Information Bottleneck-based Causal Attention for Multi-label Medical Image Recognition

Xiaoxiao Cui[1], Yiran Li[2], Kai He[2], Shanzhi Jiang[2], Mengli Xue[2], Wentao Li[1], Junhong Leng[4], Zhi Liu(✉)[2], Lizhen Cui(✉)[1], and Shuo Li[5]

[1] Joint SDU-NTU Centre for Artificial Intelligence Research (C-FAIR), Shandong University, Jinan, Shandong 250101, China
[2] School of Information Science and Engineering, Shandong University, Qingdao, Shandong 266237, China
{clz, liuzhi}@sdu.edu.cn
[3] Department of ultrasound, Jinan Maternity and Child Care Hospital Affiliated to Shandong First Medical University, Jinan, Shandong 250012, China
[4] Case Western Reserve University, Cleveland, OH 44106, USA

Abstract. Multi-label classification (MLC) of medical images aims to identify multiple diseases and holds significant clinical potential. A critical step is to learn class-specific features for accurate diagnosis and improved interpretability effectively. However, current works focus primarily on causal attention to learn class-specific features, yet they struggle to interpret the true cause due to the inadvertent attention to class-irrelevant features. To address this challenge, we propose a new structural causal model (SCM) that treats class-specific attention as a mixture of causal, spurious, and noisy factors, and a novel Information Bottleneck-based Causal Attention (IBCA) that is capable of learning the discriminative class-specific attention for MLC of medical images. Specifically, we propose learning Gaussian mixture multi-label spatial attention to filter out class-irrelevant information and capture each class-specific attention pattern. Then a contrastive enhancement-based causal intervention is proposed to gradually mitigate the spurious attention and reduce noise information by aligning multi-head attention with the Gaussian mixture multi-label spatial. Quantitative and ablation results on Endo and MuReD show that IBCA outperforms all methods. Compared to the second-best results for each metric, IBCA achieves improvements of 6.35% in CR, 7.72% in OR, and 5.02% in mAP for MuReD, 1.47% in CR, and 1.65% in CF1, and 1.42% in mAP for Endo.

Keywords: Multi-label Classification · Vision Transformer · Variational Information Bottleneck · Gaussian Mixture Model.

1 Introduction

Multi-label classification (MLC) of medical images is crucial for the clinical diagnosis of diseases, as multiple diseases or conditions can co-occur within a single image [8, 16, 18, 20]. Instead of training separate classifiers for each label, MLC

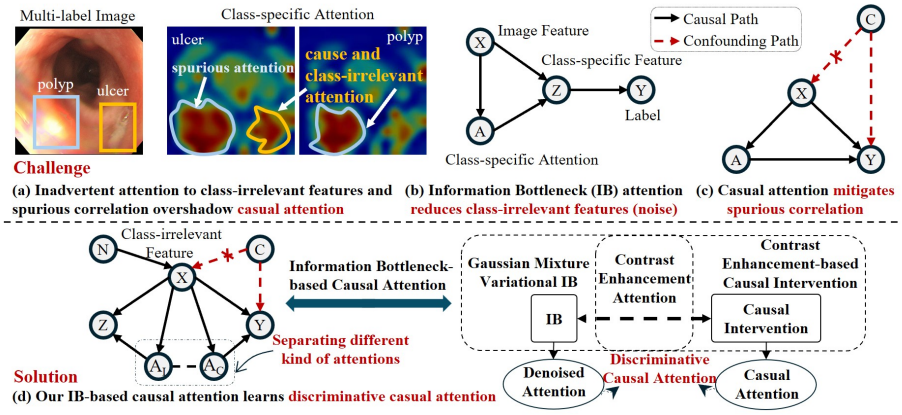

Fig. 1. Challenge and our solution. (a) The task-irrelevant features and spurious attention in spatial attention hinder causal class-specific patterns for MLC of medical images. While Information Bottleneck (IB) attention cannot mitigate the confounder factor (b) and causal attention cannot reduce noise information (c), we propose IB-based causal attention to address this challenge (d).

seeks to learn a single model that predicts multiple labels simultaneously for a given input, thus promoting more efficient training. Therefore, it is crucial to learn comprehensive representations and effectively separate them into class-specific and class-irrelevant within each instance for accurate predictions.

