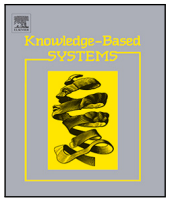
Knowledge-Based Systems 299 (2024) 111993

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys



# Multi-attentional causal intervention networks for medical image diagnosis

Shanshan Huang, Lei Wang, Jun Liao, Li Liu *

*School of Big Data & Software Engineering, Chongqing University, Chongqing, 400044, China*

ARTICLE INFO

*Keywords:*
Causal inference
Causal intervention
Medical image classification
Weakly supervised learning
Multi-head attention

ABSTRACT

Medical image diagnosis has developed rapidly under the impetus of the deep network. Previous works mainly focus on improving the diagnostic accuracy of models, i.e., first use a backbone network to extract image global features and then feed it into the classifier for diagnosis. However, these methods do not fully explore the transparent and reasonable decision-making process of the final classification results, which is crucial for medical diagnosis. In this paper, we propose a framework called Causal Intervention-based Multi-head Attention network (CaIMA) to enhance the explainability of medical diagnosis from a causal inference perspective, by exploring the inherent causal relationship between multi-region attention and diagnosis results. Specifically, it consists of three key components: (1) The multi-region attention module enables the network to focus on the distinct discriminative lesion regions that hold causal relationships with the predicted outcome. (2) The attention-driven data augmentation module provides accurate localization of discriminative regions and enhances model explainability. (3) The causal intervention module aims to explore the intrinsic causal relationship between the attention map and the predicted outcome, encouraging the network to learn more useful attention maps for medical image diagnosis. Besides, to address the learning difficulty of this network, we further introduce a non-overlapping multiple attentional guidance loss that encourages the learned multiple attention maps to focus on specific lesion regions without overlapping. We compare the proposed CaIMA with state-of-the-art methods on multimedia medical datasets, including three public medical image datasets (Kvasir, ISIC2018, COVID-19) and one private dataset (CLC), and the experimental results substantiate the effectiveness of CaIMA in terms of diagnosis accuracy and explainability.

## 1. Introduction

Medical image classification plays a pivotal role in modern healthcare, aiding clinicians in diagnosing ailments and making informed decisions. Previous work [1–3] often presents as black-box models, which only output the diagnostic results, but do not provide information on which factors guide the network to make the final prediction. Explainability and transparency are extremely significant in disease diagnosis because a patient requires not only an accurate prediction but also an explanation of the disease as well as a reasonable way for intervention (or treatment) on possible attributes (i.e., causes). Besides, due to the intra-class diversity and inter-class similarity of medical images, medical image diagnosis can be viewed as a fine-grained classification problem. That is, to distinguish between similar but less different classes in medical images. Medical images of different diseases may be very similar in overall appearance, but there are subtle differences in local features, which requires the model to focus on these local details to achieve an accurate and explainable diagnosis. Take the classification of gastrointestinal endoscopic images as an example, many images look very similar and can only be distinguished by some small local differences (e.g., the presence or absence of polyps), as shown in Fig. 1(a). Fortunately, the attention mechanism [4–8] provides a novel solution to address medical diagnosis in fine-grained classification by focusing on regions in medical images that are relevant to diagnosis results. However, the plain attention mechanisms-based methods consider the correlation between the attention map and the diagnostic results that only reflect their co-occurrence relationship under the assumption of independence, but cannot explain the underlying causal mechanisms due to the spurious correlations caused by confounders. As shown in Fig. 1(a)–(d), the majority of images within the "dyed-lifted-polyps" category in the Kvasir dataset have an endothelial background, and previous works may treat spurious correlation regions (e.g., endothelium (red box)) as discriminatory regions while ignoring regions causally related to the diagnostic result (i.e., polypectomy regions (white box)). These spurious correlations between attention maps and the diagnostic results lead to unexplainable and unreliable decision-making. In other words, these methods fail to focus on the intrinsic causal representations (attention maps that are causally related to the classification results) of the different lesions, making disease

* Corresponding author.
*E-mail addresses:* shanshanhuang@cqu.edu.cn (S. Huang), leiwangtt@stu.cqu.edu.cn (L. Wang), liaojun@cqu.edu.cn (J. Liao), dcsliuli@cqu.edu.cn (L. Liu).

https://doi.org/10.1016/j.knosys.2024.111993
Received 4 March 2024; Received in revised form 25 April 2024; Accepted 22 May 2024
Available online 24 May 2024

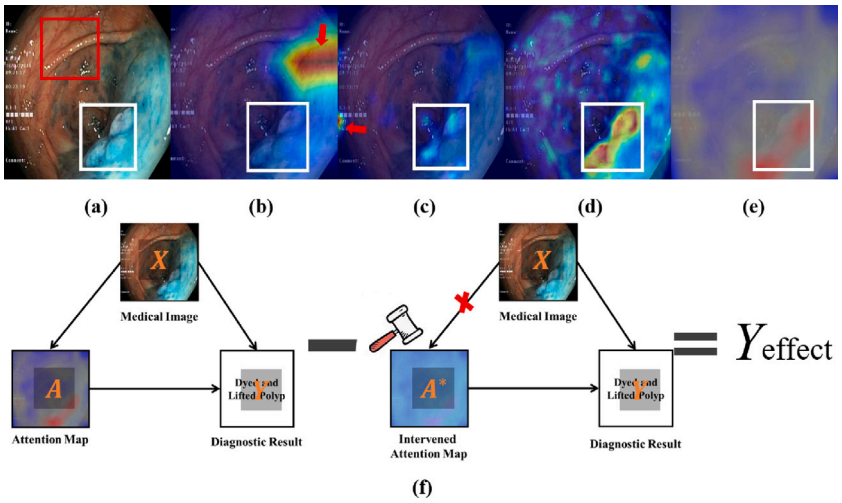



**Fig. 1.** The Visualization of Grad-CAM and causal intervention operation. We show the Grad-CAM visualization of (a) Original image, (b) HiFuse [9], (c) ConvNext [10], (d) CaIMA, and (e) the learned attention map of CaIMA. (f) The true causal effect $Y_{\text{effect}}$ of attention maps on diagnosis results is obtained by calculating the difference between the true diagnostic result and the counterfactual diagnostic result obtained with the intervened attention map (the intervention mode will be introduced in Section 4.4.2). This difference serves as a supervisory signal for encouraging the model to focus on the regions that are causally related to the diagnostic results. We observe that compared to other baseline methods, such as [9,10], our method not only focuses better on the object (white box) rather than other spuriously related areas (red arrows), but also prefers to look at the whole object rather than certain parts even if the lesion areas (i.e., polyp) are distributed in different locations.

diagnostic results difficult to explain. Therefore, intrinsic relationships between attention maps of lesions and diagnostic results should be explored to identify true causality from the input medical images and prediction results. To address the above issues, We revisit medical image diagnosis from a causal inference perspective. Specifically, we formalize the causality between the attention map, the input medical image, and the diagnostic result as the causal graph, and the causal intervention is employed to explore intrinsic relationships between learned attention maps and diagnosis results, as shown in Fig. 1(f), to improve the explainability of the prediction model. Specifically, the first term on the left-hand side of Fig. 1(f) represents the prediction of outcome under observed conditions (i.e., $Y(A = A; X = X)$), while the second term represents the prediction of outcome after causal intervention (i.e., $Y(do(A = A^*); X = X)$).

Moreover, in medical image diagnosis tasks, we face the challenge of diverse distribution of disease lesions, which are often not confined to a single region but spread across multiple regions of the image. As shown in Fig. 1(a), polyps may be distributed in multiple locations. The attention map should focus on multiple regions simultaneously as shown in Fig. 1(e), to provide more reliable visual explanations for medical diagnostic results. Motivated by the above observations, we argue that decomposing attention into multiple regions (i.e., multi-region attention) may be more effective in capturing multiple discriminative local features in medical images. However, multi-attention-based models [11,12] are often trained unsupervised or weakly supervised due to the lack of region-level labels (e.g., segmentation masks or bounding boxes). This trivial learning strategy tends to degrade the learned multi-attention to single attention, i.e., only one attention region produces strong responses, while the rest of the regions are suppressed.

To address the above problem, we propose an explainable medical diagnosis method by incorporating causal interventions with the multi-attention framework. Specifically, in this paper, we formalize the medical image diagnosis problem as a causal graph, as shown in Fig. 2(b), and propose a multi-region attention causal intervention module that encourages the network to learn the attention maps that are causally related to the diagnosis result by maximizing the causal effect (i.e., the difference between the effects of learned attention and the intervened attention maps on the final prediction). Our contributions are summarized as follows:

(1) We explore the explainability of medical image classification models from the perspective of causal inference. Causal intervention is introduced to compute the causal effect of attention maps on diagnostic results, this causal effect then serves as the supervision to eliminate spurious correlations arising from data selection bias and provides a reliable basis for diagnostic results.

(2) We propose a novel multi-region attention framework, called CaIMA, that incorporates non-overlapping multiple attentional guidance loss and attention-driven data augmentation modules to capture discriminative features from multiple regions of medical images.

(3) Experiments demonstrate that our CaIMA outperforms state-of-the-art methods on multimedia medical datasets, including three public datasets (i,e., Kvasir, ISIC2018, and COVID-19) and one private dataset (CLC).

The rest of this paper is organized as follows. Section 2 reviews related work, including causal inference in vision, medical image classification and causal inference in medical image analysis, and then a detailed description of the proposed method is given in Section 3. In Section 4, extensive experiments are performed to demonstrate the effectiveness of our method. Finally, conclusions are drawn in Section 5.

