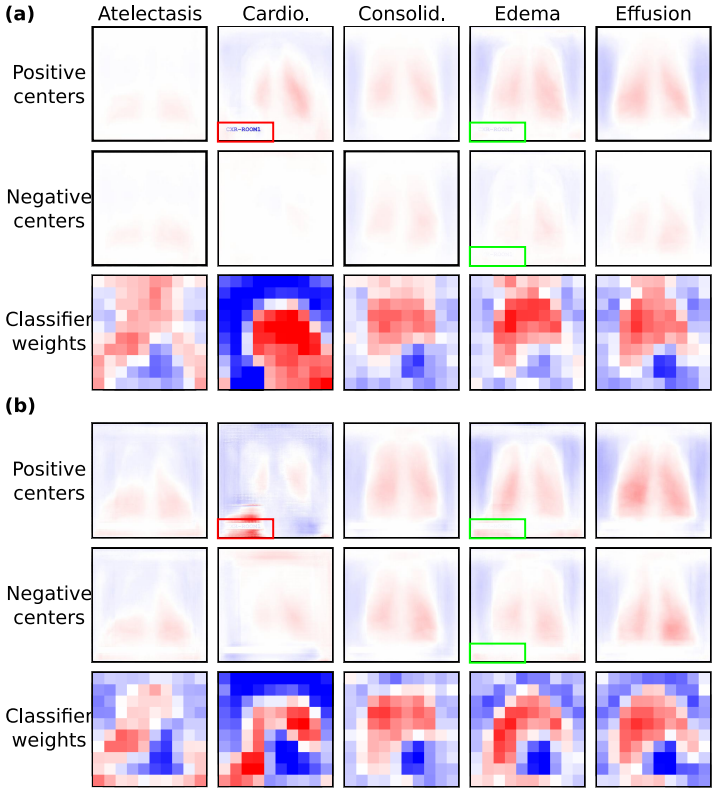


Figure 11: Global explanations produced by Attri-Net trained on the contaminated CheXpert dataset: (a) global explanation of the model trained without guidance, (b) global explanation of the model trained with guidance. Red boxes highlight the spurious signal detected by the class centers of cardiomegaly, and green boxes highlight the spurious signal captured by the class centers of edema.

explanations. Our experiments showed that Attri-Net can produce high-quality local explanations that substantially outperform all baselines regarding class sensitivity and disease sensitivity while retaining classification performance comparable to state-of-the-art black-box models.

Our further experiments showed that explanations of the black-box model can be highly dependent on which post-hoc technique is used, and fundamentally differ from each other even on the same sample. This erodes trust in their capacity to provide faithful explanations in high-stakes applications and shows the need for inherently interpretable models such as ours, where the predictions are formed directly and linearly from visually interpretable class attribution maps.

Apart from providing transparent local explanations

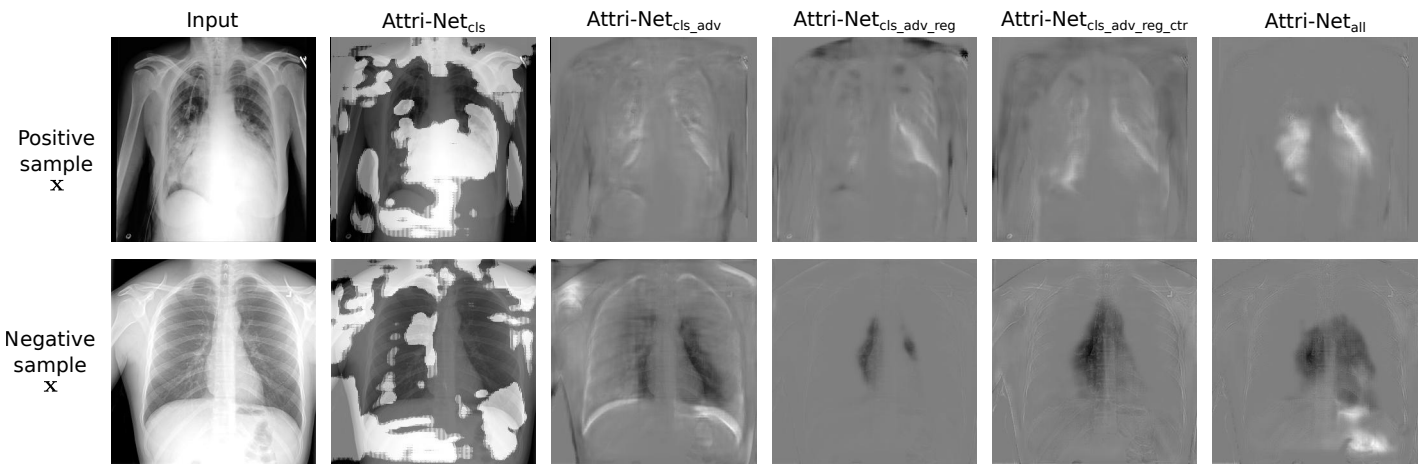


Figure 12: Qualitative results of ablation study. Attribution maps generated by models trained with different subsets of our proposed losses for samples from the ChestX-ray8 dataset with cardiomegaly (top row) and without cardiomegaly (bottom row).

for individual predictions, Attri-Net can provide a global explanation of the model at the dataset level. Notably, the global explanation produced by Attri-Net was shown to faithfully capture synthetically induced shortcut learning behavior, which highlights the potential of our proposed approach to assist ML practitioners in improving the quality of the model and datasets. In addition, we demonstrated that Attri-Net can be combined with model guidance. If human annotations are available this allows to enforce that the correct features are used for making predictions and that the explanations are aligned with human knowledge. Our experiments showed that even very few human annotations can achieve significant improvement in disease localization.

The qualitative and quantitative assessments in this paper suggest that our method provides useful explanations; however, several limitations remain. Specifically, since Attri-Net performs classification and generates explanations by learning the differences between disease-positive and disease-negative classes, it is particularly suited for imaging modalities that capture relatively fixed anatomical structures and lesions. Therefore, detecting tiny tumors or nodules would be a challenging task for Attri-Net, especially when such lesions can appear in various locations. However, since many medical imaging modalities typically focus on specific parts of the human body with relatively fixed anatomical structures, the design concept of Attri-Net can potentially be extended to other imaging modalities, including 3D domains such as CT or MRI. In future work, we will evaluate the suitability of Attri-Net to other medical imaging domains with fixed structures (e.g. fundus imaging) as well as domains with varying structure (e.g. histopathological imaging). Since pixel-level lesion annotation often requires extensive expert effort, we will further explore the trade-off between annotation effort and the benefits of model guidance. Systematically quantifying global explanations remains a challenging task. In this study, the shortcut learning scenarios used for evaluation are limited to systematic artifacts with relatively fixed spatial locations. However, in real-world settings, shortcuts can arise in more diverse and spatially variable patterns. In future work, we aim to investigate more complex shortcut learning conditions and develop improved methods for quantifying global explanations. Finally, we believe it is also crucial to assess the utility of Attri-Net and other explanation methods in human-in-the-loop settings, which is a crucial step toward clinical deployment.

Acknowledgments

This work was funded by the Deutsche Forschungsgemeinschaft (DFG) – EXC number 2064/1 – Project number 390727645 and the Carl Zeiss Foundation in the project "Certification and Foundations of Safe Machine Learning Systems in Healthcare". The authors acknowledge the support of the Hertie Foundation. The authors thank the International Max Planck Research School for Intelligent Systems (IMPRS-IS) for supporting Susu Sun and Stefano Woerner.

