Chest X-ray Image Classification: A Causal Perspective

Weizhi Nie[1], Chen Zhang[1], Dan Song[1(✉)], Yunpeng Bai[2], Keliang Xie[3], and An-An Liu[1]

[1] Tianjin University, Tianjin 300072, China
{weizhinie,zhangchen001,dan.song}@tju.edu.cn
[2] Department of Cardiac Surgery, Chest Hospital, Tianjin University, and Clinical school of Thoracic, Tianjin Medical University, Tianjin 300052, China
[3] Department of Critical Care Medicine, Department of Anesthesiology, and Tianjin Institute of Anesthesiology, Tianjin Medical University General Hospital, Tianjin 300052, China

Abstract. The chest X-ray (CXR) is a widely used and easily accessible medical test for diagnosing common chest diseases. Recently, there have been numerous advancements in deep learning-based methods capable of effectively classifying CXR. However, assessing whether these algorithms truly capture the cause-and-effect relationship between diseases and their underlying causes, or merely learn to map labels to images, remains a challenge. In this paper, we propose a causal approach to address the CXR classification problem, which involves constructing a structural causal model (SCM) and utilizing backdoor adjustment to select relevant visual information for CXR classification. Specifically, we design various probability optimization functions to eliminate the influence of confounding factors on the learning of genuine causality. Experimental results demonstrate that our proposed method surpasses the performance of two open-source datasets in terms of classification performance. To access the source code for our approach, please visit: https://github.com/zc2024/Causal_CXR.

Keywords: Medical image processing · Causal inference · Chest X-ray image classification

1 Introduction

Chest X-ray (CXR) is a non-invasive diagnostic test frequently utilized by medical practitioners to identify thoracic diseases. In clinical practice, the interpretation of CXR results is typically performed by expert radiologists, which can be time-consuming and subject to individual medical abilities [1]. Consequently, researchers have sought automated and accurate CXR classification technologies based on machine learning, aiming to assist physicians in achieving more precise diagnoses [6,7,17,18,21]. However, there are some inherent problems with

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
H. Greenspan et al. (Eds.): MICCAI 2023, LNCS 14222, pp. 25–35, 2023.
https://doi.org/10.1007/978-3-031-43898-1_3

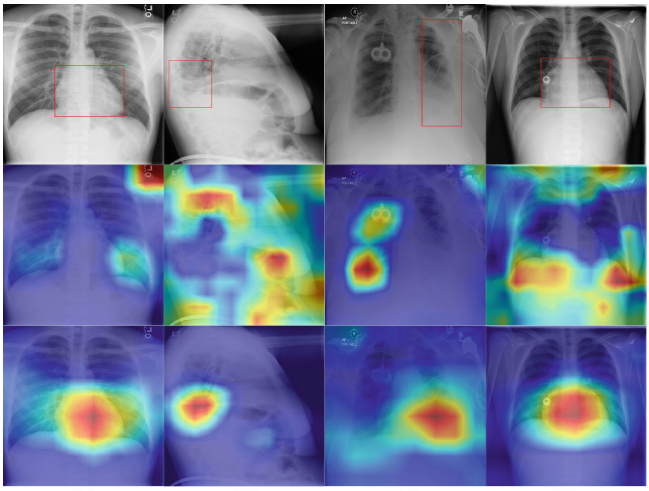

Fig. 1. Some tough cases in NIH dataset. Each column is the same CXR image, and each row from top to bottom shows the original image with a pathological bounding box, weighted heat maps of traditional CNN-based deep learning, and our method. Four difficult situations such as (a): letters on images, (b): irregular images, (c): medical devices on images, and (d): easily confused between classes.

CXR images that are difficult to solve, such as high interclass similarity [16], dirty atypical data, complex symbiotic relationships between diseases [21], and long-tailed or imbalanced data distribution [26].

Some examples are shown in Fig. 1 from the NIH dataset, we can find previous methods performed not stable when dealing with some tough cases. For example, the label of Fig. 1(d) is cardiomegaly but the predicting results generated by a traditional CNN-based model is infiltration, which fits the statistical pattern of symbiosis between these two pathologies [21]. The black-box nature of deep learning poses challenges in determining whether the learned representations truly capture causality, even when the proposed models demonstrate satisfactory performance. Unfortunately, some recent efforts such as [9,17] already notice part of the above problems but only try to solve it by data pre-processing or designing complicated model, these approaches have not succeeded in enabling deep models to effectively capture genuine causality.

