Multilevel Causality Learning for Multi-label Gastric Atrophy Diagnosis

Xiaoxiao Cui[1], Shanzhi Jiang[2], Baolin Sun[2], Yiran Li[2], Yankun Cao[1], Zhen Li[3], Chaoyang Lv[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 Gastroenterology, Qilu Hospital of Shandong University, Jinan, Shandong 250012, China
[4] Department of Gastroenterology, Linyi County People's Hospital, Dezhou 251599, China
[5] Case Western Reserve University, Cleveland, OH 44106, USA

Abstract. No studies have formulated endoscopic classification (EG) of gastric atrophy (GA) as a multi-label classification (MLC) problem, which requires the simultaneous detection of GA and its gastric sites during an endoscopic examination. Accurate EG of GA is crucial for assessing the progression of early gastric cancer. However, the strong visual interference in endoscopic images is caused by various inter-image differences and subtle intra-image differences, leading to confounding contexts and hindering the causalities between class-aware features (CAFs) and multi-label predictions. We propose a multilevel causality learning approach for multi-label gastric atrophy diagnosis for the first time, to learn robust causal CAFs by de-confounding multilevel confounders. Our multilevel causal model is built based on a transformer to construct a multilevel confounder set and implement a progressive causal intervention (PCI) on it. Specifically, the confounder set is constructed by a dual token path sampling module that leverages multiple class tokens and different hidden states of patch tokens to stratify various visual interference. PCI involves attention-based sample-level re-weighting and uncertainty-guided logit-level modulation. Comparative experiments on an endoscopic dataset demonstrate the significant improvement of our model, such as IDA (0.95% on OP, and 0.65% on mAP) and TS-Former (1.11% on OP, and 1.05% on mAP).

Keywords: Multi-label Classification · Causal Intervention · Gastric Atrophy Detection.

1 Introduction

No studies have investigated computer-aided endoscopic grading (EG) of gastric atrophy (GA) according to the Kimura-Takemoto classification (KTc) [7] using a

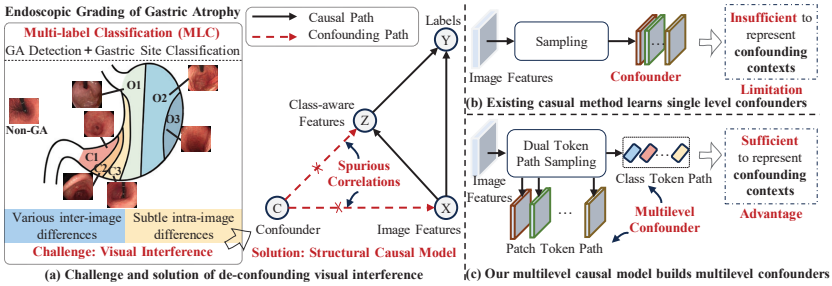

Fig. 1. Problem setting and our method. (a) Visual interference in endoscopic images leads to spurious correlation between image features and labels, which can be deconfounded by a causal model. (b) Limitation of existing method and (c) advantage of our multilevel causal model on building confounders to capture the visual interference.

multi-label classification (MLC) approach [4, 9]. EG of GA is crucial for screening patients with an increased risk of progression to malignancy by detecting GA and classifying its gastric site while scanning the entire gastric mucosa via endoscopy. Despite the demonstrated superiority of deep learning-based methods in GA detection or gastric site classification [1, 21, 8], accurate EG of GA remains challenging due to strong visual interference during endoscopic examination. Furthermore, training independent classifiers for each task is expensive and difficult [23], while MLC holds potential in clinical applications [17, 22].

Although the effectiveness of MLC has been demonstrated [14], the learned features are prone to stem from spurious correlations among different labels. On one hand, label correlation methods [24, 2, 19] ignore the true relationship between features and labels due to the interdependencies among labels [3, 13]. On the other hand, attention in image region-level [20] or semantic-level [25] may not always capture meaningful factors for improvement. As it can inadvertently attend to confounding contexts when training samples are insufficient [10], leading to incorrect causalities between learned features and predictions. Therefore, learning a robust class-aware feature (CAF) in MLC is crucial and can enhance the explainability of the model. This can be addressed by causality theory [15], which provides a theoretical perspective by studying the inherent relationship between an initial event (the cause) and a subsequent event (the effect).