Ethical Standards

The work follows appropriate ethical standards in conducting research and writing the manuscript, following all applicable laws and regulations regarding treatment of animals or human subjects.

Conflicts of Interest

We declare we don't have conflicts of interest.

Data availability

We used public datasets for our experiments, and we made the code for reproducing the results publicly available at `https://github.com/ss-sun/Attri-Net-V2`.

References

Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. *Advances in neural information processing systems*, 31, 2018.

David Alvarez Melis and Tommi Jaakkola. Towards robust interpretability with self-explaining neural networks. *Advances in neural information processing systems*, 31, 2018.

Elaine Angelino, Nicholas Larus-Stone, Daniel Alabi, Margo Seltzer, and Cynthia Rudin. Learning certifiably optimal rule lists for categorical data. *Journal of Machine Learning Research*, 18(234):1–78, 2018.

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International conference on machine learning*, pages 214–223. PMLR, 2017.

Nishanth Arun, Nathan Gaw, Praveer Singh, Ken Chang, Mehak Aggarwal, Bryan Chen, Katharina Hoebel, Sharut Gupta, Jay Patel, Mishka Gidwani, et al. Assessing the trustworthiness of saliency maps for localizing abnormalities in medical imaging. *Radiology: Artificial Intelligence*, 3(6):e200267, 2021.

Matan Atad, Vitalii Dmytrenko, Yitong Li, Xinyue Zhang, Matthias Keicher, Jan Kirschke, Bene Wiestler, Ashkan Khakzar, and Nassir Navab. Chexplaining in style: Counterfactual explanations for chest x-rays using stylegan. *arXiv preprint arXiv:2207.07553*, 2022.

Maximilian Augustin, Valentyn Boreiko, Francesco Croce, and Matthias Hein. Diffusion visual counterfactual explanations. *Advances in Neural Information Processing Systems*, 35:364–377, 2022.

Dor Bank, Noam Koenigstein, and Raja Giryes. Autoencoders. *Machine learning for data science handbook: data mining and knowledge discovery handbook*, pages 353–374, 2023.

Alina Jade Barnett, Fides Regina Schwartz, Chaofan Tao, Chaofan Chen, Yinhao Ren, Joseph Y Lo, and Cynthia Rudin. Interpretable mammographic image classification using case-based reasoning and deep learning. *arXiv preprint arXiv:2107.05605*, 2021.

Shahaf Bassan, Guy Amir, and Guy Katz. Local vs. global interpretability: A computational complexity perspective. *arXiv preprint arXiv:2406.02981*, 2024.

Christian F Baumgartner, Konstantinos Kamnitsas, Jacqueline Matthew, Tara P Fletcher, Sandra Smith, Lisa M Koch, Bernhard Kainz, and Daniel Rueckert. Sononet: real-time detection and localisation of fetal standard scan planes in freehand ultrasound. *IEEE transactions on medical imaging*, 36(11):2204–2215, 2017.

Christian F Baumgartner, Lisa M Koch, Kerem Can Tezcan, Jia Xi Ang, and Ender Konukoglu. Visual feature attribution using wasserstein gans. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8309–8319, 2018.

Hasan A Bedel and Tolga Çukur. Dreamr: Diffusion-driven counterfactual explanation for functional mri. *IEEE Transactions on Medical Imaging*, 2024.

Moritz Bohle, Mario Fritz, and Bernt Schiele. Convolutional dynamic alignment networks for interpretable classifications. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10029–10038, 2021.

Moritz Böhle, Navdeeppal Singh, Mario Fritz, and Bernt Schiele. B-cos alignment for inherently interpretable cnns and vision transformers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

Valentyn Boreiko, Indu Ilanchezian, Murat Seçkin Ayhan, Sarah Müller, Lisa M Koch, Hanna Faber, Philipp Berens, and Matthias Hein. Visual explanations for the detection of diabetic retinopathy from retinal fundus images. In *International conference on medical image computing and computer-assisted intervention*, pages 539–549. Springer, 2022.

Wieland Brendel and Matthias Bethge. Approximating cnns with bag-of-local-features models works surprisingly well on imagenet. *arXiv preprint arXiv:1904.00760*, 2019.

Chaofan Chen, Oscar Li, Daniel Tao, Alina Barnett, Cynthia Rudin, and Jonathan K Su. This looks like that: deep learning for interpretable image recognition. *Advances in neural information processing systems*, 32, 2019.

Jianbo Chen, Le Song, Martin J Wainwright, and Michael I Jordan. L-shapley and c-shapley: Efficient model interpretation for structured data. *arXiv preprint arXiv:1808.02610*, 2018.

Zhi Chen, Yijie Bei, and Cynthia Rudin. Concept whitening for interpretable image recognition. *Nature Machine Intelligence*, 2(12):772–782, 2020.

Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8789–8797, 2018.

Molnar Christoph. Interpretable machine learning: A guide for making black box models explainable. 2020.

Joseph Paul Cohen, Rupert Brooks, Sovann En, Evan Zucker, Anuj Pareek, Matthew P Lungren, and Akshay Chaudhari. Gifsplanation via latent shift: a simple autoencoder approach to counterfactual generation for chest x-rays. In *Medical Imaging with Deep Learning*, pages 74–104. PMLR, 2021.

Joseph Paul Cohen, Louis Blankemeier, and Akshay Chaudhari. Identifying spurious correlations using counterfactual alignment, 2025. URL `https://arxiv.org/abs/2312.02186`.

Berkeley J Dietvorst, Joseph P Simmons, and Cade Massey. Algorithm aversion: people erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General*, 144(1):114, 2015.

Arthur Dodek, Donald G Kassebaum, and J David Bristow. Pulmonary edema in coronary-artery disease without cardiomegaly: paradox of the stiff heart. *New England Journal of Medicine*, 286(25):1347–1350, 1972.

Kerol R Djoumessi Donteu, Indu Ilanchezian, Laura Kühlewein, Hanna Faber, Christian F Baumgartner, Bubacarr Bah, Philipp Berens, and Lisa M Koch. Sparse activations for interpretable disease grading. In *Medical Imaging with Deep Learning*, 2023.

Rachel Lea Draelos and Lawrence Carin. Use hirescam instead of grad-cam for faithful explanations of convolutional neural networks. *arXiv preprint arXiv:2011.08891*, 2020.

Gabriel Erion, Joseph D Janizek, Pascal Sturmfels, Scott M Lundberg, and Su-In Lee. Improving performance of deep learning models with axiomatic attribution priors and expected gradients. *Nature machine intelligence*, 3(7): 620–631, 2021.

Nima Fathi, Amar Kumar, Brennan Nichyporuk, Mohammad Havaei, and Tal Arbel. Decodex: Confounder detector guidance for improved diffusion-based counterfactual explanations. *arXiv preprint arXiv:2405.09288*, 2024.

Tanmay Garg, Deepika Vemuri, and Vineeth N Balasubramanian. Advancing ante-hoc explainable models through generative adversarial networks. *arXiv preprint arXiv:2401.04647*, 2024.