To effectively address the aforementioned challenges, we approach the task of CXR image classification from a causal perspective. Our approach involves elucidating the relationships among causal features, confounding features, and the classification outcomes. In essence, our fundamental idea revolves around the concept of "borrowing from others." To illustrate this concept, let us consider an example involving letters in an image. Suppose a portion of the image contains marked letters, which can impact the classification of the unmarked portion. We perceive these letters as confounders. By borrowing the mark from the marked

portion and adding it to the unmarked part, we effectively eliminate the confounding effect: "If everyone has it, it's as if no one has it." The same principle applies to other confounding assumptions we have mentioned.

Towards this end, we utilize causal inference to minimize the confounding effect and maximize the causal effect to achieve a stable and decent performance. Specifically, we utilize CNN-based modules to extract the feature from the input CXR images, and then apply Transformer based cross-attention mechanism [20] to produce the estimations of the causal and confounding features from the feature maps. After that, we parameterize the backdoor adjustment by causal theory [14], which combines every causal estimation with different confounding estimations and encourages these combinations to remain a stable classification performance via the idea of "borrowing from others". It tends to facilitate the invariance between the causal patterns and the classification results.

We evaluate our method on multiple datasets and the experimental results consistently demonstrate the superior performance of our approach. The contributions of our work can be summarized as follows:

- We take a casual look at the chest X-ray multi-label classification problem and model the disordered or easily-confused part as the confounder.
- We propose a framework based on the guideline of backdoor adjustment and presented a novel strategy for chest X-ray image classification. It allows our properly designed model to exploit real and stable causal features while removing the effects of filtrable confounding patterns.
- Extensive experiments on two large-scale public datasets justify the effectiveness of our proposed method. More visualizations with detailed analysis demonstrate the interpretability and rationalization of our proposed method.

2 Methodology

In this section, we first define the causal model, then identify the strategies to eliminate confounding effects.

2.1 A Causal View on CXR Images

From the above discussion, we construct a Structural Causal Model (SCM) [2] in Fig. 2 to solve the spurious correlation problems in CXR. It contains the causalities about four elements: Input CXR image D, confounding feature C, causal feature X, and prediction Y, where the arrows between elements stand for cause and effect: cause $\rightarrow$ effect. We have the following explanations:

- $C \leftarrow D \rightarrow X$: X denotes the causal feature which really contributes to the diagnosis, C denotes the confounding feature which may mislead the diagnosis and is usually caused by data bias and other complex situations mentioned above. The arrows denote feature extraction process, C and X usually coexist in the medical data D, these causal effects are built naturally.

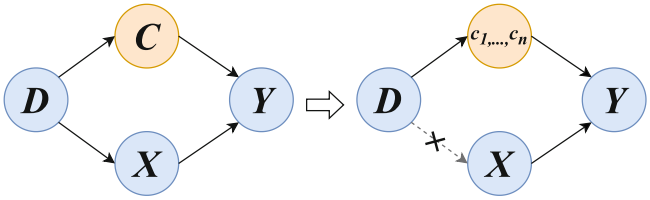

Fig. 2. SCM for CXR image classification. "D" is the input data, "C" denotes the confounding features, "X" is the causal features and "Y" is the prediction results. Confounding factors can block backdoor path between causal variables, so after adjustment, the path is blocked, shown in right part.

- $C \rightarrow Y \leftarrow X$: We denote Y as the classification result which should have been caused only by X but inevitably disturbed by confounding features. The two arrows can be implemented by classifiers.

The goal of the model should capture the true causality between X and Y, avoiding the influence of C. However, the conventional correlation $P(Y|X)$ fails to achieve that because of the backdoor path [13] $X \leftarrow D \rightarrow C \rightarrow Y$. Therefore, we apply the causal intervention to cut off the backdoor path and use $P(Y|do(X))$ to replace $P(Y|X)$, so the model is able to exploit causal features.

2.2 Causal Intervention via Backdoor Adjustment

Here, we propose to use the backdoor adjustment [2] to implement $P(Y|do(X))$ and eliminate the backdoor path, which is shown on the right of Fig. 2. The backdoor adjustment assumes that we can observe and stratify the confounders, *i.e.*, $C = \{c_1, c_2, ..., c_n\}$, where each c is a stratification of the confounder feature. We can then exploit the powerful **do-calculus** on causal feature X by estimating $P_b(Y|X) = P(Y|do(X))$, where the subscript b denotes the backdoor adjustment on the SCM. Causal theory [14] provides us with three key conclusions:

- $P(c) = P_b(c)$: the marginal probability is invariant under the intervention, because C will remain unchanged when cutting the link between D and X.
- $P_b(Y|X, c) = P(Y|X, c)$: Y's response to X and C has no connection with the causal effect between X and C.
- $P_b(c|X) = P_b(c)$: X and C are independent after backdoor adjustment.