## 2. Related work

In this section, we present the related work of our study from three perspectives, including causal inference in vision, medical image classification, and causal inference in medical image analysis.

### 2.1. Causal inference in vision

Causal inference [13,14] has been gradually applied to better understand and explain computer vision tasks. In recent years, such as facial action unit recognition [15], person re-identification [16], noisy image classification [17], and fake news detection [18]. Causal inference gives models the ability to take into account naturally occurring causal effects in a task and to distinguish between direct and indirect effects. Rao et al. [19] designed counterfactual attention learning methods to enhance attention learning. Chen et al. [16] proposed a gesture-guided attention network for the re-identification of occluded persons, which explores causal relationships between predicted identities and input cues to mitigate the negative effects of occlusion bias. Inspired by these works, we model medical images, attention maps, and diagnosis results (i.e., label) as a causal graph as shown in Fig. 2(b), and we introduce a

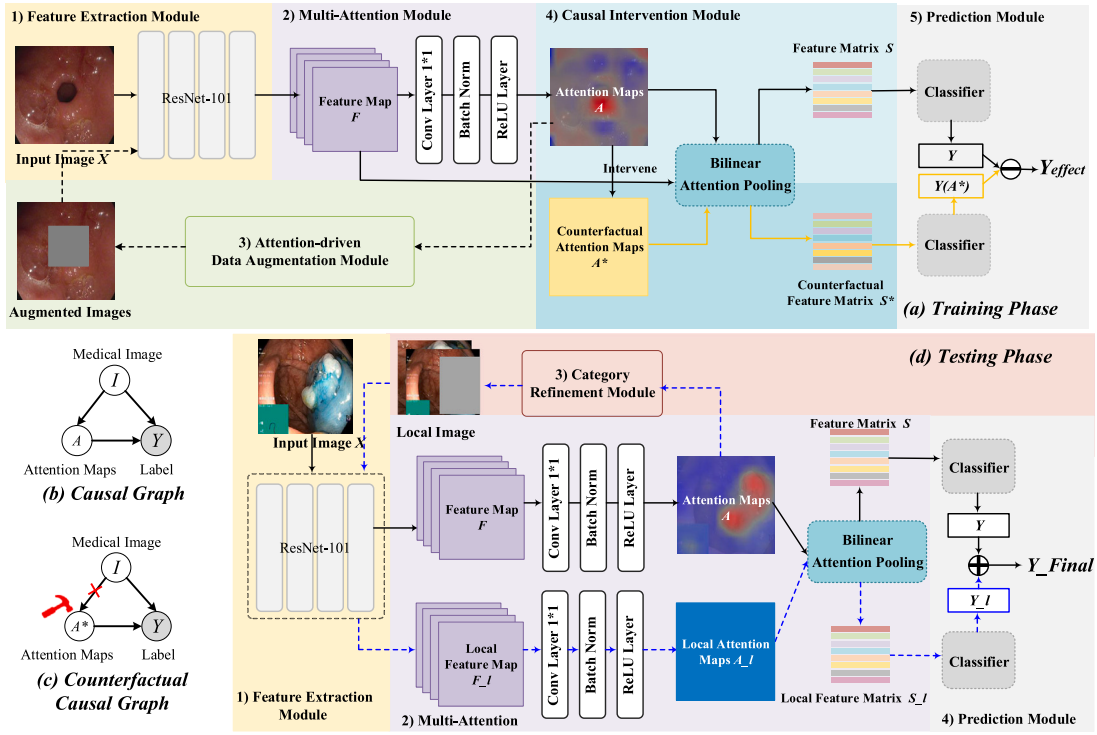



**Fig. 2.** The overview of our CaIMA and causal graph for medical image diagnosis. Solid lines represent the processing flow of the source image, where yellow lines denote the processing flow of the intervened attention map, while black lines represent the processing flow of the pre-intervention attention map. Dashed lines indicate the processing flow of images obtained through the ADDA module (black) or CRM module (blue). Note that for simplicity, we only show the workflow (Solid lines) for the original input image; in fact, the augmented image shares the workflow with the original input image, with the difference that the attention map of the augmented image does not go through the causal intervention module.

causal intervention operation into the multi-region attention network to quantify the causal effect between attention maps and diagnosis results to improve the quality of the attention maps and hence provide reliable explanations of the model's decision-making results.

### 2.2. Medical image classification

Medical image classification aims at extracting salient features from medical images for high-precision diagnosis tasks [20,21]. Previous studies have often relied on hand-designed features covering underlying image properties (e.g., edges or corner points) as well as high-level image properties (e.g., complex boundaries of cancers) for specific classification tasks [22–25]. Recently, with the powerful feature extraction capability of deep learning, medical image classification methods based on deep learning have made significant progress. These methods [7,26,27] often utilize neural networks to automatically learn image features, and thus achieve significant improvements in classification accuracy and computational efficiency. However, despite their excellent performance, these methods are often regarded as "black-box" models because their decision-making process is difficult to explain.

At this point, several methods [28–30], called explainable AI (XAI), have been proposed to help understand the decision-making process of black-box models. These methods use gradient [31], guided backpropagation [32], gradient-based class activation maps (Grad-CAMs) [33], class activation maps (CAMs), smooth grad [34] and deconvolution techniques [35] to explain the image classification tasks. Cohen et al. [36] used Grad-CAM, CAM, and heat maps to localize the lungs of patients with COVID-19-infected areas to predict the severity and extent of the disease and to produce explainable outputs. Cheng et al. [37] used a hybrid attention mechanism of channel attention and spatial attention modules to visualize skin lesion regions, providing explainability for skin lesion type prediction. Manzari et al. [38] used Crad-CAM as an attentional visualization tool to provide a visual explanation of the model's decision-making results, with significant results in the classification of various types of diseases. However, previous methods have failed to equip models with the ability to distinguish between causal and spurious correlations. The generated attention maps tend to focus on regions that are spuriously correlated with predicted labels, leading to opacity and unreliability of the model's decision-making process. Besides, these methods tend to focus only on the most salient regions while ignoring other regions that may affect the prediction results, in fact, the attention maps causally related to the disease type should focus on multiple regions simultaneously, as shown in Fig. 1. Therefore, in this paper, we combine causal intervention operations with multi-region attention networks to learn multiple causal attention maps in an image that affect the decision-making process, thus improving the visual explainability of the decision-making of the classification model.

### 2.3. Causal inference in medical image analysis

Recently, there has been significant interest in improving the performance of medical image analysis models from the perspective of causal inference. Causal inference-based methods have been applied to discover causal links of neural processes [39,40], provide explanations for network performance [41–43], generate counterfactual medical images [44,45], and improve fairness [46]. For example, Kayser et al. [41] analyzed how artifacts affect automated polyp detection from a causal perspective and proposed a multi-task learning technique for polyp detection. Pawlowski et al. [44] proposed a unified framework for structural causal models based on modular deep mechanisms, DeepSCM, and applied it to counterfactual medical image generation. Further, Ribeiro et al. [45] delved into the practical constraints of DeepSCM in real-world applications and proposed a general causal generative modeling

framework based on deep structural causal models. This framework aims to accurately estimate high-fidelity image counterfactuals.

Additionally, there are other methods that utilize causal inference for medical image segmentation tasks and have achieved outstanding segmentation performance [47–49]. Chen et al. [47] pioneered the integration of causality into weakly supervised semantic segmentation of medical images by introducing the causal CAM (C-CAM), which incorporates two causal chains to generate precise pseudo segmentation masks. In contrast to C–CAM, CauSSL [49] models the medical image segmentation task as a novel causal graph and introduces a causality-inspired semi-supervised learning approach on top of it to enhance semi-supervised learning for medical image segmentation. Despite the widespread integration of causality into different areas of medical analysis, there exists a noticeable gap in applying causality-based inference methods to medical image diagnostic tasks. This observation has also inspired the focus of this paper.

## 3. Method

This section initially states the motivation of the proposed methodology and gives a brief overview of our framework. Subsequently, it offers detailed descriptions of each module and loss function employed in our proposed method.

### 3.1. Motivation & model overview

As aforementioned, most existing classification methods focus the learned visual attention on regions spuriously related to category labels (e.g., endodermic gut) while neglecting regions causally related to medical image labels (e.g., lesion types), i.e., the causal relationship between visual attention and lesion types. This inspired us to introduce a causal intervention operation that encourages visual attention to be focused on the lesion region with which it is causally related. Furthermore, the intra-class discrepancy of medical images is typically subtle and occurs in multiple regions (as shown in Fig. 4(j)), which makes it challenging for a single attention structure network to effectively capture. Therefore, we argue that decomposing the attention into multiple regions can be more efficient for collecting local features for medical image classification. Therefore, We employ a multi-region attention network to tackle the above challenges.

However, multi-attention networks can only be trained using unsupervised or weakly supervised learning methods due to the lack of region-level labels in medical image datasets. This may cause multi-attention networks to usually degenerate into single-attention networks, where multiple attention maps focus on the same regions and ignore other regions which may provide discriminative information. To solve this issue, on the one hand, we design an attention-driven data augmentation module to encourage the network to focus on other attention regions that are relevant to the diagnostic outcome. On the other hand, we develop a non-overlapping multiple-attentional guidance loss that aims to ensure that each attention map is localized on a specific lesion area.