Susanne Gaube, Harini Suresh, Martina Raue, Alexander Merritt, Seth J Berkowitz, Eva Lermer, Joseph F Coughlin, John V Guttag, Errol Colak, and Marzyeh Ghassemi. Do as ai say: susceptibility in deployment of clinical decision-aids. *NPJ digital medicine*, 4(1):31, 2021.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

Marzyeh Ghassemi, Luke Oakden-Rayner, and Andrew L Beam. The false hope of current approaches to explainable artificial intelligence in health care. *The Lancet Digital Health*, 3(11):e745–e750, 2021.

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.

Thomas Grote and Philipp Berens. On the ethics of algorithmic decision-making in healthcare. *Journal of medical ethics*, 46(3):205–211, 2020.

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.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 590–597, 2019.

Amir Jamaludin, Timor Kadir, and Andrew Zisserman. Spinenet: automatically pinpointing classification evidence in spinal mris. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 166–175. Springer, 2016.

Guillaume Jeanneret, Loïc Simon, and Frédéric Jurie. Adversarial counterfactual visual explanations. In *Proceedings*

of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 16425–16435, 2023.

Shalmali Joshi, Oluwasanmi Koyejo, Been Kim, and Joydeep Ghosh. xgems: Generating examplars to explain black-box models. *arXiv preprint arXiv:1806.08867*, 2018.

Ashkan Khakzar, Pedram Khorsandi, Rozhin Nobahari, and Nassir Navab. Do explanations explain? model knows best. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10244–10253, 2022.

Been Kim, Rajiv Khanna, and Oluwasanmi O Koyejo. Examples are not enough, learn to criticize! criticism for interpretability. *Advances in neural information processing systems*, 29, 2016.

Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, et al. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *International conference on machine learning*, pages 2668–2677. PMLR, 2018.

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

Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In *International conference on machine learning*, pages 5338–5348. PMLR, 2020.

Himabindu Lakkaraju, Stephen H Bach, and Jure Leskovec. Interpretable decision sets: A joint framework for description and prediction. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1675–1684, 2016.

Zhe Li, Chong Wang, Mei Han, Yuan Xue, Wei Wei, Li-Jia Li, and Li Fei-Fei. Thoracic disease identification and localization with limited supervision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8290–8299, 2018.

Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen Awm Van Der Laak, Bram Van Ginneken, and Clara I Sánchez. A survey on deep learning in medical image analysis. *Medical image analysis*, 42: 60–88, 2017.

Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

Dimitry Mindlin, Malte Schilling, and Philipp Cimiano. Abc-gan: Spatially constrained counterfactual generation for image classification explanations. In *World Conference on Explainable Artificial Intelligence*, pages 260–282. Springer, 2023.

Ha Q Nguyen, Khanh Lam, Linh T Le, Hieu H Pham, Dat Q Tran, Dung B Nguyen, Dung D Le, Chi M Pham, Hang TT Tong, Diep H Dinh, et al. Vindr-cxr: An open dataset of chest x-rays with radiologist's annotations. *Scientific Data*, 9(1):429, 2022.

Sérgio Pereira, Raphael Meier, Richard McKinley, Roland Wiest, Victor Alves, Carlos A Silva, and Mauricio Reyes. Enhancing interpretability of automatically extracted machine learning features: application to a rbm-random forest system on brain lesion segmentation. *Medical image analysis*, 44:228–244, 2018.

Gustav Grund Pihlgren and Kary Främling. Segmentation and smoothing affect explanation quality more than the choice of perturbation-based xai method for image explanations, 2025. URL `https://arxiv.org/abs/2409.04116`.

Vipin Pillai, Soroush Abbasi Koohpayegani, Ashley Ouligian, Dennis Fong, and Hamed Pirsiavash. Consistent explanations by contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10213–10222, 2022.

Pedro O Pinheiro and Ronan Collobert. From image-level to pixel-level labeling with convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1713–1721, 2015.

Xuyin Qi, Zeyu Zhang, Aaron Berliano Handoko, Huazhan Zheng, Mingxi Chen, Ta Duc Huy, Vu Minh Hieu Phan, Lei Zhang, Linqi Cheng, Shiyu Jiang, et al. Projectedex: Enhancing generation in explainable ai for prostate cancer. *arXiv preprint arXiv:2501.01392*, 2025.

Sukrut Rao, Moritz Böhle, Amin Parchami-Araghi, and Bernt Schiele. Studying how to efficiently and effectively guide models with explanations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1922–1933, 2023.

Mauricio Reyes, Raphael Meier, Sérgio Pereira, Carlos A Silva, Fried-Michael Dahlweid, Hendrik von Tengg-Kobligk, Ronald M Summers, and Roland Wiest. On the interpretability of artificial intelligence in radiology: challenges and opportunities. *Radiology: artificial intelligence*, 2(3):e190043, 2020.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016.

Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence*, 1(5):206–215, 2019.

Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv preprint arXiv:1911.08731*, 2019.

Adriel Saporta, Xiaotong Gui, Ashwin Agrawal, Anuj Pareek, Steven QH Truong, Chanh DT Nguyen, Van-Doan Ngo, Jayne Seekins, Francis G Blankenberg, Andrew Y Ng, et al. Benchmarking saliency methods for chest x-ray interpretation. *Nature Machine Intelligence*, 4(10):867–878, 2022.

Kathryn Schutte, Olivier Moindrot, Paul Hérent, Jean-Baptiste Schiratti, and Simon Jégou. Using stylegan for visual interpretability of deep learning models on medical images. *arXiv preprint arXiv:2101.07563*, 2021.

Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. In *International conference on machine learning*, pages 3145–3153. PMLR, 2017.

Sumedha Singla, Brian Pollack, Junxiang Chen, and Kayhan Batmanghelich. Explanation by progressive exaggeration. *arXiv preprint arXiv:1911.00483*, 2019.

Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. *Advances in neural information processing systems*, 30, 2017.

Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin Riedmiller. Striving for simplicity: The all convolutional net. *arXiv preprint arXiv:1412.6806*, 2014.

Guolei Sun, Thomas Probst, Danda Pani Paudel, Nikola Popović, Menelaos Kanakis, Jagruti Patel, Dengxin Dai, and Luc Van Gool. Task switching network for multi-task learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8291–8300, 2021.

Susu Sun, Lisa M Koch, and Christian F Baumgartner. Right for the wrong reason: Can interpretable ml techniques detect spurious correlations? In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 425–434. Springer, 2023a.

Susu Sun, Stefano Woerner, Andreas Maier, Lisa M Koch, and Christian F Baumgartner. Inherently interpretable multi-label classification using class-specific counterfactuals. *arXiv preprint arXiv:2303.00500*, 2023b.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International conference on machine learning*, pages 3319–3328. PMLR, 2017.

Philipp Tschandl, Christoph Rinner, Zoe Apalla, Giuseppe Argenziano, Noel Codella, Allan Halpern, Monika Janda, Aimilios Lallas, Caterina Longo, Josep Malvehy, et al. Human–computer collaboration for skin cancer recognition. *Nature Medicine*, 26(8):1229–1234, 2020.

Haofan Wang, Zifan Wang, Mengnan Du, Fan Yang, Zijian Zhang, Sirui Ding, Piotr Mardziel, and Xia Hu. Score-cam: Score-weighted visual explanations for convolutional neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 24–25, 2020.

Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2097–2106, 2017.

Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao. A discriminative feature learning approach for deep face recognition. In *Computer vision–ECCV 2016: 14th European conference, amsterdam, the netherlands, October 11–14, 2016, proceedings, part VII 14*, pages 499–515. Springer, 2016.

Wenwu Ye, Jin Yao, Hui Xue, and Yi Li. Weakly supervised lesion localization with probabilistic-cam pooling. *arXiv preprint arXiv:2005.14480*, 2020.

Zhuoning Yuan, Yan Yan, Milan Sonka, and Tianbao Yang. Large-scale robust deep auc maximization: A new surrogate loss and empirical studies on medical image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3040–3049, 2021.

Yujia Zhang, Kuangyan Song, Yiming Sun, Sarah Tan, and Madeleine Udell. "why should you trust my explanation?" understanding uncertainty in lime explanations. *arXiv preprint arXiv:1904.12991*, 2019.

Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929, 2016.

Wentao Zhu, Qi Lou, Yeeleng Scott Vang, and Xiaohui Xie. Deep multi-instance networks with sparse label assignment for whole mammogram classification. In *Medical Image Computing and Computer Assisted Intervention-MICCAI 2017: 20th International Conference, Quebec City, QC, Canada, September 11-13, 2017, Proceedings, Part III 20*, pages 603–611. Springer, 2017.

Appendix A. Additional examples of counterfactual attribution maps

Examples of counterfactual images obtained by adding the class-specific attribution map to the input image, i.e. $\hat{\mathbf{x}} = \mathbf{x} + M_c(\mathbf{x})$, are shown in Fig. 13.

Appendix B. Additional results of classification

Table 5 shows the disease-wise classification performance of all compared models for reference.

Appendix C. Ablation study of the model guidance

We performed an ablation study of model guidance. As we discuss in the main paper, there are various approaches to generate pseudo guidance from the limited number of ground truth pixel-wise annotations of disease. We evaluated model guidance with the pseudo bounding box, pseudo-binary mask, and weighted pseudo mask and found that the pseudo binary mask perform best among the three kinds of pseudo guidance. We show the pseudo masks for three datasets and the result in the main paper. Here, we show the weighted pseudo masks (see Fig. 14 a, b, c)and pseudo bounding box (see Fig. 14 d, e, f) as an additional reference. It can be clearly observed that the weighted pseudo masks have very strong focuses while the pseudo bounding boxes cover much larger regions.

With Table 6 and Fig. 15, we show the quantitative and qualitative results of using different model guidance strategies. We have the following observations.

First, the classification performance and class sensitivity do not change much with different model guidance approaches. However, adding guidance to the model significantly improved the disease sensitivity, indicating that the model focuses more on the correct disease-relevant region.

Second, the disease sensitivity score and the model guidance effect vary a lot between different diseases. For diseases with large lesion regions, such as cardiomegaly, the disease sensitivity score and the improvement of the score after adding guidance are much larger than those with smaller lesion regions, such as atelectasis.

Third, the model guidance approach affects disease localization a lot. The "box only" model is guided only using very limited pixel-wise annotations. The 30% improvement in disease sensitivity of the "box only" model compared with the model "without guidance" shows that even very few annotations can achieve nice guidance. Since chest X-ray images have relative fixed structures, the model "pseudo bbox" trained only with pseudo bounding boxes achieved better performance than the model using only ground truth annotation. The improvement of the model "pseudo mask" which is trained only with pseudo masks compared with model "pseudo bbox" shows that more focused guidance have better effects. The "mixed" model is trained using the strategy we describe in the main paper, i.e., for samples with ground truth annotations, use ground truth as guidance, otherwise, using pseudo masks as guidance. The best disease sensitive score achieved by the "mixed" model leads us to the conclusion that we should make full use of the available limited ground truth annotations. Since the pseudo masks are not ground truth annotation, therefore, model "mixed weighted mask" that guided by more focused weighted pseudo masks does not achieve the best performance in disease sensitivity.

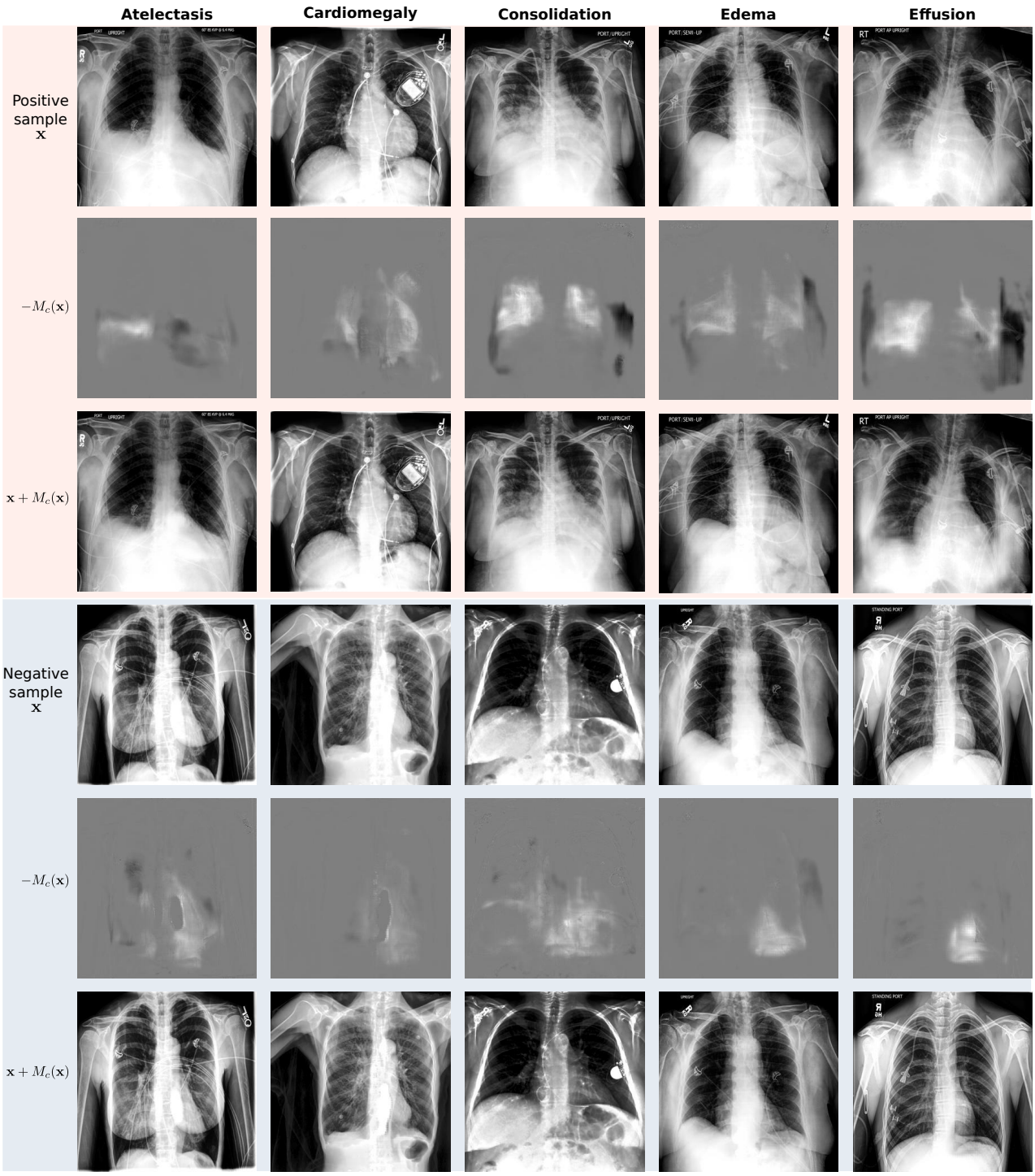


Figure 13: Counterfactual image generation. Samples from the CheXpert dataset. The examples in the top group of rows (in red) show input images that are positive for different classes c. For those, the evidence for specific disease c is strong, as shown in attribution maps $M_c(\mathbf{x})$. After adding $M_c(\mathbf{x})$ with inputs, the strong disease effects are removed, as shown in the counterfactuals in the third row. The bottom group of rows (in blue) shows images that are negative for class c. For these images, the disease effects in attribution maps are light, and those images remain mostly unchanged by adding the output of $M_c(\mathbf{x})$.