Based on the conclusions, the backdoor adjustment for the SCM in Fig. 2 is:

$$\begin{aligned} P(Y|do(X)) = P_b(Y|X) &= \sum_{c \in \mathcal{C}} P_b(Y|X, c) P_b(c|X) \\ &= \sum_{c \in \mathcal{C}} P_b(Y|X, c) P_b(c) = \sum_{c \in \mathcal{C}} P(Y|X, c) P(c), \end{aligned} \tag{1}$$

where $\mathcal{C}$ is the confounder set, $P(c)$ is the prior probability of c. We approximate the formula by a random sample operation which will be detailed next.

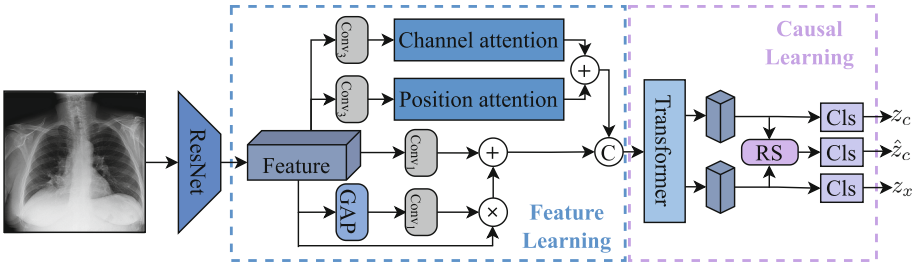

Fig. 3. Overview of our network. Firstly, we apply CNN with modified attention to extract the image feature, where the n in Conv_n denotes the kernel size of the convolutional operation, "+", "×", and "C" denote add, multiply, and concatenate operations, respectively. "GAP" means global average pooling, "RS" is the random sample operation, and "Cls" denotes the classifier. The cross-attention module inside the transformer decoder disentangles the causal and confounding feature, then we can apply parameterized backdoor adjustment to achieve causal inference.

2.3 Training Object

Till now, we need to provide the implementations of Eq. (1) in a parameterized method to fit the deep learning model. However, in the medical scenario, $\mathcal{C}$ is complicated and hard to obtain, so we simplify the problem and assume a uniform distribution of confounders. Traditionally, the effective learning of useful knowledge in deep models heavily relies on the design of an appropriate loss function. Then, towards effective backdoor adjustment, we utilize different loss functions to drive our deep model to learn causal and spurious features respectively. Figure 3 illustrates the proposed network. Note that the channel and position attention is implemented by adopting an efficient variant of self-attention [12]. We will break the whole framework down in detail below.

Given $x \in \mathbb{R}^{H_0 \times W_0 \times 3}$ as input, we extract its spatial feature $F \in \mathbb{R}^{H \times W \times v}$ using the backbone, where $H_0 \times W_0$, $H \times W$ represent the height and width of the CXR image and the feature map respectively, and v denotes the hidden dimension of the network. Then, we use zero-initialized $Q_0 \in \mathbb{R}^{k \times v}$ as the queries in the cross-attention module inside the transformer, where k is the number of categories, each decoder layer l updates the queries Q_{l-1} from its previous layer. Here, we denote Q as the causal features and $\overline{Q}$ as the confounding features:

$$Q_l = softmax(\widetilde{Q}_{l-1}\widetilde{F}/\sqrt{dim_{\widetilde{F}}})F, \overline{Q}_l = (1 - softmax(\widetilde{Q}_{l-1}\widetilde{F}/\sqrt{dim_F}))F, \quad (2)$$

where the tilde means position encodings, the disentangled features yield two branches, which can be fed separately into a point-wise Multi-layer perceptron (MLP) network and get corresponding classification logits via a sigmoid function.

Disentanglement. As shown in Fig. 3, we try to impel the model to learn both causal and confounding features via the designed model structure and loss

function. Specifically, we adopt a CNN-based model to extract the feature of input images, then capture the causal feature and confounding feature by cross-attention mechanism. Thus we can make the prediction via MLP and classifiers:

$$\begin{aligned} h_c &= MLP_{confounding}(\overline{Q_l}), z_c = \Phi_c(h_c), \\ h_x &= MLP_{causal}(Q_l), z_x = \Phi_x(h_x), \end{aligned} \tag{3}$$

where $h \in \mathbb{R}^{v \times k}$, $\Phi(\cdot)$ represents classifier, and z denotes logits.