Attention has demonstrated its effectiveness in learning class-specific representations to improve classification accuracy in MLC tasks. By capturing different regions [25, 26] or semantics [24, 28] occupied by objects from various categories, attention facilitates the learning of robust class-specific features for MLC. However, attention cannot always access the causal factors and may incorrectly focus on the wrong context regions when training samples are insufficient [9]. This contextual bias is considered confounder, leading to spurious correlations. This can be mitigated by causality learning [2, 9] by highlighting the underlying causal regions in an image for more robust class-specific features.

Despite promising results of causality-based attention for class-specific feature learning, they still struggle to minimize redundancy information to ensure the most discriminative causal attention for each class. Redundancy is introduced when the attention mechanism captures class-irrelevant features, such as background and inherent anatomical structures in medical images, reducing its ability to interpret the cause. For example, polyps and ulcers exhibit similar textures under varying lighting conditions (Fig. 1 (a)). This results in both causal factors and class-irrelevant features being incorporated into the attention, which inevitably introduces noise. While the Information Bottleneck (IB) principle [17] can be applied to restrict information bypassed by attention [6], it overly relies on label-related attention and ignores its causality. This limitation motivates the incorporation of IB into causal attention mechanisms. However, applying a

spherical Gaussian prior to the latent class-specific features reduces the model's ability to differentiate between multi-label class-specific attentions in MLC tasks.

To address this challenge, a new structural causal model (SCM) is constructed to regard class-specific attention as a mixture of causal, spurious, and noisy factors, and Information Bottleneck-based Causal Attention (IBCA) is proposed to learn discriminative causal class-specific attention for MLC of medical images (Fig. 1 (d)). Specifically, class-irrelevant (noise) information is eliminated by Gaussian mixture variational IB to learn class-specific attention, which is incorporated into contrastive enhancement-based causal intervention to mitigate spurious and noise attention for discriminative causal class-specific features. Building upon this SCM, our **contributions** are: (1) For the first time, IBCA incorporates IB into causality learning for MLC of medical images, effectively separating spurious and noisy information from the causal class-specific attention. (2) IBCA advances causality learning by gradually incorporating Gaussian mixture multi-label spatial attention into causal intervention through contrastive learning. This enables the learning of discriminative causal attention with enhanced interpretability.

2 Information Bottleneck-based Causal Attention

2.1 Causal Inference for MLC of Medical Images

Attention Information Bottleneck. For MLC tasks, the image feature X, class-specific attention A, latent class-specific feature Z, and prediction Y follow the joint conditional distribution of the model shown in Fig. 1 (b). The mutual information between Z and Y is $I(Z;Y) = \int_{z\in Z}\int_{y\in Y} p(z,y)\log\frac{p(z,y)}{p(z)p(y)}\,dz\,dy$. To remove label-irrelevant information in A, IB is applied to optimize the objective: $\min_A -I(Z;Y) + \beta I(Z;A,X)$, where $\beta > 0$ controls the trade-off between two MI terms, encouraging A to learn minimal redundancy attention that is most predictive of Z. Based on variational approximation [1], the lower bound of the attentive variational information bottleneck can be obtained by:

$$\mathcal{L}_{VIB} = I(Z;Y) - \beta I(Z;A,X). \tag{1}$$

The first term is the negative loglikelihood of the prediction, aiming to learn Z that can maximally obtain the correct prediction. The second term aims to minimize the MI between X and Z with given A.