Table 5: Classification performance for each disease, measured by area under the ROC curve (AUC).

	CheXpert					
Model	**Atelectasis**	**Cardio.**	**Consolid.**	**Edema**	**Effusion**	**Avg.**
ResNet50	0.782	0.902	0.867	0.877	0.950	0.875
B-cos ResNet50	0.791	0.909	0.846	0.855	0.928	0.866
B-cos ResNet50 (guided)	0.719	0.904	0.820	0.840	0.915	0.839
ours	0.807	0.914	0.870	0.841	0.935	0.873
ours (guided)	0.763	0.852	0.870	0.839	0.917	0.848
	ChestX-ray8					
Model	**Atelectasis**	**Cardio.**	**Consolid.**	**Edema**	**Effusion**	**Avg.**
ResNet50	0.723	0.857	0.708	0.804	0.799	0.778
B-cos ResNet50	0.697	0.825	0.694	0.802	0.766	0.757
B-cos ResNet50 (guided)	0.692	0.826	0.696	0.794	0.760	0.754
ours	0.713	0.858	0.711	0.823	0.792	0.779
ours (guided)	0.718	0.847	0.699	0.810	0.794	0.774
	Vindr-CXR					
Model	**Aortic enlarg.**	**Cardio.**	**Pulmon. fib.**	**Pleu. thicken.**	**Pleu. Effusion**	**Avg.**
ResNet50	0.795	0.827	0.685	0.720	0.794	0.764
B-cos ResNet50	0.877	0.942	0.736	0.781	0.842	0.836
B-cos ResNet50 (guided)	0.876	0.926	0.726	0.773	0.840	0.828
ours	0.758	0.872	0.723	0.753	0.839	0.789
ours (guided)	0.812	0.860	0.702	0.743	0.793	0.782

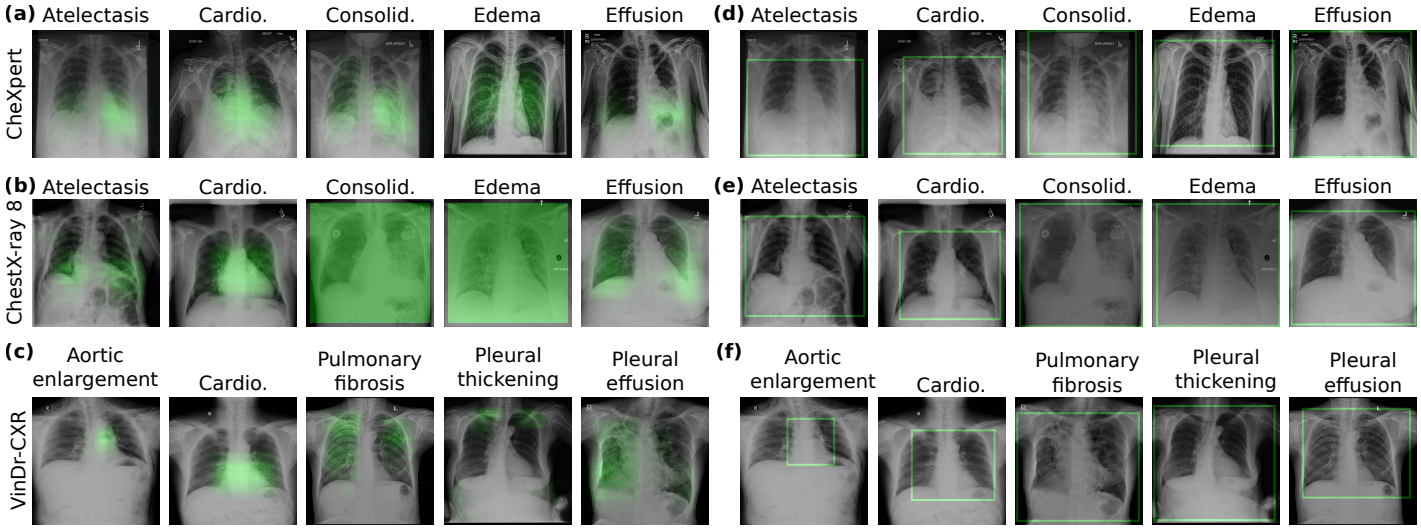


Figure 14: Weighted pseudo masks and pseudo bounding box.

Table 6: Ablation study on model guidance. Evaluated on ChestX-ray8 dataset.

Guidance	Classification AUC	Class sensitivity	Disease sensitivity average over disease	Disease sensitivity Atelectasis	Disease sensitivity Cardiomegaly	Disease sensitivity Effusion
without guidance	0.779	0.951	0.158	0.047	0.343	0.084
bbox only	**0.782**	0.942	0.201	0.067	0.447	0.087
pseudo bbox	0.775	0.950	0.249	0.069	0.582	0.097
pseudo mask	0.775	0.944	0.313	**0.139**	0.689	0.110
mixed	0.774	**0.965**	**0.327**	0.120	0.733	**0.129**
mixed weighted mask	0.780	0.948	0.305	0.071	**0.753**	0.092