The causal part aims to estimate the really useful feature, so we apply the supervised classification loss in a cross-entropy format:

$$\mathcal{L}_{sl} = -\frac{1}{|D|} \sum_{d \in D} y^{\top} \log(z_x), \tag{4}$$

where d is a sample and D is the training data, y is the corresponding label. The confounding part is undesirable for classification, so we follow the work in CAL [19] and push its prediction equally to all categories, then the confounding loss is defined as:

$$\mathcal{L}_{conf} = -\frac{1}{|D|} \sum_{d \in D} KL(y_{uniform}, z_c), \tag{5}$$

where KL is the KL-Divergence, and $y_{uniform}$ denotes a predefined uniform distribution.

Causal Intervention. The idea of the backdoor adjustment formula in Eq. (1) is to stratify the confounder and combine confounding and causal features manually, which is also the implementation of the random sample in Fig. 3. For this propose, we stratify the extracted confounding feature and randomly add it to the other CXR images' features, then feed into the classifier as shown in Eq. (6), and get a "intervened graph", then we have the following loss guided by causal inference:

$$\hat{z}_c = \Phi(h_x + \hat{h}_c), \tag{6}$$

$$\mathcal{L}_{bd} = -\frac{1}{|D| \cdot |\hat{D}|} \sum_{d \in D} \sum_{\hat{d} \in \hat{D}} y^{\top} \log(\hat{z}_c), \tag{7}$$

where $\hat{z}_c$ is the prediction from a classifier on the "intervened graph", $\hat{h}_c$ is the stratification feature via Eq. (3), $\hat{D}$ is the estimated stratification set contains trivial features. The training objective of our framework can be defined as:

$$\mathcal{L} = \mathcal{L}_{sl} + \alpha_1 \mathcal{L}_{conf} + \alpha_2 \mathcal{L}_{bd}, \tag{8}$$

where α_1 and α_2 are hyper-parameters, which decide how powerful disentanglement and backdoor adjustment are. It pushes the prediction stable because of the shared image features according to our detailed results in the next section.

3 Experiments

3.1 Experimental Setup

We evaluate the common thoracic diseases classification performance on the NIH ChestX-ray14 [21] and CheXpert [6] data sets. NIH consists of 112,120 frontal-view CXR images with 14 diseases and we follow the official data split for a fair comparison, and the latter dataset consists of 224,316 images.

In our experiments, we adopt ResNet101 [5] as the backbone. Our experiment is operated by using NVIDIA GeForce RTX 3090 with 24 GB memory. We use the Adam optimizer [8] with a weight decay of $1e$-2 and the max learning rate is $1e$-3. On the NIH data set, we resize the original images to 512 $\times$ 512 as the input and 320 $\times$ 320 on CheXpert. We evaluate the classification performance of our method with the area under the ROC curve (AUC) for the whole test set.

3.2 Results and Analysis

Table 1 illustrates the overall performance of the NIH Chest-Xray14 dataset of our proposed method compared with other previous state-of-art works, the best performance of each pathology is shown in bold. From the experiments on the NIH data set, we can conclude that we eliminate some spurious relationships within and among CXR images from the classification results. Specifically, we

Table 1. Comparation of AUC scores with previous SOTA works. We report the AUC with a 95% confidence interval (CI) of our method.

Abnormality	DNetLoc [4]	Xi *et al.* [11]	ImageGCN [10]	DGFN [3]	Ours
Atelectasis	0.77	0.77	0.80	**0.82**	0.81 (0.81, 0.82)
Cardiomegaly	0.88	0.87	0.89	0.93	**0.94** (0.93, 0.95)
Effusion	0.83	0.83	0.87	0.88	**0.91** (0.91, 0.92)
Infiltration	0.71	0.71	0.70	**0.75**	**0.75** (0.74, 0.77)
Mass	0.82	0.83	0.84	0.88	**0.89** (0.88, 0.90)
Nodule	0.76	**0.79**	0.77	**0.79**	0.76 (0.74, 0.79)
Pneumonia	0.73	**0.82**	0.72	0.78	**0.82** (0.80, 0.83)
Pneumothorax	0.85	0.88	0.90	0.89	**0.91** (0.91, 0.93)
Consolidation	0.75	0.74	0.80	0.81	**0.82** (0.81, 0.83)
Edema	0.84	0.84	0.88	0.89	**0.90** (0.89, 0.90)
Emphysema	0.90	**0.94**	0.92	**0.94**	**0.94** (0.93, 0.95)
Fibrosis	0.82	0.83	0.83	0.82	**0.84** (0.84, 0.85)
Pleural_Thicken	0.76	0.79	0.79	**0.81**	0.77 (0.75, 0.78)
Hernia	0.90	0.91	**0.94**	0.92	**0.94** (0.92, 0.95)
Mean AUC	0.807	0.819	0.832	0.850	**0.857** (0.849, 0.864)