Fig. 2 shows the proposed CaIMA, which consists of five key components: (1) The feature extraction module that is used to extract high-level semantic representations in medical images. (2) The multi-region attention module is utilized to generate multiple attention maps from the input representations, each attention map reflecting a region that causes the diagnostic result. (3) The attention-driven data augmentation module (ADDA) is employed to facilitate the attention map to pinpoint the lesion area that causes the diagnostic result. (4) The causal intervention module is designed to improve the quality of the attention map by quantifying the causal effect between the visual attention maps and the lesion type and provide explainability for medical diagnosis. (5) The prediction module is used to give the final diagnosis of the disease. Specifically, CaIMA first uses a pre-trained model to extract medical image representations, the extracted representations are then fed into a multi-region attention module to generate multiple discriminative attention maps. The attention-driven data augmentation module is then employed to facilitate the attention maps to pinpoint the lesion area that causes the diagnostic result. Then, the causal intervention module is designed to work synergistically with the prediction module to quantify the causal effects of the learned visual attention on the final diagnostic results for removing spurious correlations between the learned attention maps and the diagnostic results and improving the model's explainability and predictive performance.

Furthermore, to prevent multiple attentions from degenerating into single attention, where multiple attention maps simultaneously focus on the same region while ignoring other regions that may provide discriminative information, we propose a non-overlapping multiple-attentional guidance loss. The design goal of this loss is to ensure that each attention map accurately localizes to a specific lesion region while avoiding multiple attention maps from repeatedly localizing in the same region.

### 3.2. Multiple attention maps generation

Consider a medical image $X$ as input, our framework first uses feature extraction network $G$ to extract high-level image features $F = G(X) \in \mathbb{R}^{C\times H\times W}$, where $W$, $H$, $C$ denote the width, height, and number of channels of the feature maps, respectively. Then $F$ are fed into the multi-region attention module $\varphi$ to generate $M$ attention maps $A = \varphi(F) \in \mathbb{R}^{W\times H\times M}$ that represent multiple discriminative regions causally related to the lesion type. The multi-region attention module is constructed by a 1*1 convolutional layer, a batch norm (BN), and a ReLU layer. The obtained attention map $A$ is then multiplied with the feature map $F$ and fed into the bilinear attention pooling (BAP) module to obtain the corresponding feature matrix $S = \{S_1, \ldots, S_m, \ldots, S_M\}$, as

$$S_m = \phi(F \odot A_m) = \frac{1}{HW}\sum_{h=1}^{H}\sum_{w=1}^{W} F^{h,w}A_m^{h,w}, \tag{1}$$

where $\odot$ denotes the element level multiplication operation and $\phi$ denotes the BAP operation. The ultimate representation $S$ can be fed to the classifier for medical image diagnosis.

### 3.3. Attention-driven data augmentation

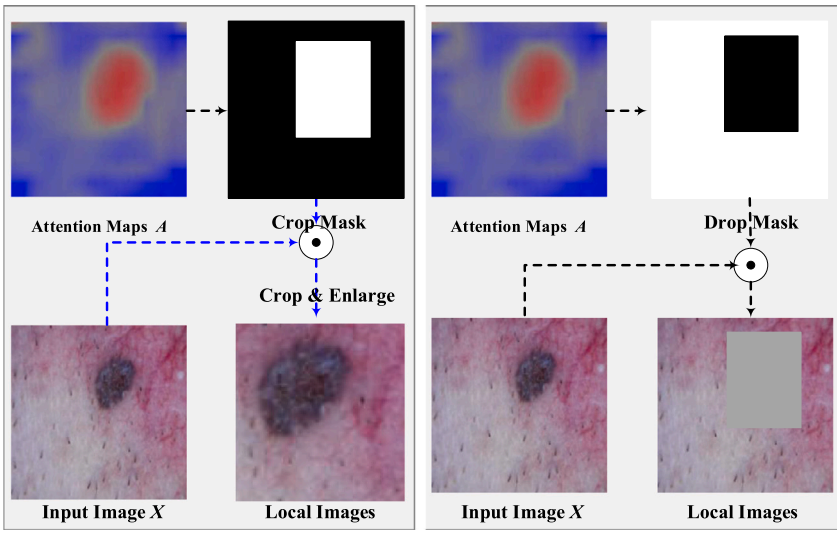
We employ the ADDA module instead of the random image augmentation methods (e.g., rotation, cropping, flipping, etc.) to prevent the degradation of the multi-region attention to single attention due to the lack of fine-grained labels. This module first select an attention map $A_k$ randomly and normalize it to obtain the augmented attention map $A_k^{\#}$, then the elements $A_k^{\#}(i,j)$ are compared with the threshold $\varsigma$ to generate the drop mask $D_k$, and finally $D_k$ is multiplied with the original image $X$ to obtain the final augmented image $X_a$, as

$$X_a = X \odot D_k, \text{where } D_k(i,j) = \begin{cases} 0, if\, \mathrm{A}_k^{\#}(i,j) > \varsigma \\ 1, otherwise \end{cases}. \tag{2}$$

In addition, we add the category refinement module (CRM) during the testing phase of CaIMA to refine the diagnostic results by augmenting the original image. Fig. 3 shows two ways of CRM. It consists of two ways, attention cropping and attention dropping. The former method is to use the ADDA module to obtain the augmented image, and the latter is to first crop the image that is obtained by element-wise multiplication of the original image and crop mask $C_k$, as shown in Eq. (3) and then enlarge the cropped image to the size of the original image by an upsampling operation.

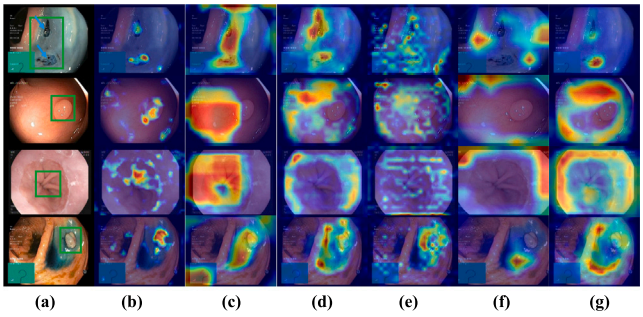
$$X_{lc} = X \odot C_k, \text{where} C_k(i,j) = \begin{cases} 1, if\, \mathrm{A}_k^{\#}(i,j) > \vartheta \\ 0, otherwise \end{cases}, \tag{3}$$

where $C_k$ is the crop mask, which is generated by comparing the attention map $A_k^{\#}$ with the threshold $\vartheta$. Then the final diagnosis result is obtained by the weighted average of the diagnosis result of the local

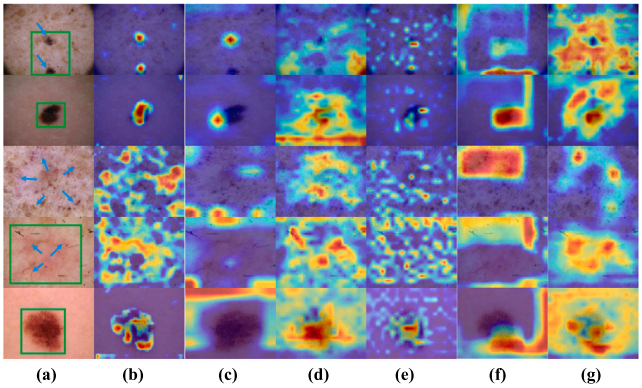



(a) Attention Cropping (b) Attention Dropping

**Fig. 3.** Two ways of CRM. (a) Attention cropping. (b) Attention dropping. .




(i) kvasir




(j) ISIC2018

**Fig. 4.** Visual explanation of different methods for predicting disease diagnosis using Grad-CAM on different datasets. (a) Original image. (b) CaIMA (c) HiFuse [9]. (d) FastViT [50]. (e) RepViT [51]. (f) DaViT [52]. (g) Skresnet50 [53]. Note that the regions that actually cause the diagnostic result are marked with green boxes in column (a), and the blue arrows indicate that the lesion areas are distributed in different regions.

images (i.e., augmented images) and the diagnosis result of the original image.

Furthermore, we propose a non-overlapping multiple attentional guidance loss to encourage the attention map to focus on multiple discriminative lesion regions, which can be calculated by

$$\mathcal{L}_N = \underbrace{\sum_{b=1}^{B}\sum_{m=1}^{M}\max\left(\|S_m - c_m\|_2 - m_{\text{intra}}, 0\right)}_{\text{Intra-class loss}} + \underbrace{\sum_{i,j\in(M,M),i\neq j}\max\left(m_{\text{inter}} - \|c_i - c_j\|_2, 0\right)}_{\text{Inter-class loss}}, \tag{4}$$

where $B$, and $M$ denote the batch size and the number of attentions, respectively. $m_o$ denotes the margin value between the feature matrix $S_m$ and the corresponding feature center $c_m$, and $m_i$ denotes the margin value between different feature centers. Specifically, the first term, as intra-class loss, promotes compact clustering of samples within the same category in the feature space to heighten the model's sensitivity to intra-class variations. The second term, inter-class loss, encourages greater dispersion of samples from distinct categories in the feature space to mitigate inter-class confusion. The feature centers $c^t$ after each training round can be updated by

$$c^t \leftarrow c^t - \kappa(c^t - \frac{1}{B}\sum_{b=1}^{B}F^b), \tag{5}$$

where $\kappa$ is the update rate of feature centers. $B$ and $F$ represent the batch size and feature maps respectively.

### 3.4. Causal intervention module

As shown in Fig. 2(b), we formalize the causality between the attention map $A$, the input medical image $X$, and the predicted label $Y$ as the causal graph, where Link $X \rightarrow (A, Y)$ indicates that $X$ is the common cause of $A$ and $Y$, and Link $(X, A) \rightarrow Y$ indicates that $X$, $A$ is the cause of $Y$. The causal intervention module collaborates with the prediction module to quantify the causal effect of the attention maps on diagnostic outcomes by mimicking the intervention operation. Fig. 2(c) shows the counterfactual causal graph after intervention on $A$. We perform an intervention operation on the "attention maps $A$", the value of $A$ is fixed and all edges pointing to the $A$ are removed.