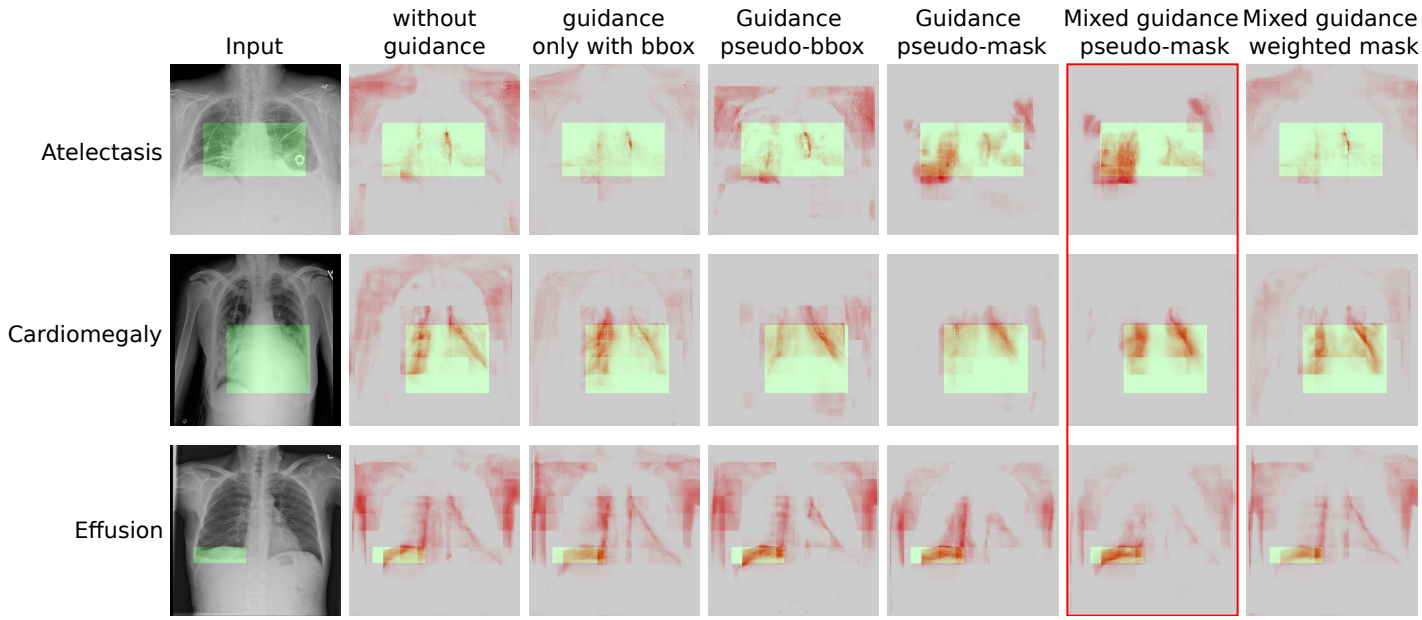


Figure 15: Local explanations from models trained with different model guidance approaches. Samples are from ChestX-ray8 dataset. The red box highlights the explanations from the best-performed "mixed" model that is trained with both ground truth annotation and pseudo binary masks.

Appendix D. Additional qualitative results of local explanation

Fig. 16 and Fig. 17 contain additional examples of local explanations using all compared methods derived from the CheXpert and Vindr-CXR datasets, respectively.

Appendix E. Additional quantitative results of local explanation.

We present disease-specific quantitative results, including 95% confidence intervals for class sensitivity and disease sensitivity, across all three datasets in the Tables 7 and 8 for further reference.

As discussed in Appendix C, disease sensitivity scores vary greatly between different diseases. On three datasets, the disease sensitivity scores of cardiomegaly are much higher than those scores of other diseases. The improvement of disease sensitivity in cardiomegaly after adding guidance is also much bigger. This can be explained by the relatively large and fixed lesion region of the disease cardiomegaly. The big improvement after adding model guidance also happens with disease arotic enlargement, which has a quite focused lesion region which can be seen from Fig. 15. Notably, on three datasets, the disease sensitivity scores of pleural effusion do not improve much after adding guidance. By comparing the weighted pseudo-masks of effusion on three datasets in Fig. 15, it is clear that the lesion regions of effusion vary much between different datasets. Therefore, the pseudo guidance mechanism does not perform well in this disease.

Appendix F. Implementation details

F.1 Discriminator training

The Attri-Net framework requires training a discriminator function D in parallel to the class attribution generator M. The weight parameters θ of the discriminator are computed in separate gradient update steps using the Wasserstein GAN objective. The full discriminator optimisation objective is then given by

$$\min_{\theta} \sum_{c} \mathop{\mathbb{E}}_{\mathbf{x}\sim p(\mathbf{x}|y_c=0)} [D_c(\mathbf{x}|\theta)] + \mathop{\mathbb{E}}_{\mathbf{x}\sim p(\mathbf{x}|y_c=1)} [D_c(\mathbf{x}+M_c(\mathbf{x})|\theta)] \,,$$

where we omitted the gradient penalty loss which ensures the discriminator fulfills the Lipschitz-1 constraint dictated by the Wasserstein GAN objective.

F.2 Network architectures

The network architecture of the attribution map generator and the discriminator of the Attri-Net framework are shown in Table 9 and Table. 10, respectively. L refers to the length of input/output features, N is the number of output channels, and K is the kernel size.

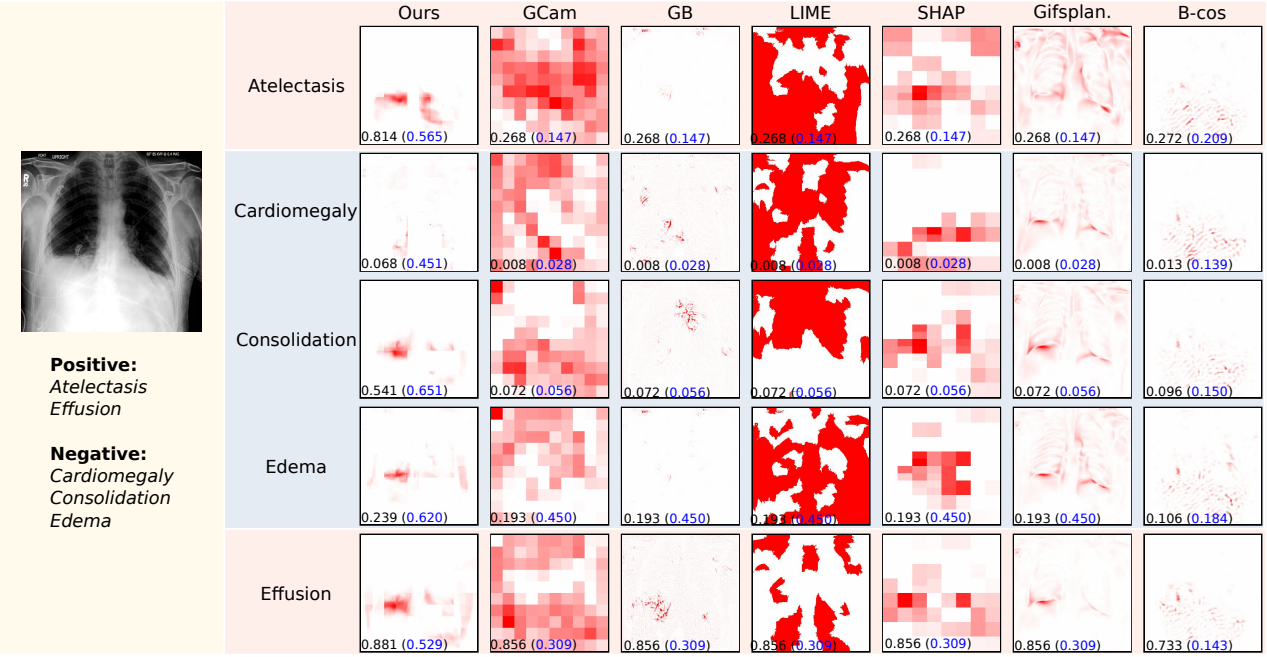


Figure 16: Local explanations for an example image from the CheXpert dataset.