Causal Attention. An SCM is built as a causal view to describe the causality among four variables in MLC of medical images: X, A, Y, and confounder C in Fig. 1 (c). C influences X, which in turn affects A, introducing a spurious correlation into the desired causal effect: $X \rightarrow A \rightarrow Y$, resulting in spurious correlations: $C \rightarrow X \rightarrow A \rightarrow Y$. To eliminate the confounding effect caused by C, causal intervention cuts off the links from C to X and A to identify the causal attention. This is implemented by a backdoor adjustment with do-calculus [15]: $\mathrm{P}(Y \mid \mathrm{do}(X))$. To approximate the confounding effects, sampling all possible confounding impacts can be approximated by sampling N times on the observed

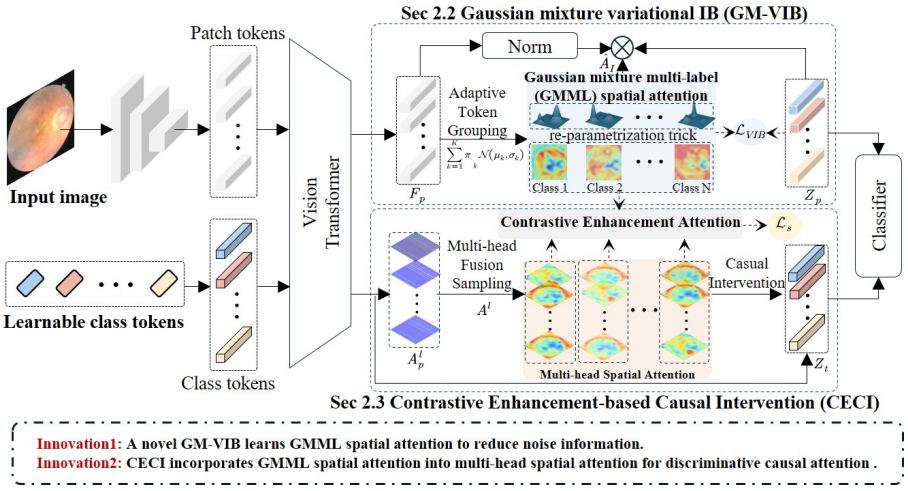

Fig. 2. Overview of our IBCA model, which is built on a transformer to learn Gaussian mixture multi-label spatial attention for noise reduction and is further integrated with contrastive enhancement-based causal intervention to achieve discriminative causal attention learning.

data (x, y) [14]. In this way, different class-specific attention a_k^n for class k of sample n is obtained, and the causal intervention can be modeled as the sigmoid activated classification probability of the class-specific features:

$$\begin{aligned} P(Y=k|do(X=x)) &= \sum_{n=1}^{N} \frac{P(Y=k|X=x^n)P(X=x^n)}{P(X=x^n|C=c)} \\ &= \sum_{n=1}^{N} \frac{\text{Sigmoid}(\text{Clf}(a_k^n x^n))P(a_k^n)}{P(a_k^n|c)}, \end{aligned} \tag{2}$$

where $\text{Clf}(\cdot)$ denotes the classifier corresponding to each class k, and $a_k^n x^n$ represents the class-specific features. The term $\frac{P(a_k^n)}{P(a_k^n|c)}$ indicates the weight of each spatial attention sample, which is set to $1/N$ due to uniform sampling in the multi-head attention mechanism. Further implementation details are provided in Section 2.3.

Information Bottleneck-based Causal Attention. We construct an SCM of IB-based causal attention by incorporating noise **N** into the causal attention mechanism (Fig. 1 (d)). Although **N** introduces no spurious correlation with Y, it mitigates the discriminative ability of causal attention A_c. We propose explicitly separating attention into noise, causal, and spurious components, and design a simple IBCA approach that introduces contrastive enhancement attention with IB and causal intervention. Details will be provided in the following two sections.

2.2 Gaussian Mixture Variational Information Bottleneck

Our Gaussian Mixture Variational Information Bottleneck (GM-VIB) learns Gaussian mixture multi-label (GMML) spatial attention to effectively filter out class-irrelevant information. The distribution parameters of each Gaussian component are learned by adaptive token grouping and the VIB constraint.