**Table 1**
The statistics of datasets.

| Dataset | Availability | Data volume | Categories | Resolution | Image type |
|---|---|---|---|---|---|
| Kvasir | Public | 4,000 | 8 | Various Resolutions | Gastrointestinal Endoscopy images |
| ISIC2018 | Public | 10,015 | 7 | 600 × 450 | Dermoscopic images |
| COVID-19 | Public | 746 | 2 | Various Resolutions | CT images |
| CLC | Private | 19,403 | 3 | 512 × 512 | CT images |

Specifically, we first perform intervention operation (i.e., $do()$) on the attention maps $\mathbf{A}$ to generate the intervened attention maps $\mathbf{A}^*$ and the corresponding prediction label $Y(\mathbf{A}^*)$ as

$$Y(\mathbf{A}^*) = Y(do(A = \mathbf{A}^*), X = \mathbf{X}) = C(\phi(\mathbf{A}^*_{\mathbf{m}} \odot G(X))), m = 1, \ldots, M. \quad (6)$$

Here, $C$ represents the classifier. The intervened attention map $\mathbf{A}^*$ is generated randomly by sampling the attention values at each location from a uniform distribution $U(0, 2)$. Note that the optimal intervention mode was determined experimentally, the details of the experiment will be presented in Section 4.4.

Following [54,55], we then estimate the causal effect $Y_{\text{effect}}$ of the attention map $\mathbf{A}$ on the predicted outcome to motivate the model to generate high-quality attention maps by

$$Y_{\text{effect}} = \mathbb{E}_{A^* \sim \tau}[Y(\mathbf{A}) - Y(\mathbf{A}^*)], \quad (7)$$

where $\tau$ denotes the distribution of counterfactual attention. $Y(A)$ represents the predicted outcome obtained from the actual attention map $A$, $Y(A^*)$ represents the predicted outcome obtained from the counterfactual attention map $A^*$. The acquisition of $Y(A)$ is similar to Eq. (6), with the exception of replacing $do(A = A^*)$ with $A = A$, which can be expressed as $Y(\mathbf{A}) = Y(A = \mathbf{A}, X = \mathbf{X}) = C(\phi(\mathbf{A}_{\mathbf{m}} \odot G(X))),\ m = 1, \ldots, M$.

Further, we use the causal effect $Y_{\text{effect}}$ as the supervision signal to guide the attentional learning process to learn the attention map focused on more causally localized regions and to enhance model explainability. This loss function can be defined as

$$\mathcal{L}_C = CE(Y, Y_{\text{effect}}), \quad (8)$$

where $CE()$ refers to the cross-entropy loss.

### 3.5. Loss function

We design a composite loss function to guide the model training without any bounding box/part annotations, which consists of three components: standard classification loss, causal-effect loss, and non-overlapping multiple attentional guidance loss. i.e.,

$$\mathcal{L} = \omega_1 L_S + \omega_2 \mathcal{L}_C + \omega_3 \mathcal{L}_N, \quad (9)$$

where $\omega_1$, $\omega_2$, and $\omega_3$ are the weights used to balance the importance of each loss function. In our work, $\omega_1 = \omega_3 = 1$ and $\omega_1 = 0.5$. Of these, the classification loss $L_S$ encourages the network to extract informative features that focus on global discriminative regions and can be defined as

$$\mathcal{L}_S = \mathcal{L}_o + \mathcal{L}_a = CE(Y, Y') + CE(Y, Y'_a), \quad (10)$$

where $\mathcal{L}_o$ ($\mathcal{L}_a$) is employed to constrain the difference between the prediction results for the original image (augmented image) and the true category. $Y$ and $Y'$ ($Y'_a$) denote the original label and the predicted label for the original image (augmented image), respectively. Causal-effect loss $L_C$ (Eq. (8)), which is used as a supervised signal to guide the attentional learning process to learn the attention map focused on the most discriminative local regions and to enhance model explainability. Non-overlapping multiple attentional guidance loss $L_N$ (Eq. (4)) is used to encourage the attention map to focus on multiple discriminative lesion regions.

## 4. Experiments

In this section, we conduct experiments on three public datasets and one private dataset to answer the following research questions:

- **RQ1:** How does CaIMA perform compared to state-of-the-art methods in medical diagnosis?
- **RQ2:** How do different components affect model performance?
- **RQ3:** Does the learned attention map provide explainability for diagnostic results?
- **RQ4:** Does the effectiveness of CaIMA attribute to using more attention heads? Does the efficiency of the model depend on the different causal interventions?

### 4.1. Experimental setup

In this section, we provide detailed descriptions of the datasets used in the experiments, the baseline methods for comparison, as well as the implementation details of our model and the performance evaluation metrics.

#### 4.1.1. Datasets

We perform experiments on four medical datasets, including three public datasets, Kvasir [56], ISIC2018 [57], and COVID-19 [58] as well as one private dataset, CLC. The summary statistics of these datasets are shown in Table 1. For the publicly available datasets used in our experiments, each dataset presents unique challenges. Specifically, Kvasir is a multi-class medical dataset where lesions typically concentrate in fewer than two regions; ISIC2018 is a multi-class dataset with a large number of samples, and the lesion areas were usually distributed over multiple regions.; while COVID-19 is a binary classification dataset with fewer samples but more distributed lesions.

**Kvasir** [56]: This dataset is an internal image of the gastrointestinal tract collected by endoscopic equipment at the Vestre Viken Health Trust in Norway. These images show anatomical landmarks of the GI tract, pathologic findings, and endoscopic procedures. The anatomical landmarks include the Z-line, pylorus, and cecum, and the pathologic findings include esophagitis, polyps, and ulcerative colitis. The endoscopic procedure provides images related to the resection of the lesion, i.e., dyed and lifted polyp and dyed resection margins. The resolution of these images ranged from 720 × 576 to 1920 × 1072 pixels.

**ISIC2018** [57]: This dataset is a large-scale dermoscopic image dataset published by the International Skin Imaging Collaboration (ISIC), which is designed to address the problem of classifying malignant melanoma. These images contain seven skin lesion categories: melanoma, melanocytic nevus, basal cell carcinoma, actinic keratosis/Bowen's disease, benign keratosis, dermatofibroma, and vascular lesions, all with a resolution of 600 × 450.

**COVID-19** [58]: This dataset includes 349 COVID-19-positive CT scan images and 397 normal or negative CT scans containing other types of disease. The resolution of these images ranged from 236 × 184 to 512 × 450.

**CLC**: This dataset is a collection of lung cancer images from one hospital between May 2018 and September 2019, containing a total of 466 computed tomography scans of lung cancer patients. It includes three types of small cell lung cancer, namely adenocarcinoma (ADC),

**Table 2**
Performance comparison on Kvasir and ISIC2018 datasets. RI indicates relative improvement of CaIMA over the best baseline result. The best results are highlighted in bold and suboptimal results are underlined.

| Network structure | Method \ Dataset | Kvasir | | | | ISIC2018 | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | ACC | F1 | P | R | ACC | F1 | P | R |
| CNNs-based | VGG-19 [59] | 0.7775 | 0.7775 | 0.7786 | 0.7783 | 0.7925 | 0.6183 | 0.6371 | 0.6089 |
| | Skresnet50 [53] | 0.7475 | 0.7391 | 0.7475 | 0.7544 | 0.6286 | 0.3753 | 0.3683 | 0.4349 |
| | ConvNeXt-B [10] | 0.7460 | 0.7461 | 0.7478 | 0.7464 | 0.7652 | 0.5094 | 0.6684 | 0.5052 |
| | RepViT [51] | 0.8267 | 0.8259 | 0.8267 | 0.8275 | 0.6166 | 0.4474 | 0.4483 | 0.5238 |
| Transformer-based | XCiT [60] | 0.7458 | 0.7421 | 0.7458 | 0.7449 | 0.6970 | 0.4534 | 0.4505 | 0.4903 |
| | Twins_svt_Large [8] | 0.7058 | 0.7010 | 0.7058 | 0.7031 | 0.6605 | 0.3821 | 0.3608 | 0.4521 |
| | BEiT_base [61] | 0.6950 | 0.6780 | 0.6950 | 0.7081 | 0.6530 | 0.3912 | 0.3791 | 0.4520 |
| | DaViT [52] | 0.7667 | 0.7651 | 0.7667 | 0.7672 | 0.6570 | 0.2618 | 0.2473 | 0.4762 |
| MLP-based | Mixer-L [62] | 0.7430 | 0.7414 | 0.7443 | 0.7434 | 0.7892 | 0.5988 | 0.6136 | 0.5916 |
| Hybrid | Conformer [1] | 0.8425 | 0.8427 | 0.8445 | 0.8437 | 0.8266 | 0.7244 | 0.7331 | 0.7166 |
| | HiFuse [9] | 0.8083 | 0.8080 | 0.8083 | 0.8245 | **0.8412** | 0.7532 | **0.7652** | 0.7474 |
| | ConvMixer [63] | 0.7225 | 0.7116 | 0.7225 | 0.7250 | 0.6845 | 0.4006 | 0.3907 | 0.4449 |
| | FastViT [50] | 0.8592 | 0.8587 | 0.8592 | 0.8592 | 0.5981 | 0.5166 | 0.5184 | 0.6033 |
| The proposed | CaIMA(Inception v3) | 0.9183 | 0.9182 | 0.9183 | 0.9185 | 0.7544 | 0.7397 | 0.7191 | 0.7910 |
| | CaIMA(ResNet101) | **0.9358** | **0.9358** | **0.9358** | **0.9375** | 0.7813 | **0.7536** | 0.7517 | **0.8012** |
| | RI | **8.92%** | **8.98%** | **8.92%** | **9.11%** | **-7.12%** | **0.05%** | **-1.76%** | **7.20%** |

squamous carcinoma (SCC), and non-small cell carcinoma. These CT images were serial chest images in Digital Imaging and Communications in Medicine (DICOM) format, and all images with a resolution of $512 \times 512$.