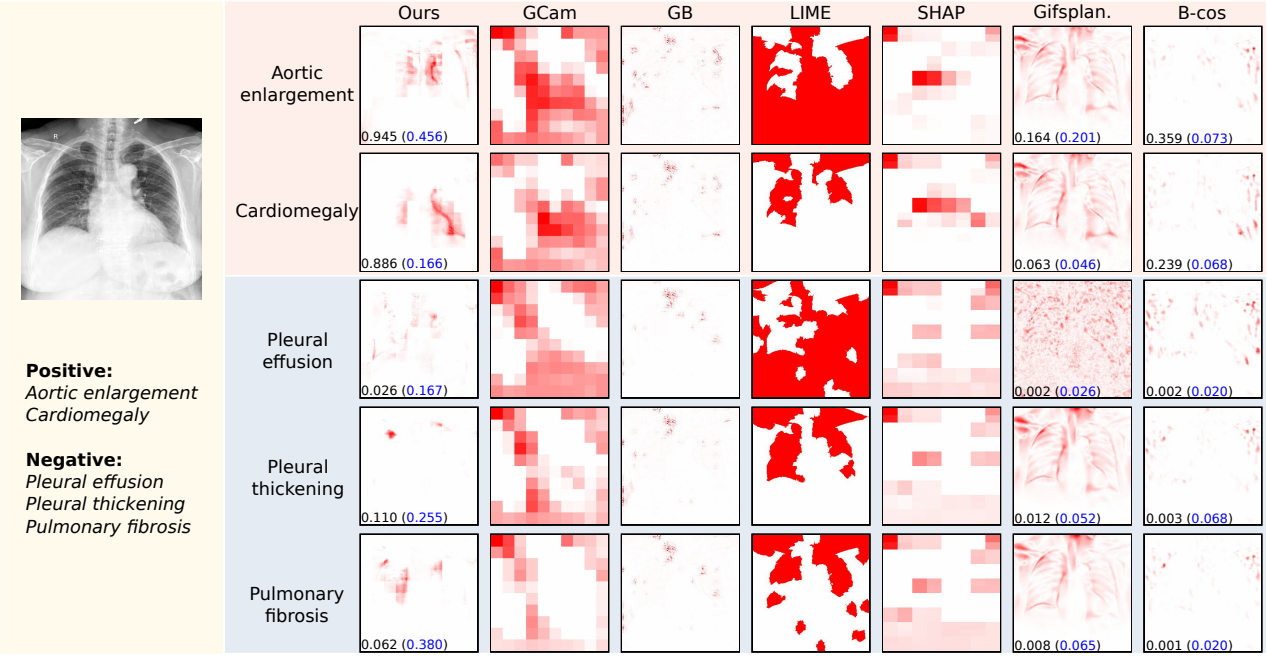


Figure 17: Local explanations for an example image from the Vindr-CXR dataset.

Table 7: Comparison of Class Sensitivity.

Model	Atelectasis	Cardio.	CheXpert Consolid.	Edema	Effusion
GB	0.267 ± 0.197	0.376 ± 0.213	0.329 ± 0.211	0.262 ± 0.126	0.293 ± 0.249
GCam	0.177 ± 0.198	0.189 ± 0.209	0.245 ± 0.213	0.182 ± 0.144	0.160 ± 0.216
LIME	0.248 ± 0.147	0.254 ± 0.132	0.263 ± 0.046	0.244 ± 0.049	0.256 ± 0.165
SHAP	0.290 ± 0.160	0.325 ± 0.162	0.319 ± 0.116	0.382 ± 0.132	0.431 ± 0.181
Gifsp.	0.259 ± 0.273	0.497 ± 0.593	0.461 ± 0.390	0.167 ± 0.271	0.108 ± 0.269
B-cos ResNet50	0.230 ± 0.200	0.360 ± 0.207	0.211 ± 0.091	0.230 ± 0.103	0.294 ± 0.140
B-cos ResNet50 (guided)	0.223 ± 0.134	0.320 ± 0.115	0.286 ± 0.123	0.249 ± 0.110	0.286 ± 0.204
ours	**0.533** ± 0.353	**0.743** ± 0.320	**0.663** ± 0.239	**0.715** ± 0.266	**0.797** ± 0.226
ours (pseudo guidance)	0.499 ± 0.186	0.659 ± 0.241	0.533 ± 0.270	0.588 ± 0.230	0.796 ± 0.238
ours (full guidance)	-	-	-	-	-
Model	**Atelectasis**	**Cardio.**	**ChestX-ray8 Consolid.**	**Edema**	**Effusion**
GB	0.302 ± 0.123	0.266 ± 0.129	0.228 ± 0.086	0.310 ± 0.100	0.205 ± 0.088
GCam	0.205 ± 0.127	0.226 ± 0.103	0.232 ± 0.145	0.219 ± 0.136	0.249 ± 0.143
LIME	0.231 ± 0.065	0.207 ± 0.075	0.227 ± 0.059	0.247 ± 0.113	0.226 ± 0.075
SHAP	0.453 ± 0.078	0.483 ± 0.083	0.395 ± 0.097	0.367 ± 0.104	0.468 ± 0.095
Gifsp.	0.511 ± 0.555	0.830 ± 0.549	0.696 ± 0.608	0.785 ± 0.585	0.520 ± 0.690
B-cos ResNet50	0.308 ± 0.163	0.263 ± 0.147	0.280 ± 0.140	0.253 ± 0.144	0.277 ± 0.158
B-cos ResNet50 (guided)	0.325 ± 0.202	0.282 ± 0.204	0.362 ± 0.189	0.360 ± 0.182	0.278 ± 0.162
ours	0.925 ± 0.045	0.959 ± 0.024	0.944 ± 0.024	0.954 ± 0.023	**0.972** ± 0.028
ours (pseudo guidance)	**0.954** ± 0.021	**0.971** ± 0.019	**0.958** ± 0.029	**0.972** ± 0.020	0.968 ± 0.013
ours (full guidance)	-	-	-	-	-
Model	**Aortic enlarg.**	**Cardio.**	**Vindr-CXR Pulmon. fib.**	**Pleu. thicken.**	**Pleu. Effusion**
GB	0.236 ± 0.140	0.250 ± 0.135	0.274 ± 0.112	0.296 ± 0.139	0.268 ± 0.139
GCam	0.175 ± 0.105	0.174 ± 0.126	0.160 ± 0.140	0.152 ± 0.132	0.219 ± 0.074
LIME	0.255 ± 0.047	0.230 ± 0.144	0.267 ± 0.087	0.258 ± 0.106	0.245 ± 0.063
SHAP	0.362 ± 0.137	0.350 ± 0.106	0.254 ± 0.074	0.270 ± 0.091	0.288 ± 0.083
Gifsp.	0.217 ± 0.502	0.170 ± 0.444	0.297 ± 0.411	0.270 ± 0.398	0.672 ± 0.650
B-cos ResNet50	0.251 ± 0.137	0.271 ± 0.145	0.217 ± 0.122	0.227 ± 0.118	0.242 ± 0.122
B-cos ResNet50 (guided)	0.252 ± 0.154	0.230 ± 0.117	0.241 ± 0.147	0.218 ± 0.142	0.248 ± 0.103
ours	0.885 ± 0.130	0.883 ± 0.146	0.850 ± 0.210	0.866 ± 0.170	0.821 ± 0.290
ours (pseudo guidance)	**0.938** ± 0.079	0.912 ± 0.145	**0.918** ± 0.079	**0.908** ± 0.109	**0.933** ± 0.087
ours (full guidance)	0.885 ± 0.132	**0.914** ± 0.114	0.818 ± 0.258	0.889 ± 0.088	0.856 ± 0.195

Table 8: Comparison of Disease Sensitivity.