can find that we are not only making progress in most categories but also dealing with some pathologies with high symbiotic dependence such as cardiomegaly and infiltration [21]. The visualization results in Fig. 1 prove that the issues raised were addressed.

Table 2. Ablation study on NIH data set.

Model	Feature Learning	Causal Learning	AUC
1	-	-	0.812
2	-	+	0.833
3	+	-	0.824
4	+	+	**0.857**

We conduct experiments on the random addition ratio of "confounding features" and found that the ratio of 30% to 40% is appropriate. Besides, the α_1 in Eq. 8 works well around 0.4 to 0.7, and α_2 works well around 0.4 to 0.5.

Ablation studies on the NIH data set are shown in Table. 2. Where "+" denotes utilizing the module whereas "-" denotes removing the module. We demonstrate the efficiency of our method from the ablation study, and we can find that our feature extraction and causal learning module play significant roles, respectively. Besides, during the training process, Fig. 4 shows the fluctuation of the classification effect of three classifiers, where the three lines in the diagram correspond to the three classifiers in Fig. 3. We can find the performance of the confounding classifier goes up at first and then down. At the same time, the other two classifiers' performance increased gradually, which is in line with our expectations. After visualization, we found that confounding factors could be "beneficial" for classification in some cases (e.g., certain diseases require patients to wear certain medical devices during X-rays), but this is the wrong shortcut, we expect the model to get causal features. Our causal learning framework successfully discards the adverse effect of confounding features and makes the prediction stable.

The results on CheXpert also prove the superiority of our method, we achieve the mean AUC of 0.912 on the five challenging pathologies [6], which surpasses the performance of previous SOTA works such as [6] and [15].

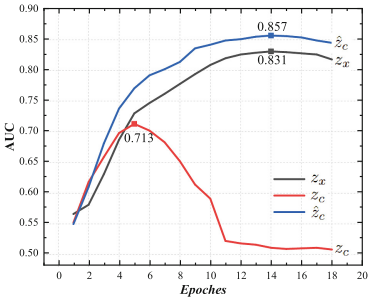

Fig. 4. Fluctuation of classification effect of three classifiers.

Our method may be general and can be applied to many other medical scenario such as glaucoma [24,25] and segmentation task [22,23]. We will apply contrast learning or self supervised learning in our future works inspired by above-mentioned papers.

4 Conclusion

In conclusion, we present a novel causal inference-based chest X-ray image multi-label classification framework from a causal perspective, which comprises a feature learning module and a backdoor adjustment-based causal inference module. We find that previous deep learning based strategies are prone to make the final prediction via some spurious correlation, which plays a confounder role then damages the performance of the model. We evaluate our proposed method on two public data sets, and experimental results indicate that our proposed framework and method are superior to previous state-of-the-art methods.

Acknowledgement. This work was supported by the National Natural Science Foundation of China (62272337).

References

1. Brady, A., Laoide, R.Ó., McCarthy, P., McDermott, R.: Discrepancy and error in radiology: concepts, causes and consequences. Ulster Med. J. **81**(1), 3 (2012)
2. Glymour, M., Pearl, J., Jewell, N.P.: Causal Inference in Statistics: A Primer. Wiley, Hoboken (2016)
3. Gong, X., Xia, X., Zhu, W., Zhang, B., Doermann, D., Zhuo, L.: Deformable Gabor feature networks for biomedical image classification. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 4004–4012 (2021)
4. Gündel, S., Grbic, S., Georgescu, B., Liu, S., Maier, A., Comaniciu, D.: Learning to recognize abnormalities in chest x-rays with location-aware dense networks. In: Vera-Rodriguez, R., Fierrez, J., Morales, A. (eds.) CIARP 2018. LNCS, vol. 11401, pp. 757–765. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-13469-3_88

5. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
6. Irvin, J., et al.: Chexpert: a large chest radiograph dataset with uncertainty labels and expert comparison. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, pp. 590–597 (2019)
7. Ke, A., Ellsworth, W., Banerjee, O., Ng, A.Y., Rajpurkar, P.: CheXtransfer: performance and parameter efficiency of ImageNet models for chest x-ray interpretation. In: Proceedings of the Conference on Health, Inference, and Learning, pp. 116–124 (2021)
8. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
9. Liu, H., Wang, L., Nan, Y., Jin, F., Wang, Q., Pu, J.: SDFN: segmentation-based deep fusion network for thoracic disease classification in chest x-ray images. Comput. Med. Imaging Graph. **75**, 66–73 (2019)
10. Mao, C., Yao, L., Luo, Y.: ImageGCN: multi-relational image graph convolutional networks for disease identification with chest x-rays. IEEE Trans. Med. Imaging **41**(8), 1990–2003 (2022)
11. Ouyang, X., et al.: Learning hierarchical attention for weakly-supervised chest x-ray abnormality localization and diagnosis. IEEE Trans. Med. Imaging **40**(10), 2698–2710 (2020)
12. Pan, X., et al.: On the integration of self-attention and convolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 815–825 (2022)
13. Pearl, J.: Interpretation and identification of causal mediation. Psychol. Methods **19**(4), 459 (2014)
14. Pearl, J., et al.: Models, reasoning and inference. Cambridge, UK: Cambridge University Press 19(2) (2000)
15. Pham, H.H., Le, T.T., Tran, D.Q., Ngo, D.T., Nguyen, H.Q.: Interpreting chest x-rays via CNNs that exploit hierarchical disease dependencies and uncertainty labels. Neurocomputing **437**, 186–194 (2021)
16. Rajaraman, S., Antani, S.: Training deep learning algorithms with weakly labeled pneumonia chest x-ray data for covid-19 detection. MedRxiv (2020)
17. Rocha, J., Pereira, S.C., Pedrosa, J., Campilho, A., Mendonça, A.M.: Attention-driven spatial transformer network for abnormality detection in chest x-ray images. In: 2022 IEEE 35th International Symposium on Computer-Based Medical Systems (CBMS), pp. 252–257. IEEE (2022)
18. Saleem, H.N., Sheikh, U.U., Khalid, S.A.: Classification of chest diseases from x-ray images on the CheXpert dataset. In: Mekhilef, S., Favorskaya, M., Pandey, R.K., Shaw, R.N. (eds.) Innovations in Electrical and Electronic Engineering. LNEE, vol. 756, pp. 837–850. Springer, Singapore (2021). https://doi.org/10.1007/978-981-16-0749-3_64
19. Sui, Y., Wang, X., Wu, J., Lin, M., He, X., Chua, T.S.: Causal attention for interpretable and generalizable graph classification. In: Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, pp. 1696–1705 (2022)
20. Vaswani, A., et al.: Attention is all you need. In: Advances in Neural Information Processing Systems, vol. 30 (2017)

21. Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., Summers, R.M.: 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, pp. 2097–2106 (2017)
22. Wu, J. et al.: SeATrans: learning segmentation-assisted diagnosis model via transformer. In: Wang, L., Dou, Q., Fletcher, P.T., Speidel, S., Li, S. (eds.) Medical Image Computing and Computer Assisted Intervention - MICCAI 2022. MICCAI 2022. LNCS, vol. 13432, pp 677–687. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-16434-7_65
23. Wu, J., et al.: Calibrate the inter-observer segmentation uncertainty via diagnosis-first principle. arXiv preprint arXiv:2208.03016 (2022)
24. Wu, J. et al.: Opinions vary? Diagnosis first!. In: Wang, L., Dou, Q., Fletcher, P.T., Speidel, S., Li, S. (eds.) Medical Image Computing and Computer Assisted Intervention - MICCAI 2022. MICCAI 2022. LNCS, vol. 13432, pp. 604–613. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-16434-7_58
25. Wu, J., et al.: Leveraging undiagnosed data for glaucoma classification with teacher-student learning. In: Martel, A.L., et al. Medical Image Computing and Computer Assisted Intervention - MICCAI 2020. MICCAI 2020. LNCS, vol. 12261, pp. 731–740. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-59710-8_71
26. Zhang, Y., Kang, B., Hooi, B., Yan, S., Feng, J.: Deep long-tailed learning: a survey. arXiv preprint arXiv:2110.04596 (2021)