Adaptive Token Grouping. Given an input image x, its multi-label spatial attention maps A_I are derived from the patch tokens F_p learned by a multi-class token transformer [21]. Each spatial attention map follows a Gaussian mixture distribution, with the number of Gaussian components equal to the number of multi-label classes N_c. For each Gaussian component, the reparameterization trick [1] is used to calculate its mean vector μ_k, covariance matrix σ_k, and coefficient π_k. This is implemented by applying three different multiple liner projection layers to F_p. In this way, the Gaussian mixture representation is learned as $\sum_{k=1}^{K} \pi_k \mathcal{N}(\mu_k, \sigma_k), k \in \{1, 2, \cdots, N_c\}$. π_k satisfies $0 \leq \pi_k \leq 1, \sum_{k=1}^{N_c} \pi_k = 1$.

Variational Information Bottleneck. To ensure that the learned spatial attention is class-specific, different Gaussian mixture multi-label spatial attention samples are obtained by resampling from the Gaussian mixed distribution to derive Z in Eq. 1. Specifically, the attention samples $\hat{A}_I$ are obtained by sampling a new set of $\Pi = \{\pi_1, \pi_2, \cdots, \pi_K\}$ from a Gamma distribution for each class [4], and sampling each vector independently from a component of the prior Gaussian distribution G^p. The class-aware features Z_p are then learned by multiplying F_p with $\hat{A}_I$ as: $Z_p = Norm(F_p) \times \hat{A}_I, Z_p \in \mathbb{R}^{C \times D}$, where $Norm(\cdot)$ denotes normalization operation. To produce class scores, Z_p is passed through a Global Average Pooling (GAP) layer, which is supervised by a Multi-Label Soft Margin (MLSA) loss. Thus, the first term in Eq. 1 can be rewritten as $L_{mlsm}\{GAP(Z_p), y\}$. We set G^p as a fixed spherical Gaussian: $G^p = \mathcal{N}(0, 1)$, and the second term in Eq. 1 can be substituted by calculating the Kullback-Leibler (KL) divergence between the approximate posterior G_p and the true posterior $\mathcal{N}(\mu_k, \sigma_k)$ by:

$$KL(G_k^q \parallel G^p) = \frac{1}{2} \sum_{k=1}^{N} (\mu_k^2 + \sigma_k^2 - \log(\sigma_k) - n). \tag{3}$$

We assume $n = 1$ to simplify the derivation process. Eq. (1) is rewritten as: $\mathcal{L}_{VIB} = L_{mlsm}\{GAP(Z_p), y\} + \beta KL(G_k^q \parallel G^p)$. This constrains spatial attention by reducing noise to highlight the most contributing regions for each class.

2.3 Contrastive Enhancement-based Causal Intervention

Our Contrastive Enhancement-based Causal Intervention (CECI) gradually incorporates GMML spatial attention into Multi-Head Fusion Sampling (MHFS)-based causal intervention, constraining noise information to enable discriminative causal attention by a Contrastive Enhancement Attention (CAE) loss.

MHFS-based Causal Intervention. Multi-head self-attention is incorporated into the causal intervention due to the embedded class-specific information in it for a multi-class token transformer. To generate class-specific attentions of

MHFS, for each head self-attention map $A_p^l \in \mathbb{R}^{(N_p^2+N_c)\times(N_p^2+N_c)}$ from the last transformer block, the class-to-patch attention $A_{c2p}^l = A_p^l[1:N_c, N_c+1:N_c+N^2]$ and patch-to-patch affinity $A_{p2p}^l = A_p^l[N_c+1:N_c+N_p^2, N_c+1:N_c+N_p^2]$ are multiplied to generate a class-specific attention sample $A^l \in \mathbb{R}^{N_c\times N_p^2}$, where N_p is the patch size. Instead of directly applying each class-specific attention sample to the classification to ensure causality, all token embeddings are incorporated into the classification [21]. Since patch-tokens are incorporated in the classification in the VIB loss, a GAP is applied to multiple class-tokens F_t to learn a class-specific feature Z_t, with MLSA loss L_t to constrain the training.
CAE aligns each A^l with Gaussian mixture spatial attention to reduce noise information in the causal intervention. Specifically, each attention is activated by a sigmoid function, and the pairwise cosine similarity between A^l and $\hat{A}_I$ is calculated and averaged across all heads to formulate the CAE loss $\mathcal{L}_s$ by:

$$\mathcal{L}_s = \frac{1}{H}\sum_{l=1}^{H}\mathcal{L}_s^l, \quad \mathcal{L}_s^l = 1 - \frac{\hat{A}_I A^l}{\|\hat{A}_I\|\|A^l\|}. \tag{4}$$

H is the number of attention heads. This forces each A^l to receive stronger and more frequent guidance to learn discriminative causal class-specific information.

The final objective to minimize is simply the summation of different losses: $\mathcal{L} = \mathcal{L}_{VIB} + \mathcal{L}_t + \lambda_s\mathcal{L}_s$, where λ_s is the trade-off weights and we set 0.01 here.

3 Experiments and Results

3.1 Dataset and Implementation Details

Dataset and Evaluation. We conduct experiments on two publicly available datasets: Endo [20] and MuReD [8]. Endo consists of 3865 colonoscopy images, annotated with 4 types of lesions. MuReD includes 2,208 samples collected from three different sources (ARIA [3], STARE [5], and RFMiD [13]), covering a wider range of diseases with 20 label types. Both datasets are randomly split into training, validation, and test sets in an 8:1:1 ratio.

In addition, we compare our methods with several recent state-of-the-arts: 1) Basic methods: TA-DCL [23] with intra-pool contrastive learning, TS-Former [27], MLSL-Net [22], C-Tran [7], and Q2L [11]), 2) Causality learning-based methods: CCD [10], IDA [9], and Multilevel Causality (ML-C) [2]. Mean average precision (mAP), class-wise recall (CR), class-wise F1 score (CF1), overall recall (OR), and overall F1 score (OF1) are used as evaluation metrics.
Implementation Details. We train our IBCA model on an NVIDIA Tesla A40 GPU. A hybrid ViT Backbone (ResNet-50+ViT-B_16) [19] with multiple class tokens is employed in our experiments, and all ViT-based methods share this backbone. Adam optimizer with an initial learning rate of $1e-4$ is adopted for a batch size of 64. The training epoch is set to 200, and β in $\mathcal{L}_{VIB}$ is set to 0.001. During inference, the mean average of patch-token and class-token-based predictions is used for the final prediction. Our code is available on GitHub [5].

[5] https://github.com/rabbittsui/IBCA

Table 1. Experiment results in comparisons with SOTAs on MuRed and Endo.

Models	MuRed					Endo				
	CR↑	CF1↑	OR↑	OF1↑	mAP↑	CR↑	CF1↑	OR↑	OF1↑	mAP↑
CTRAN [7]	39.44	40.26	51.58	60.12	59.11	53.36	54.99	51.42	53.43	61.28
Q2L [11]	61.32	48.28	71.93	61.84	61.42	61.48	65.37	61.32	64.52	67.95
TS-Former [27]	39.16	44.99	54.04	62.86	61.97	39.16	44.99	54.04	62.86	61.97
MLSL-Net [22]	52.04	54.91	64.91	66.43	60.35	62.98	67.53	65.09	68.15	71.37
TA-DCL∗ [23]	46.05	50.50	52.63	60.48	56.54	63.92	64.86	60.38	62.29	64.34
IDA [9]	55.74	58.18	63.86	67.41	63.42	64.71	70.04	65.09	69.52	74.96
CCD [10]	60.39	57.15	64.21	66.79	64.90	61.59	70.40	62.26	69.47	74.03
ML-C [2]	59.69	59.59	64.91	69.16	68.41	60.95	66.48	63.21	67.17	74.21
Ours	**66.74**	**59.90**	**72.63**	**69.22**	**73.43**	**66.18**	**72.05**	**66.04**	**70.00**	**76.38**

Table 2. Ablation study of our proposed modules on MuRed and Endo.