#### 4.1.2. Baseline model

To demonstrate the effectiveness of CaIMA, we compare our approach with four kinds of state-of-the-art image classification methods.

**• Convolutional Neural Networks (CNNs)-based Methods**

- VGG-19 [59]. It is a major milestone in deep learning and represents a benchmark for deep CNNs in image classification.
- Skresnet50 [53]. It is a variant of ResNet [64] and is widely used for tasks such as image classification and target detection. We follow the official codebase[1] for implementation.
- ConvNeXt-B [10]. It employs a multi-scale information fusion strategy that aims to improve the feature extraction of image data. We follow the official codebase[2] for implementation.
- RepViT [51]. It revisits the design of lightweight CNNs from the perspective of a Vision transformer (ViT), which is a model constructed fully of convolutions. We follow the official codebase[3] for implementation.

**• Transformer-based Methods**

- XCiT [60]. It is an image classification model based on the attention mechanism which draws on the idea of ViT. We follow the official codebase[4] for implementation.
- Twins_svt_Large [8]. It designs a spatially separable attention mechanism to improve the performance of image classification. We follow the official codebase[5] for implementation.
- BEiT_base [61]. It is a two-phase model with a standard transformer for its backbone network. We follow the official codebase[6] for implementation.
- DaViT [52]. It introduces a dual-attention visual transformer that alternately applies spatial attention and channel attention to improve the performance of image classification. We follow the official codebase[7] for implementation.

**• Multilayer Perceptron (MLP)-based methods**

- Mixer-L [62]. Its main feature is to process images entirely using the MLP, instead of using the common convolution or self-attention mechanisms. We follow the official codebase[8] for implementation.

**• Hybrid Architectures-based Methods**

- Conformer [1]. It is a parallel two-branch network structure, where the CNN branch uses the ResNet structure and the Transformer branch uses the ViT structure. We follow the official codebase[9] for implementation.
- HiFuse [9]. It fuses the advantages of transformers and CNNs from multi-scale hierarchies to improve the classification accuracy of various medical images. We follow the official codebase[10] for implementation.
- ConvMixer [63]. It is similar in spirit to ViT and Mixer-L [62], but it uses only standard convolution for the mixing step. We follow the official codebase[11] for implementation.
- FastViT [50]. It is a generalized CNN and Transformer hybrid vision model. We follow the official codebase[12] for implementation.

#### 4.1.3. Implementation details & evaluation protocol

We adopt the standard ResNet101 [64] and Inception v3 [65] as backbone networks for our multi-region attention causal intervention framework. The training epochs for the proposed model varied across datasets: 20 epochs for ISIC2028 (ResNet) and 25 epochs for ISIC2028 (Inception v3), 50 epochs for Kvasir (ResNet) and 55 epochs for Kvasir (Inception v3), 70 epochs for NCSLC, and 40 epochs for COVID-19. The learning rate is set to 1e−3, the batch size is 32, and the update rate of feature centers $\kappa$ is 5e−2. In the $L_N$ loss function, the intra-class margin $m_{\text{intra}}$ and inter-class margin $m_{\text{inter}}$ are set to 0.05 and 0.2 respectively. In the attention-guided data augmentation module, the threshold $\varsigma$ is set to 0.5. For the number of attention heads, we explore {1, 2, 3, 4, 5, 6, 7, 8} and determine them to be 5, 4, and 4 for the Kvasir, ISIC2018, and COVID-19 datasets, respectively. All experiments are implemented based on PyTorch using NVIDIA RTX 3090 GPU.

[1] https://github.com/implus/SKNet
[2] https://github.com/facebookresearch/ConvNeXt
[3] https://github.com/THU-MIG/RepViT
[4] https://github.com/facebookresearch/xcit
[5] https://github.com/Meituan-AutoML/Twins
[6] https://github.com/microsoft/unilm/tree/master/beit
[7] https://github.com/dingmyu/davit
[8] https://github.com/lucidrains/mlp-mixer-pytorch
[9] https://github.com/pengzhiliang/Conformer
[10] https://github.com/huoxiangzuo/HiFuse
[11] https://github.com/tmp-iclr/convmixer
[12] https://github.com/apple/ml-fastvit

**Table 3**
Performance comparison on COVID-19 and CLC datasets. RI indicates relative improvement of CaIMA over the best baseline result. The best results are highlighted in bold and suboptimal results are underlined.

| Network structure | Method \ Dataset | COVID-19 | | | | CLC | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | ACC | F1 | P | R | ACC | F1 | P | R |
| CNN-based | VGG-19 [59] | 0.5914 | 0.5755 | 0.5904 | 0.5813 | 0.7070 | 0.5179 | 0.5903 | 0.4661 |
| | Skresnet50 [53] | 0.5764 | 0.5740 | 0.5796 | 0.5822 | 0.9969 | 0.9964 | 0.9958 | 0.9969 |
| | ConvNeXt-B [10] | 0.5538 | 0.5468 | 0.5495 | 0.5481 | 0.7070 | 0.5179 | 0.5903 | 0.4661 |
| | RepViT [51] | 0.6355 | 0.6229 | 0.6299 | 0.6456 | 0.9948 | 0.9934 | 0.9938 | 0.9931 |
| Transformer-based | XCiT [60] | 0.6108 | 0.6062 | 0.6153 | 0.6229 | 0.8330 | 0.8015 | 0.8034 | 0.8004 |
| | Twins_svt_Large [8] | 0.5369 | 0.4765 | 0.5259 | 0.5442 | 0.9632 | 0.9582 | 0.9570 | 0.9598 |
| | BEiT_base [61] | 0.6108 | 0.5967 | 0.6051 | 0.6187 | 0.7150 | 0.6387 | 0.6520 | 0.6872 |
| | DaViT [52] | 0.5172 | 0.3409 | 0.5000 | 0.2586 | 0.6408 | 0.4703 | 0.5383 | 0.4238 |
| MLP-based | Mixer-L [62] | 0.7043 | 0.7012 | 0.7038 | 0.7006 | 0.9075 | 0.8936 | 0.9064 | 0.8850 |
| Hybrid | Conformer [1] | 0.7581 | 0.7560 | 0.7681 | 0.7781 | 0.8668 | 0.8402 | 0.8373 | 0.8505 |
| | HiFuse [9] | 0.7340 | 0.7238 | 0.7282 | 0.7597 | 0.9940 | 0.9923 | 0.9928 | 0.9917 |
| | ConvMixer [63] | 0.6502 | 0.6490 | 0.6531 | 0.6565 | 0.9985 | 0.9979 | 0.9979 | 0.9979 |
| | FastViT [50] | 0.6552 | 0.6480 | 0.6510 | 0.6610 | 0.9972 | 0.9966 | 0.9968 | 0.9964 |
| The proposed | CaIMA(Inception v3) | 0.7980 | 0.7980 | 0.7990 | 0.7990 | **0.9990** | **0.9986** | **0.9989** | 0.9983 |
| | CaIMA(ResNet101) | **0.8571** | **0.8568** | **0.8565** | **0.8575** | **0.9990** | **0.9986** | 0.9987 | **0.9984** |
| | RI | **13.06%** | **13.33%** | **11.51%** | **10.20%** | **0.05%** | **0.07%** | **0.10%** | **0.05%** |

In all experiments, we choose 80% of the images in each dataset randomly for training, while the remaining 20% are employed for testing. Each image was scaled down to 224 × 224 pixels. We adopt accuracy (ACC), precision (P), recall (R), and F1 score (F1) as the evaluation metrics for extensive experiments. Remarkably, for multiclassification datasets (i.e., Kvasir, ISIC2018, and CLC dataset), macro-F1, macro-P and macro-R were employed. For all metrics, the larger values, the better.

### 4.2. Performance comparison (RQ1 & RQ3)

Tables 2 and 3 show the comparison results on three public real-world datasets and a private dataset. Fig. 4 presents the visual explanation of different methods for predicting disease diagnosis using Grad-CAM on ISIC2018 and Kvasir. From which We have the following observations:

First, in term of model prediction performance, our method CaIMA significantly outperforms other state-of-the-art methods on both public and private datasets, verifying the effectiveness of our method. Specifically, our CaIMA method improves the R by 9.11%, 7.20%, 10.20%, and 0.05% on the Kvasir, ISIC2018, COVID-19, and CLC datasets, respectively, compared to the best baseline methods. This is because our proposed method enables the model to learn more features causally related to diagnostic outcomes through causal intervention operations.

Second, in term of model explainability, our method provides more reliable visual explanations. As shown in Fig. 4, most state-of-the-art methods either expand regions relevant to diagnostic outcomes or focus on regions spuriously or even irrelevantly related to diagnostic outcomes. Our CaIMA method focuses more on regions causally related to diagnostic outcomes. This indicates that our proposed method can improve the accuracy and explainability of medical diagnosis by focusing on the regions of the image that actually cause the classification results and eliminating the effects of spurious correlations.

Third, the hybrid method is superior to the other three types of methods, which can be attributed to its amalgamation of CNNs' robust feature extraction abilities with the Transformer's attention mechanism. Thus it also produces competitive results in terms of diagnostic performance and explainability. However, since these methods are vulnerable to the interference of spurious correlations in the dataset during the training process, which leads to the model being more biased towards focusing on regions superficially correlated with the diagnostic results but with no actual causality and ignoring the regions that are truly causally related to the disease. our approach addresses this issue by integrating CNNs and attention mechanisms while introducing causal intervention operations. This integration allows the model to maintain reliable interpretability while improving performance.