However, existing causal intervention methods often construct confounders at the same feature level, which is insufficient to represent confounding contexts due to the visual interference in endoscopic images. Intuitively, the graphic difference interference is caused by the inter-image difference due to various gastric sites and endoscopic conditions in endoscopic examinations (Fig. 1(a)). Moreover, the graphic similarity interference in GA detection stems from the subtle intra-image differences between normal and abnormal mucosa. Therefore, relying solely on a single level of features may not capture these visual interferences very well, hindering the inference of causality between CAFs and labels (Fig. 1(b)).

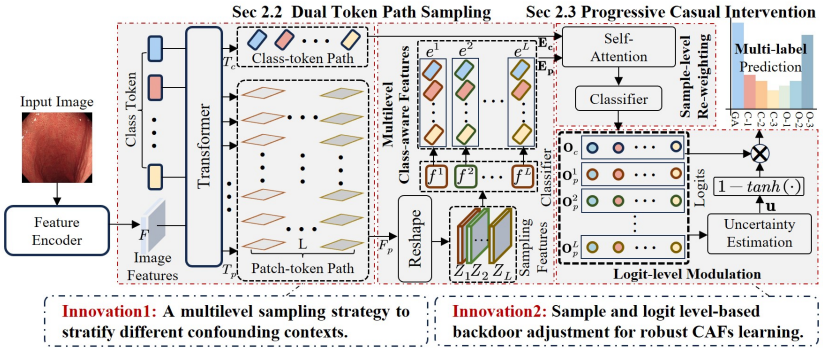

Fig. 2. Illustration of our multilevel causal model, which is built upon a transformer to capture multilevel confounding contexts and implement progressive intervention.

A structural causal model (SCM) is constructed for the EG of GA, and a novel multilevel causality learning approach is proposed to learn the robust causal relationship between CAFs and labels in MLC (Fig. 1(c)). Our multilevel causality learning enriches confounders with multilevel feature representation and implements a progressive causal intervention (PCI) (Fig. 2). Building upon this SCM, our **contributions** can be summarized as follows: **(1)** For the first time, a multilevel causality learning approach for multi-label gastric atrophy diagnosis is proposed to address the strong visual interference in endoscopic images. **(2)** Our multilevel causal model innovatively constructs multilevel confounders to stratify various confounding contexts and advances causal intervention to derive CAFs that are robust to visual interference. **(3)** An endoscopic dataset is collected to study the effectiveness of our multilevel causal model.

2 Multilevel Causality Multi-label Learning

2.1 Causal Inference for Endoscopic Grading of Gastric Atrophy

A causal view of visual interference in EG of GA is formulated by using SCM. In this SCM, causality among four variables (Fig. 1(a)): image feature X, confounder C, class-aware feature Z, and prediction Y are denoted by direct links. The existence of C introduces interference with the desired causal effect: $X \to Y$, resulting in spurious correlations between X and Y: $C \to X \to Z \to Y$, $C \to Z \to Y$, and $C \to X \to Y$. The way to eliminate the confounding effect is causal intervention. It cuts off the links from C to X and Z to build a beneficial causal effect for robust classification: $X \to Z \to Y$ and is implemented by a backdoor adjustment with do-calculus [16] by:

$$\begin{aligned} \mathrm{P}(Y \mid \mathrm{do}(X)) &= \sum_{c} \mathrm{P}(Y \mid X, C=c)\mathrm{P}(C=c) \\ &= \sum_{c}\sum_{z} \mathrm{P}(Y \mid X=x, C=c, Z=z)\mathrm{P}(Z \mid X=x)\mathrm{P}(C=c) \end{aligned} \tag{1}$$

$\mathrm{P}(Y \mid X, C = c)$ is the prediction of the classifier trained in possible confounders c. $\mathrm{P}(Y \mid X = x, Z = z, C = c)$ denotes the classification probability of Z from X. This process is implemented by sigmoid activated classification of spatial grouping of X. According to [15], sampling on c can be approximated by sampling on the observed data (x, y). Instead of sampling C in a fixed feature level to implement the confounder, a multilevel sampling strategy is proposed to deconfound visual interference in endoscopic images. To calculate $\mathrm{P}(Z \mid X = x)\mathrm{P}(C = c)$, sample-level re-weighting and logit-level modulation are proposed. The details will be described in the following two sections.