Methods	MuRed					Endo				
	CR↑	CF1↑	OR↑	OF1↑	mAP↑	CR↑	CF1↑	OR↑	OF1↑	mAP↑
Basic	59.26	57.80	63.16	68.77	72.81	63.00	68.44	64.62	69.02	75.34
Single VIB	66.11	58.71	68.42	67.36	73.28	63.65	68.63	65.57	69.50	76.14
GMM VIB	65.11	59.64	69.82	**69.82**	72.59	65.70	70.62	65.57	**70.03**	76.38
Ours	**66.74**	**59.90**	**72.63**	69.22	**73.43**	**66.18**	**72.05**	**66.04**	70.00	**76.38**

3.2 Comparison with State-of-the-art Methods

We validated the effectiveness of the proposed IBCA on the MuRed and Endo datasets in Table 1. It is observed that the causality learning-based method outperforms the basic methods on most metrics, due to its efficient de-confounding of the contextual bias by different causal intervention strategies. However, our method incorporates the information bottleneck theory into the causal attention of ViT, and possesses superior capability on both datasets. Compared to ML-C [2], our method brings improvements with increments of 7.05% in CR, 7.72% in OR, and 4.95% in mAP for MuRed, respectively. Compared to IDA [9], our method brings improvements with increments of 1.74% in CR, 2.01% in CF1, and 1.42% in mAP for Endo, respectively. These results demonstrate the effectiveness of our approach in learning minimal redundancy and the most discriminative causal class-aware features for different labels.

3.3 Ablation Study

An ablation study is conducted to evaluate the effectiveness of Gaussian mixture multi-label spatial attention, VIB loss, and contrastive enhancement attention in our method. Results are shown in Table 2. In the first column of Table 2, Basic refers to removing all three key components, with a fully connected layer to learn spatial mapping in the patch-token path. Single VIB substitutes the Gaussian Mixture Model (GMM) distribution with a single Gaussian distribution without $\mathcal{L}_s$, and GMM VIB represents our GMM-based spatial attention with

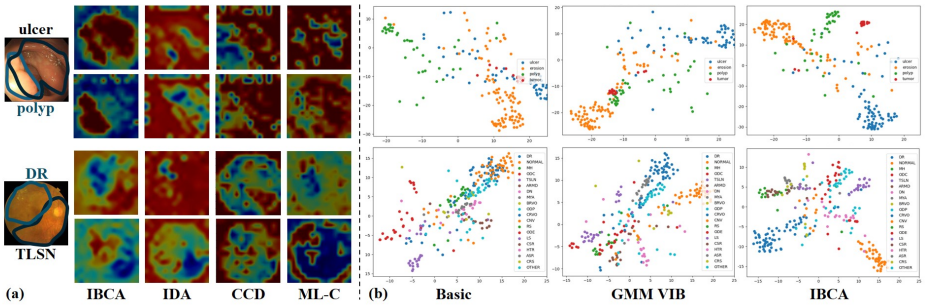

Fig. 3. Visualization results demonstrated the effectiveness of our method. (a) Visualization of class-specific spatial attention from different methods. (b) T-SNE results of class-specific features from different settings in ablation study.

VIB without $\mathcal{L}_s$. Compared to the basic architecture, both Single VIB and GMM VIB show significant improvements, particularly on MuReD, with increases of 6.85% and 5.26% in CR and OR, respectively. This suggests the effectiveness of the IB in filtering out task-irrelevant information, particularly when the number of predefined class labels is smaller. Notably, GMM VIB outperforms Single VIB across most metrics. For MuReD, CF1 improves by 0.93%, OR by 1.40%, and OF1 by 2.46%; for Endo, CR increases by 2.05%, CF1 by 1.99%, and OF1 by 0.47%. These results highlight the superiority of our GMM-based multi-label spatial attention in reducing class-irreverent features. Finally, $\mathcal{L}_s$ incorporates the IB constraint to reduce the noise factor from causal intervention, achieving the best performance in most metrics.