Moreover, we conducted a comprehensive analysis of different backbones and found that our method performs better than other models on both ResNet101 and Inception v3 backbones. This indicates that our proposed method has good generality and applicability, being able to flexibly adapt to different datasets and backbone choices. Overall, CaIMA not only surpasses other baseline methods in diagnostic accuracy but also exhibits superior performance in terms of explainability.

Overall, CaIMA not only surpasses other baseline methods in diagnostic accuracy but also exhibits superior performance in terms of explainability.

### 4.3. Ablation study (RQ2 & RQ3)

In ablation studies, we first investigate the effectiveness of CaIMA and its variants, called CaIMA_N, CaIMA_C, CaIMA_A, CaIMA_L, CaIMA_CA, CaIMA_AL, and CaIMA_CL, respectively. The difference between these variants and the CaIMA is that key components of the model are either partially or fully deactivated during the training process. Then we visualize the attention maps of CaIMA and its variants on the Kvasir and ISIC2018 datasets, as shown in Fig. 5, to explore whether the learned attention maps can provide reliable explanations for diagnostic outcomes.

We consider seven variants of CaIMA in the ablation study: CaIMA_N denotes that we deactivate three key components simultaneously: causal intervention operation (CI), attention-driven data augmentation (ADDA) module, and non-overlapping multiple attentional guidance loss $L_N$. CaIMA_C, CaIMA_A, CaIMA_L denote models with CI operation, ADDA module, and $L_N$, respectively. CaIMA_CA, CaIMA_AL, CaIMA_CL denote that the models deactivate $L_N$, CI operations, and ADDA modules, respectively. The experiment results are shown in Table 4. These results allow us to draw the following conclusions.

**(i)** In all cases, CaIMA significantly outperforms CaIMA_N. A possible explanation is that the presence of sample selection or confounding bias, combined with the fact that CaIMA_N only uses likelihood loss to train the model, makes the trained model subject to spurious correlation attention maps.

**(ii)** Compare CaIMA_N, CaIMA_C and CaIMA, although the introduction of the CI module (i.e., CaIMA_C) helps the model focus on lesion regions causally related to diagnostic results, it does not address the inherent degradation issue in multi-attention networks, where the model captures only one instead of multiple lesion regions. This is

**Table 4**
Performance comparison of CaIMA and its eight variants. The best results are highlighted in bold.

| Index | Components | | | Kvasir | | | | ISIC2018 | | | | COVID-19 | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | CI | ADDA | $L_N$ | ACC | F1 | P | R | ACC | F1 | P | R | ACC | F1 | P | R |
| 1 | | | | 0.9208 | 0.9205 | 0.9208 | 0.9222 | 0.7728 | 0.6658 | 0.6343 | 0.7535 | 0.8079 | 0.8076 | 0.8075 | 0.8077 |
| 2 | ✓ | | | 0.9283 | 0.9283 | 0.9283 | 0.9287 | 0.7629 | 0.6917 | 0.6934 | 0.7421 | 0.7980 | 0.7971 | 0.7966 | 0.7995 |
| 3 | | ✓ | | 0.9267 | 0.9266 | 0.9267 | 0.9272 | 0.7504 | 0.6995 | 0.6985 | 0.766 | 0.8030 | 0.8016 | 0.8010 | 0.8061 |
| 4 | | | ✓ | 0.9333 | 0.9332 | 0.9333 | 0.9342 | 0.6650 | 0.6208 | 0.6086 | 0.7426 | 0.7291 | 0.7157 | 0.7224 | 0.7643 |
| 5 | ✓ | ✓ | | 0.9283 | 0.9217 | 0.9217 | 0.9222 | 0.7534 | 0.6791 | 0.6643 | 0.7591 | 0.8227 | 0.8211 | 0.8204 | 0.8277 |
| 6 | | ✓ | ✓ | 0.9283 | 0.9282 | 0.9283 | 0.9297 | **0.7823** | 0.7129 | 0.7155 | 0.7602 | 0.8079 | 0.8074 | 0.8071 | 0.8080 |
| 7 | ✓ | | ✓ | 0.9300 | 0.9300 | 0.9300 | 0.9305 | 0.7409 | 0.6836 | 0.6567 | 0.7869 | 0.8030 | 0.8021 | 0.8017 | 0.8040 |
| 8 | ✓ | ✓ | ✓ | **0.9358** | **0.9358** | **0.9358** | **0.9375** | 0.7813 | **0.7536** | **0.7517** | **0.8012** | **0.8571** | **0.8568** | **0.8565** | **0.8575** |

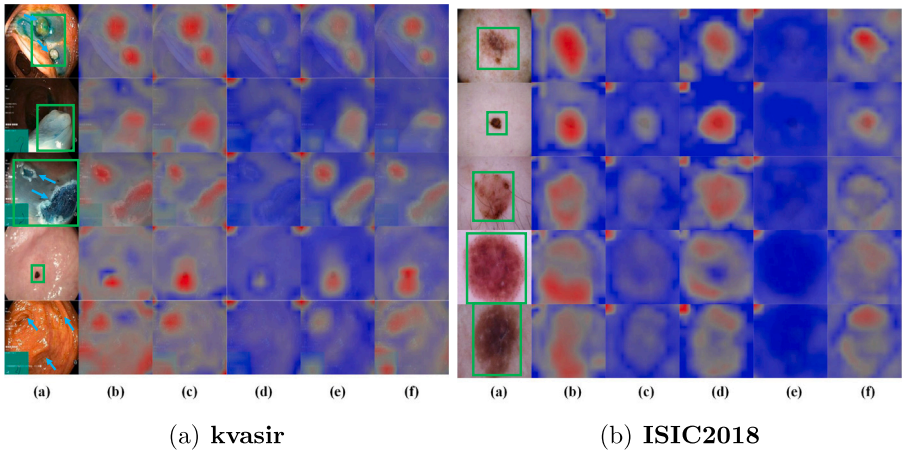



**Fig. 5.** The attention maps obtained by CaIMA and its variants. (a) Original image. (b) CaIMA. (c) CaIMA_CL. (d) CaIMA_AL. (e) CaIMA_L. (f) CaIMA_CA. Note that the regions that cause the diagnostic result are marked with green boxes in column (a), and the blue arrows indicate that the lesion areas are distributed in different regions.

particularly notable in datasets like COVID-19 and ISIC2018, where lesions typically spread across multiple regions. Hence, compared to the CaIMA_N, the performance of the CaIMA_C shows a slight decrease. When further incorporating the ADDA and $L_N$ loss (i.e., CaIMA), the model's performance significantly improves, indicating the synergistic effect of the various modules in the proposed method.

**(iii)** Compare to CaIMA_AL, the performance of CaIMA is significantly better than CaIMA_AL. Specifically, for the ISIC2018, COVID-19, and Kvasir datasets, the F1 score was improved by 4.07%, 4.94%, and 0.76%, respectively. Furthermore, it can be observed in Fig. 5 that the introduction of the CI operation drives the model to focus more on regions that are causally related to lesion types, rather than other spurious-correlated regions (as shown in Col. 4 of Fig. 5). This provides explainability for disease prediction while confirming the efficiency of the CI operation.

**(iv)** We compare CaIMA and CaIMA_CL to verify the effectiveness of the ADDA module. The results show that CaIMA is obviously superior to CaIMA_CL. The visualization results in Cols. 2 and 3 of Fig. 5 also show that the absence of the ADDA module leads to an over-expansion of individual attention regions. In contrast, the introduction of the ADDA module encourages different attention maps to focus on different regions by blurring some of the most salient regions to ensure that the model learns more robust features from other regions.

**(v)** Comparing CaIMA and CaIMA_CA, CaIMA outperforms CaIMA_CA on all three publicly available datasets. In particular, the ACC, F1-score, P, and R values of CaIMA are improved by 2.79%, 7.45%, 8.74%, and 4.21%, respectively, over CaIMA_CA in the ISIC2018 dataset. This demonstrates the effectiveness of $L_N$ in the model. Moreover, the attention maps obtained by the model without $L_N$ tend to focus on the most distinct regions in the image, as shown in the last column of Fig. 5, which does not facilitate multiple attention maps to capture different information from different regions. The introduction of $L_N$ prompts the attention map to show responses in regions with different semantic representations as shown in col. 2 of Fig. 5.

### 4.4. In-depth analysis (RQ4)

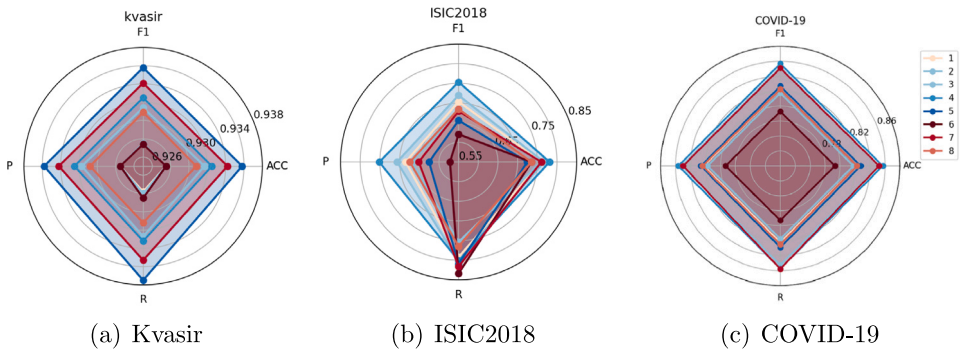
Since attention maps play a pivotal role in our method, in this subsection, we conduct sensitive analysis with different hyper-parameters and the mode of causal intervention for attention maps. Firstly, we investigate the performance of CaIMA with different $M$ values of the attention head, since $M$ value is very important and is highly related to the expressive and explainable power of the model. Secondly, we discuss how the mode of intervention on attention maps affects the performance of the model.