			CheXpert		
Model	**Atelectasis**	**Cardio.**	**Consolid.**	**Edema**	**Effusion**
GB	0.056 ± 0.071	0.255 ± 0.176	0.207 ± 0.297	0.210 ± 0.224	0.149 ± 0.229
GCam	0.102 ± 0.135	0.343 ± 0.371	0.141 ± 0.212	0.264 ± 0.240	0.120 ± 0.237
LIME	0.115 ± 0.128	0.099 ± 0.088	0.032 ± 0.078	0.163 ± 0.154	0.102 ± 0.190
SHAP	0.115 ± 0.139	0.236 ± 0.244	0.177 ± 0.243	0.404 ± 0.284	0.165 ± 0.317
Gifsp.	0.148 ± 0.180	0.190 ± 0.163	0.175 ± 0.190	0.301 ± 0.213	0.144 ± 0.240
B-cos ResNet50	0.187 ± 0.221	0.414 ± 0.238	0.211 ± 0.230	0.273 ± 0.243	0.211 ± 0.331
B-cos ResNet50 (guided)	0.232 ± 0.251	0.480 ± 0.240	0.226 ± 0.209	0.239 ± 0.257	**0.217** ± 0.341
ours	0.199 ± 0.341	0.223 ± 0.248	0.165 ± 0.225	0.353 ± 0.277	0.093 ± 0.246
ours (pseudo guidance)	**0.374** ± 0.451	**0.623** ± 0.356	**0.324** ± 0.414	**0.490** ± 0.329	0.194 ± 0.418
ours (full guidance)	-	-	-	-	-
			ChestX-ray8		
Model	**Atelectasis**	**Cardio.**	**Consolid.**	**Edema**	**Effusion**
GB	0.093 ± 0.205	0.345 ± 0.231	-	-	0.089 ± 0.169
GCam	0.052 ± 0.141	0.248 ± 0.157	-	-	0.077 ± 0.154
LIME	0.049 ± 0.116	0.237 ± 0.118	-	-	0.080 ± 0.132
SHAP	0.098 ± 0.217	0.598 ± 0.255	-	-	**0.139** ± 0.247
Gifsp.	0.084 ± 0.191	0.319 ± 0.239	-	-	0.121 ± 0.187
B-cos ResNet50	0.098 ± 0.227	0.480 ± 0.243	-	-	0.127 ± 0.208
B-cos ResNet50 (guided)	0.098 ± 0.244	0.518 ± 0.215	-	-	0.125 ± 0.223
ours	0.047 ± 0.112	0.343 ± 0.251	-	-	0.084 ± 0.161
ours (pseudo guidance)	**0.120** ± 0.260	**0.733** ± 0.351	-	-	0.129 ± 0.257
ours (full guidance)	-	-	-	-	-
			Vindr-CXR		
Model	**Aortic enlarg.**	**Cardio.**	**Pulmon. fib.**	**Pleu. thicken.**	**Pleu. Effusion**
GB	0.041 ± 0.097	0.089 ± 0.161	0.026 ± 0.089	0.014 ± 0.072	0.065 ± 0.160
GCam	0.036 ± 0.075	0.221 ± 0.179	0.017 ± 0.054	0.006 ± 0.020	0.028 ± 0.090
LIME	0.050 ± 0.052	0.034 ± 0.077	0.022 ± 0.053	0.007 ± 0.024	0.039 ± 0.091
SHAP	0.105 ± 0.160	0.168 ± 0.150	0.006 ± 0.022	0.005 ± 0.016	0.050 ± 0.067
Gifsp.	0.071 ± 0.090	0.127 ± 0.113	0.024 ± 0.057	0.010 ± 0.022	0.029 ± 0.077
B-cos ResNet50	0.119 ± 0.163	0.250 ± 0.196	0.012 ± 0.048	0.009 ± 0.030	0.056 ± 0.133
B-cos ResNet50 (guided)	0.120 ± 0.181	0.190 ± 0.191	0.012 ± 0.055	0.012 ± 0.032	0.042 ± 0.115
ours	0.119 ± 0.180	0.196 ± 0.164	0.017 ± 0.049	0.010 ± 0.030	0.034 ± 0.129
ours (pseudo guidance)	0.270 ± 0.433	0.345 ± 0.314	**0.053** ± 0.178	**0.035** ± 0.127	0.077 ± 0.231
ours (full guidance)	**0.363** ± 0.560	**0.504** ± 0.466	0.031 ± 0.115	0.022 ± 0.056	**0.102** ± 0.282

Table 9: Attri-Net class attribution generator network architecture.

Layers	Input → Output	Layer information
Task embedding layer	Task code $\mathbf{t}_c \rightarrow$ Task embedding $\mathbf{t}'_c$	8 × FC(L100,L100)
Down-sampling	(Input image $\mathbf{x}$, $\mathbf{t}'_c$) $\rightarrow \mathbf{out}_{\text{down}}$	Ada_Conv: CONV(N64, K7x7), AdaIN, ReLU Ada_Conv: CONV(N128, K4x4), AdaIN, ReLU Ada_Conv: CONV(N256, K4x4), AdaIN, ReLU
Bottlenecks	$(\mathbf{out}_{\text{down}}$, $\mathbf{t}'_c) \rightarrow \mathbf{out}_{\text{bn}}$	Ada_ResBlock: CONV(N256, K3x3), AdaIN, ReLU Ada_ResBlock: CONV(N256, K3x3), AdaIN, ReLU Ada_ResBlock: CONV(N256, K3x3,), AdaIN, ReLU Ada_ResBlock: CONV(N256, K3x3), AdaIN, ReLU Ada_ResBlock: CONV(N256, K3x3), AdaIN, ReLU Ada_ResBlock: CONV(N256, K3x3), AdaIN, ReLU
Up-sampling	$(\mathbf{out}_{\text{bn}}$, $\mathbf{t}'_c) \rightarrow \mathbf{out}_{\text{up}}$	Ada_DECONV(N128, K4x4), AdaIN, ReLU Ada_DECONV(N64, K4x4), AdaIN, ReLU CONV(N1, K7x7)
Output layer	$(\mathbf{x}, \mathbf{out}_{\text{up}}) \rightarrow M_c(\mathbf{x})$	$M_c(\mathbf{x}) = \tanh(\mathbf{x} + \mathbf{out}_{\text{up}}) - \mathbf{x}$

Table 10: Attri-Net discriminator network architecture.

Layers	Input → Output	Layer information
Task embedding layer	Task code $\mathbf{t}_c \rightarrow$ Task embedding $\mathbf{t}'_c$	8 × FC(L100,L100)
Input layer		Ada_Conv: CONV(N64, K4x4), AdaIN, ReLU
Hidden layers	$(\mathbf{x}/\hat{\mathbf{x}}$, $\mathbf{t}'_c) \rightarrow \mathbf{out}_{\text{hid}}$	Ada_Conv: CONV(N128, K4x4), AdaIN, ReLU Ada_Conv: CONV(N256, K4x4), AdaIN, ReLU Ada_Conv: CONV(N512, K4x4), AdaIN, ReLU Ada_Conv: CONV(N1024, K4x4), AdaIN, ReLU Ada_Conv: CONV(N2048, K4x4), AdaIN, ReLU
Output layer	$\mathbf{out}_{\text{hid}} \rightarrow \mathcal{L}^{(c)}_{\text{adv}}$	CONV(N1, K3x3)