To intuitively demonstrate the effectiveness of our IBCA, we visualize the class-specific spatial attention with correct predictions of different causal-based methods. The corresponding lesion regions are highlighted in Fig.3 to investigate the interpretability of learned attention maps. It is observed that the task-irrelevant information and spurious correlation are contained in the attention maps from IDA and CCD, while ML-C tends to contain some task-irrelevant information with a causal part. In comparison to these methods, our IBCS captures more comprehensive lesion regions, especially the emphasis on the retinal vein lesions in fundus images, which contributes to better interpretability for diagnosis. In addition, t-SNE [12] is applied to visualize the distribution of patch-token-based class-specific features, as shown in Fig. 3. Clear boundaries between classes are observed on both the Endo and MuRed datasets, demonstrating that our method effectively discriminates class-specific features by leveraging GMM-based spatial attention.

4 Conclusion

To the best of our knowledge, our IBCA is the first to incorporate IB into causality learning, enabling the learning of discriminative causal class-specific

attention for multi-label medical image recognition. IBCA effectively learns IB-based Gaussian mixture multi-label spatial attention to mitigate the influence of task-irrelevant information. This Gaussian-mixed attention is further integrated into MHFS-based causal intervention to separate the redundancy noise from causal class-specific attention for class-specific feature learning. Extensive experiments on two public datasets demonstrate that our method outperforms existing methods by remarkable margins and shows better interpretation ability.

Acknowledgments. This study was funded by Shandong Natural Science Foundation under Grant ZR2024QF209.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Alemi, A.A., Fischer, I., Dillon, J.V., Murphy, K.: Deep variational information bottleneck. arXiv preprint arXiv:1612.00410 (2016)
2. Cui, X., Jiang, S., Sun, B., Li, Y., Cao, Y., Li, Z., Lv, C., Liu, Z., Cui, L., Li, S.: Multilevel causality learning for multi-label gastric atrophy diagnosis. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 682–692. Springer (2024)
3. Farnell, D.J., Hatfield, F.N., Knox, P., Reakes, M., Spencer, S., Parry, D., Harding, S.P.: Enhancement of blood vessels in digital fundus photographs via the application of multiscale line operators. Journal of the Franklin institute **345**(7), 748–765 (2008)
4. Henderson, J., Fehr, F.J.: A vae for transformers with nonparametric variational information bottleneck. In: The Eleventh International Conference on Learning Representations (2023)
5. Hoover, A., Kouznetsova, V., Goldbaum, M.: Locating blood vessels in retinal images by piecewise threshold probing of a matched filter response. IEEE Transactions on Medical imaging **19**(3), 203–210 (2000)
6. Lai, Q., Li, Y., Zeng, A., Liu, M., Sun, H., Xu, Q.: Information bottleneck approach to spatial attention learning. arXiv preprint arXiv:2108.03418 (2021)
7. Lanchantin, J., Wang, T., Ordonez, V., Qi, Y.: General multi-label image classification with transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 16478–16488 (June 2021)
8. Li, N., Li, T., Hu, C., Wang, K., Kang, H.: A benchmark of ocular disease intelligent recognition: One shot for multi-disease detection. In: Benchmarking, Measuring, and Optimizing: Third BenchCouncil International Symposium, Bench 2020, Virtual Event, November 15–16, 2020, Revised Selected Papers 3. pp. 177–193. Springer (2021)
9. Liu, R., Huang, J., Li, T.H., Li, G.: Causality compensated attention for contextual biased visual recognition. In: The Eleventh International Conference on Learning Representations (2022)
10. Liu, R., Liu, H., Li, G., Hou, H., Yu, T., Yang, T.: Contextual debiasing for visual recognition with causal mechanisms. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12755–12765 (2022)