#### 4.4.1. Effect of attention head number

To investigate the effect of the number of attention heads on the prediction performance of the model, we conducted experiments on three datasets, Kvasir, ISIC2018, and COVID-19, respectively. Specifically, while holding the other parameters constant, we vary the number of attention heads from 1 to 8. The experimental results as shown in Fig. 6 demonstrate that CaIMA performs better under a moderate number of attention heads. When the number of attention heads is large, the model performance degrades, which may be due to the competition between the attention heads resulting in some of the attention heads not being able to effectively capture useful information. Besides, the generated attention maps may become difficult to explain and visualize. When the number of attention heads is small, the model performance is likewise degraded due to information loss caused by the model's inability to adequately focus on important parts of the input data. Therefore, either too small or too large number will harm CaIMA's performance.

#### 4.4.2. Effects of different causal interventions

To investigate the impact of different causal intervention modes on model performance, we conduct experiments on the Kvasir, ISIC2018, and COVID-19 datasets, respectively. Specifically, various intervention modes were employed to generate counterfactual attention maps,




(a) Kvasir (b) ISIC2018 (c) COVID-19

**Fig. 6.** Performance comparison with different numbers of attention heads.

**Table 5**
Effects of different causal intervention modes. The best results are highlighted in bold.

| Method \ Dataset | Kvasir | | | | ISIC2018 | | | | COVID-19 | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | ACC | F1 | P | R | ACC | F1 | P | R | ACC | F1 | P | R |
| **Aver_att** | 0.9350 | 0.9350 | 0.9350 | 0.9358 | 0.7673 | 0.7193 | 0.7020 | 0.7990 | 0.8374 | 0.8362 | 0.8354 | 0.8420 |
| **Max_att** | 0.9283 | 0.9284 | 0.9283 | 0.9290 | 0.7449 | 0.7109 | 0.6957 | 0.7931 | 0.8325 | 0.8324 | 0.8327 | 0.8323 |
| **Shuffle_att** | 0.9292 | 0.9289 | 0.9292 | 0.9305 | 0.7748 | 0.7389 | 0.7280 | 0.7956 | 0.8325 | 0.8318 | 0.8313 | 0.8338 |
| **Uniform_att** | **0.9358** | **0.9358** | **0.9358** | **0.9375** | **0.7813** | **0.7536** | **0.7517** | **0.8012** | **0.8571** | **0.8568** | **0.8565** | **0.8575** |

i.e., *Random Intervention:* Random sampling from the uniform distribution $U(0,2)$ to generate the value of each position in the counterfactual attention map. *Average Intervention:* Set the value of each position in the counterfactual attention map to the mean value of the original attention map. *Maximum Residual Intervention:* Calculate the difference between the maximum value of the original attention map and the original attention map to generate the counterfactual attention map. *Minimum Residual Intervention:* Calculate the difference between the original attention map and its minimum value to generate a counterfactual attention map. *Shuffle Intervention:* Randomly shuffle the original attention map along the batch dimension to generate a counterfactual attention map. Table 5 reports the results of the performance comparison. The experimental results show that random sampling from the uniform distribution $U(0,2)$ generates values for each position in the intervened attention map and can achieve superior performance.

## 5. Conclusion

This paper provides a fresh perspective on medical image diagnosis from a causal inference viewpoint. We propose CaIMA, a framework that combines a multi-region attention module, a causal intervention module, and an ADDA module to improve the model accuracy and explainability of diagnostic results by encouraging the attention map to focus on causally related lesion regions. For CaIMA, the feature extraction module is incorporated with a multi-region attention module to obtain multiple attention maps from high-level feature maps, and the causal intervention module is designed to quantify the causal effect of the attention map on the decision outcome to constrain the model's training process such that the learned attention map focuses on the regions causally related to the diagnostic results. Furthermore, non-overlapping multiple attentional guidance loss encourages the attention map to focus on multiple discriminative lesion regions without overlapping. Experimental results show that our CaIMA achieves significant performance improvements compared to various baselines on three public datasets and one private medical image dataset. Besides, our CaIMA can assist physicians in decision-making by providing explainable and reliable visual explanations for disease diagnosis. We believe that such explainable visual attention will become an important support tool in medical image diagnosis, providing more reliable information for clinical decision-making.

## CRediT authorship contribution statement

**Shanshan Huang:** Writing – review & editing, Writing – original draft, Methodology, Conceptualization. **Lei Wang:** Visualization, Investigation. **Jun Liao:** Funding acquisition, Conceptualization. **Li Liu:** Writing – review & editing, Supervision, Funding acquisition.

## Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Data availability

The authors do not have permission to share data.

## Acknowledgments

This work was supported by grants from the National Key R & D Program of China (grant No. 2022YFB3303302), the National Natural Science Foundation of China (grant Nos. 62377040, 62207007).

## References

[1] Z. Peng, W. Huang, S. Gu, L. Xie, Y. Wang, J. Jiao, Q. Ye, Conformer: Local features coupling global representations for visual recognition, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 367–376.
[2] S. d'Ascoli, H. Touvron, M. Leavitt, A. Morcos, G. Biroli, L. Sagun, Convit: Improving vision transformers with soft convolutional inductive biases, 2021, arXiv preprint arXiv:2103.10697.
[3] W. Lei, L. Liu, L. Liu, Spatio-temporal structure consistency for semi-supervised medical image classification, in: ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, IEEE, 2023, pp. 1–5.
[4] N. Gessert, T. Sentker, F. Madesta, R. Schmitz, H. Kniep, I. Baltruschat, R. Werner, A. Schlaefer, Skin lesion classification using cnns with patch-based attention and diagnosis-guided loss weighting, IEEE Trans. Biomed. Eng. 67 (2) (2019) 495–503.
[5] J. Wang, Y. Bao, Y. Wen, H. Lu, H. Luo, Y. Xiang, X. Li, C. Liu, D. Qian, Prior-attention residual learning for more discriminative covid-19 screening in ct images, IEEE Trans. Med. Imaging 39 (8) (2020) 2572–2583.
[6] L. Yu, W. Xiang, J. Fang, Y.-P.P. Chen, R. Zhu, A novel explainable neural network for alzheimer's disease diagnosis, Pattern Recognit. 131 (2022) 108876.

[7] L. Wang, J. Liu, P. Jiang, D. Cao, B. Pang, Ddn: Dynamic aggregation enhanced dual-stream network for medical image classification, in: ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, IEEE, 2023, pp. 1–5.

[8] X. Chu, Z. Tian, Y. Wang, B. Zhang, H. Ren, X. Wei, H. Xia, C. Shen, Twins: Revisiting the design of spatial attention in vision transformers, Adv. Neural Inf. Process. Syst. 34 (2021) 9355–9366.

[9] X. Huo, G. Sun, S. Tian, Y. Wang, L. Yu, J. Long, W. Zhang, A. Li, Hifuse: Hierarchical multi-scale feature fusion network for medical image classification, 2022, arXiv preprint arXiv:2209.10218.

[10] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, S. Xie, A convnet for the 2020s, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 11976–11986.

[11] C. He, L. Zheng, T. Tan, X. Fan, Z. Ye, Multi-attention representation network partial domain adaptation for covid-19 diagnosis, Appl. Soft Comput. 125 (2022) 109205.

[12] W. Xiong, Z. Xiong, Y. Cui, An explainable attention network for fine-grained ship classification using remote-sensing images, IEEE Trans. Geosci. Remote Sens. 60 (2022) 1–14.

[13] J. Pearl, Causal inference, causality: objectives and assessment, 2010, pp. 39–58.

[14] M.A. Hernán, J.M. Robins, Causal inference, 2010.

[15] Y. Chen, D. Chen, T. Wang, Y. Wang, Y. Liang, Causal intervention for subject-deconfounded facial action unit recognition, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 36, 2022, pp. 374–382.

[16] Y. Chen, Y. Yang, W. Liu, Y. Huang, J. Li, Pose-guided counterfactual inference for occluded person re-identification, Image Vis. Comput. 128 (2022) 104587.

[17] C.-H.H. Yang, I.-T. Hung, Y.-C. Liu, P.-Y. Chen, Treatment learning causal transformer for noisy image classification, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2023, pp. 6139–6150.

[18] L. Hu, Z. Chen, Z.Z.J. Yin, L. Nie, Causal inference for leveraging image-text matching bias in multi-modal fake news detection, IEEE Trans. Knowl. Data Eng. (2022).

[19] Y. Rao, G. Chen, J. Lu, J. Zhou, Counterfactual attention learning for fine-grained visual categorization and re-identification, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 1025–1034.

[20] X. Wu, Y. Feng, H. Xu, Z. Lin, T. Chen, S. Li, S. Qiu, Q. Liu, Y. Ma, S. Zhang, Ctranscnn: Combining transformer and cnn in multilabel medical image classification, Knowl.-Based Syst. 281 (2023) 111030.

[21] H. Zhu, W. Wang, I. Ulidowski, Q. Zhou, S. Wang, H. Chen, Y. Zhang, Meednets: Medical image classification via ensemble bio-inspired evolutionary densenets, Knowl.-Based Syst. 280 (2023) 111035.