11. Liu, S., Zhang, L., Yang, X., Su, H., Zhu, J.: Query2label: A simple transformer way to multi-label classification. arXiv preprint arXiv:2107.10834 (2021)
12. Van der Maaten, L., Hinton, G.: Visualizing data using t-sne. Journal of machine learning research **9**(11) (2008)
13. Pachade, S., Porwal, P., Thulkar, D., Kokare, M., Deshmukh, G., Sahasrabuddhe, V., Giancardo, L., Quellec, G., Mériaudeau, F.: Retinal fundus multi-disease image dataset (rfmid): a dataset for multi-disease detection research. Data **6**(2), 14 (2021)
14. Pearl, J.: Causal inference in statistics: An overview (2009)
15. Pearl, J., Glymour, M., Jewell, N.P.: Causal inference in statistics: A primer. John Wiley & Sons (2016)
16. Rodríguez, M.A., AlMarzouqi, H., Liatsis, P.: Multi-label retinal disease classification using transformers. IEEE Journal of Biomedical and Health Informatics **27**(6), 2739–2750 (2022)
17. Saxe, A.M., Bansal, Y., Dapello, J., Advani, M., Kolchinsky, A., Tracey, B.D., Cox, D.D.: On the information bottleneck theory of deep learning. In: International Conference on Learning Representations (2018)
18. Schölkopf, B., Locatello, F., Bauer, S., Ke, N.R., Kalchbrenner, N., Goyal, A., Bengio, Y.: Toward causal representation learning. Proceedings of the IEEE **109**(5), 612–634 (2021)
19. Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., Jégou, H.: Training data-efficient image transformers & distillation through attention. In: International conference on machine learning. pp. 10347–10357. PMLR (2021)
20. Wang, D., Wang, X., Wang, L., Li, M., Da, Q., Liu, X., Gao, X., Shen, J., He, J., Shen, T., et al.: A real-world dataset and benchmark for foundation model adaptation in medical image classification. Scientific Data **10**(1), 574 (2023)
21. Xu, L., Bennamoun, M., Boussaid, F., Laga, H., Ouyang, W., Xu, D.: Mctformer+: Multi-class token transformer for weakly supervised semantic segmentation. IEEE transactions on pattern analysis and machine intelligence (2024)
22. Yi, L., Zhang, L., Xu, X., Guo, J.: Multi-label softmax networks for pulmonary nodule classification using unbalanced and dependent categories. IEEE Transactions on Medical Imaging **42**(1), 317–328 (2022)
23. Zhang, Y., Luo, L., Dou, Q., Heng, P.A.: Triplet attention and dual-pool contrastive learning for clinic-driven multi-label medical image classification. Medical Image Analysis **86**, 102772 (2023)
24. Zhao, J., Yan, K., Zhao, Y., Guo, X., Huang, F., Li, J.: Transformer-based dual relation graph for multi-label image recognition. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 163–172 (2021)
25. Zhou, W., Xia, Z., Dou, P., Su, T., Hu, H.: Aligning image semantics and label concepts for image multi-label classification. ACM Transactions on Multimedia Computing, Communications and Applications **19**(2), 1–23 (2023)
26. Zhu, K., Wu, J.: Residual attention: A simple but effective method for multi-label recognition. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 184–193 (2021)
27. Zhu, X., Cao, J., Ge, J., Liu, W., Liu, B.: Two-stream transformer for multi-label image classification. In: Proceedings of the 30th ACM International Conference on Multimedia. pp. 3598–3607 (2022)
28. Zhu, X., Li, J., Cao, J., Tang, D., Liu, J., Liu, B.: Semantic-guided representation enhancement for multi-label image classification. IEEE Transactions on Circuits and Systems for Video Technology (2024)