[22] S.H. Baloch, H. Krim, Flexible skew-symmetric shape model for shape representation, classification, and sampling, IEEE Trans. Image Process. 16 (2) (2007) 317–328.

[23] Y. Song, W. Cai, Y. Zhou, D.D. Feng, Feature-based image patch approximation for lung tissue classification, IEEE Trans. Med. Imaging 32 (4) (2013) 797–808.

[24] S. Koitka, C.M. Friedrich, Traditional feature engineering and deep learning approaches at medical classification task of imageclef 2016, in: Working Notes of CLEF 2016: Conference and Labs of the Evaluation Forum, Évora, Portugal, 5-8 September 2016, 2016.

[25] P. Das, Feature-based image patch approximation for lung tissue classification using rglbp and mchog, Math. Stat. Eng. Appl. 70 (1) (2021) 261–268.

[26] F. Liu, Y. Tian, Y. Chen, Y. Liu, V. Belagiannis, G. Carneiro, Acpl: Anti-curriculum pseudo-labelling for semi-supervised medical image classification, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 20697–20706.

[27] F. Shamshad, S. Khan, S.W. Zamir, M.H. Khan, M. Hayat, F.S. Khan, H. Fu, Transformers in medical imaging: A survey, Med. Image Anal. (2023) 102802.

[28] R. Gu, G. Wang, T. Song, R. Huang, M. Aertsen, J. Deprest, S. Ourselin, T. Vercauteren, S. Zhang, Ca-net: Comprehensive attention convolutional neural networks for explainable medical image segmentation, IEEE Trans. Med. Imaging 40 (2) (2020) 699–711.

[29] B.H. Van der Velden, H.J. Kuijf, K.G. Gilhuijs, M.A. Viergever, Explainable artificial intelligence (xai) in deep learning-based medical image analysis, Med. Image Anal. 79 (2022) 102470.

[30] T. Dhar, N. Dey, S. Borra, R.S. Sherratt, Challenges of deep learning in medical image analysis—improving explainability and trust, IEEE Trans. Technol. Soc. 4 (1) (2023) 68–75.

[31] K. Simonyan, A. Vedaldi, A. Zisserman, Deep inside convolutional networks: visualising image classification models and saliency maps, in: Proceedings of the International Conference on Learning Representations, ICLR, 2014.

[32] J. Springenberg, A. Dosovitskiy, T. Brox, M. Riedmiller, Striving for simplicity: The all convolutional net, in: ICLR (Workshop Track), 2015.

[33] R.R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-cam: Visual explanations from deep networks via gradient-based localization, in: Proceedings of the IEEE International Conference on Computer Vision, 2017.

[34] D. Smilkov, N. Thorat, B. Kim, F. Viégas, M. Wattenberg, Smoothgrad: removing noise by adding noise, 2017, arXiv preprint arXiv:1706.03825.

[35] M.D. Zeiler, R. Fergus, Visualizing and understanding convolutional networks, in: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September (2014) 6-12, Proceedings, Part I 13, Springer, 2014, pp. 818–833.

[36] J.P. Cohen, L. Dao, K. Roth, P. Morrison, Y. Bengio, A.F. Abbasi, B. Shen, H.K. Mahsa, M. Ghassemi, H. Li, et al., Predicting covid-19 pneumonia severity on chest x-ray with deep learning, Cureus 12 (7) (2020).

[37] J. Cheng, S. Tian, L. Yu, C. Gao, X. Kang, X. Ma, W. Wu, S. Liu, H. Lu, Resganet: Residual group attention network for medical image classification and segmentation, Med. Image Anal. 76 (2022) 102313.

[38] O.N. Manzari, H. Ahmadabadi, H. Kashiani, S.B. Shokouhi, A. Ayatollahi, Medvit: a robust vision transformer for generalized medical image classification, Comput. Biol. Med. 157 (2023) 106791.

[39] R. Sanchez-Romero, J.D. Ramsey, K. Zhang, C. Glymour, Identification of effective connectivity subregions, 2019, arXiv preprint arXiv:1908.03264.

[40] Shanshan Huang, Qingsong Li, Lei Wang, Yuanhao Wang, Li Liu, Score-based causal feature selection for cancer risk prediction, in: 2023 IEEE International Conference on Multimedia and Expo (ICME), IEEE, 2023, pp. 198–203.

[41] M. Kayser, R.D. Soberanis-Mukul, A.-M. Zvereva, P. Klare, N. Navab, S. Albarqouni, Understanding the effects of artifacts on automated polyp detection and incorporating that knowledge via learning without forgetting, 2020, arXiv preprint arXiv:2002.02883.

[42] B. Zhang, X. Guo, Q. Lin, H. Wang, S. Xu, Counterfactual inference graph network for disease prediction, Knowl.-Based Syst. 255 (2022) 109722.

[43] M. Baniasadi, M.V. Petersen, J. Gonçalves, A. Horn, V. Vlasov, F. Hertel, A. Husch, Dbsegment: Fast and robust segmentation of deep brain structures considering domain generalization, Hum. Brain Mapp. 44 (2) (2023) 762–778.

[44] N. Pawlowski, D. Coelho de Castro, B. Glocker, Deep structural causal models for tractable counterfactual inference, Adv. Neural Inf. Process. Syst. 33 (2020) 857–869.

[45] F. De Sousa Ribeiro, T. Xia, M. Monteiro, N. Pawlowski, B. Glocker, High fidelity image counterfactuals with probabilistic causal models, in: Proceedings of the 40th International Conference on Machine Learning, 2023, pp. 7390–7425.

[46] J. Schrouff, N. Harris, O. Koyejo, I. Alabdulmohsin, E. Schnider, K. Opsahl-Ong, A. Brown, S. Roy, D. Mincu, C. Chen, et al., Maintaining fairness across distribution shift: do we have viable solutions for real-world applications? 2022, arXiv preprint arXiv:2202.01034.

[47] Z. Chen, Z. Tian, J. Zhu, C. Li, S. Du, C-cam: Causal cam for weakly supervised semantic segmentation on medical images, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 11676–11685.

[48] C. Ouyang, C. Chen, S. Li, Z. Li, C. Qin, W. Bai, D. Rueckert, Causality-inspired single source domain generalization for medical image segmentation, IEEE Trans. Med. Imaging (2022).

[49] J. Miao, C. Chen, F. Liu, H. Wei, P.-A. Heng, Caussl: Causality-inspired semi-supervised learning for medical image segmentation, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 21426–21437.

[50] P.K.A. Vasu, J. Gabriel, J. Zhu, O. Tuzel, A. Ranjan, Fastvit: A fast hybrid vision transformer using structural reparameterization, 2023, arXiv preprint arXiv:2303.14189.

[51] A. Wang, H. Chen, Z. Lin, H. Pu, G. Ding, Repvit: Revisiting mobile cnn from vit perspective, 2023, arXiv:2307.09283.

[52] M. Ding, B. Xiao, N. Codella, P. Luo, J. Wang, L. Yuan, Davit: Dual attention vision transformers, in: European Conference on Computer Vision, Springer, 2022, pp. 74–92.

[53] X. Li, W. Wang, X. Hu, J. Yang, Selective kernel networks, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 510–519.

[54] X. Wu, S. Peng, J. Li, et al., Causal inference in the medical domain: a survey, Appl. Intell. (2024).

[55] K. Kuang, L. Li, Z. Geng, L. Xu, K. Zhang, B. Liao, H. Huang, P. Ding, W. Miao, Z. Jiang, Causal inference, Engineering 6 (3) (2020) 253–263, http://dx.doi.org/10.1016/j.eng.2019.08.016.

[56] K. Pogorelov, K.R. Randel, C. Griwodz, S.L. Eskeland, T. de Lange, D. Johansen, C. Spampinato, D.-T. Dang-Nguyen, M. Lux, P.T. Schmidt, M. Riegler, P. Halvorsen, Kvasir: A multi-class image dataset for computer aided gastrointestinal disease detection, in: Proceedings of the 8th ACM on Multimedia Systems Conference, MMSys'17, ACM, New York, NY, USA, 2017, pp. 164–169, http://dx.doi.org/10.1145/3083187.3083212.

[57] N. Codella, V. Rotemberg, P. Tschandl, M.E. Celebi, S. Dusza, D. Gutman, B. Helba, A. Kalloo, K. Liopyris, M. Marchetti, et al., Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic), 2019, arXiv preprint arXiv:1902.03368.

[58] J. Zhao, Y. Zhang, X. He, P. Xie, Covid-ct-dataset: a ct scan dataset about covid-19, 2020, arXiv preprint arXiv:2003.13865.

[59] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, 2014, arXiv preprint arXiv:1409.1556.

[60] A. Ali, H. Touvron, M. Caron, P. Bojanowski, M. Douze, A. Joulin, I. Laptev, N. Neverova, G. Synnaeve, J. Verbeek, et al., Xcit: Cross-covariance image transformers, Adv. Neural Inf. Process. Syst. 34 (2021) 20014–20027.

[61] H. Bao, L. Dong, S. Piao, F. Wei, Beit: Bert pre-training of image transformers, 2021, arXiv preprint arXiv:2106.08254.

[62] I.O. Tolstikhin, N. Houlsby, A. Kolesnikov, L. Beyer, X. Zhai, T. Unterthiner, J. Yung, A. Steiner, D. Keysers, J. Uszkoreit, et al., Mlp-mixer: An all-mlp architecture for vision, Adv. Neural Inf. Process. Syst. 34 (2021) 24261–24272.
[63] A. Trockman, J.Z. Kolter, Patches are all you need? 2022, arXiv preprint arXiv:2201.09792.
[64] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.
[65] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, Z. Wojna, Rethinking the inception architecture for computer vision, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2818–2826.