Summary of Advantage: An SCM clearly illustrates how confounders bias the MLC model, and how to cut off the confounding path.

2.2 Dual Token Path Sampling For Multilevel Confounder

Our dual token path sampling (DTPS) module captures various confounding contexts in endoscopy images by leveraging multilevel features from both class-token and different patch-token paths of a transformer, while multilevel CAFs are further derived via spatial grouping.

Dual Token Path Sampling: For an image feature F generated from ResNet [6], a standard transformer is introduced to project it into a set of feature queries affected by different confounders. F is split into $N \times N$ patches and then transformed into N^2 patch tokens $T_p \in \mathbb{R}^{N^2 \times D}$, where D is the embedding dimension. N_c learnable class tokens are designed and transformed into $T_c \in \mathbb{R}^{N_c \times D}$ to model global features along the spatial dimension. T_p and T_c are concatenated and then added with position embeddings to form input tokens to the transformer, which consists of multiple consecutive encoder blocks equipped with a multi-head self-attention module and a multilayer perceptron module. After that, image features affected by the confounders can be estimated from class-token and intermediate hidden states of patch-token paths.

Multilevel CAFs: To derive CAFs, spatial grouping is proposed to cluster pixels within an image towards different groups most likely to represent specific categories by a classifier. In the class-token path, T_c is directly used to formulate different CAFs represented by $\mathbf{E_c}$. In the patch-token path, each hidden state form $F_p \in \mathbb{R}^{N^2 \times D \times L}$ is stratified to derive different CAFs by a spatial grouping module, where L is the number of encoder blocks. For each $F_p^k, k = 1, 2, \cdots, L$, a classifiers $f^k(\cdot)$ implemented by a 1×1 convolution is utilized to implement spatial grouping to generate CAFs $\left\{e_n^k\right\}$, where $n = 1, 2, \cdots, N_c$. Specifically, F_p^k is arranged into a 3D tensor $Z_k \in \mathbb{R}^{D \times N \times N}$, and the probability of each pixel belong to label n is calculated by using weight vectors W_n^k of $f^k(\cdot)$ and a softmax operator:

$$A_n^k = softmax(W_n^k Z_k) \tag{2}$$

where $A_n^k \in \mathbb{R}^{N \times N}$ and $\sum_{i,j} a_n^{i,j} = 1$ for all n. Finally, for each label n, its corresponding CAF $\left\{\mathbf{e}_n^k\right\}$ is computed by using A_k^n to weight average the spatial features in Z_k: $\mathbf{e}_n^k = Z_k A_n^k$. In this way, a set of CAFs $\mathbf{E_p} = \left\{\mathbf{e}_n^k\right\}_{k=1}^{L}$ concatenate

with $\mathbf{E_c}$ formulate multilevel $\mathbf{E} \in \mathbb{R}^{(L+1)\times N_c \times D}$.
Summary of Advantage: Our DTPS is a simple sampling manner to enrich confounders with different feature levels, while pixels in features are clustered into coherent groups based on class awareness to derive multilevel CAFs.

2.3 Progressive Causal Intervention

Our PCI progressively incorporates sample-level re-weighting and logit-level modulation into the classification, enhancing causalities between CAFs and labels. Because confounders are sampled from different transformer blocks, the status of each CAFs varies. Moreover, different channels of each CAF attend to different label-related regions [26], treating their contribution equally hinders CAF learning [10]. Based on the multilevel CAFs, attention-based re-weighting and uncertainty-guided modulation are designed to bias more weight to more crucial samples under different confounders.
Attention-based Re-weighting. Self-attention is applied to operate the weighting upon different samples in $\mathbf{E}$ to emphasize more crucial samples:

$$\mathbf{E}' = \text{softmax}\left(\frac{(W_q\mathbf{E})\,(W_k\mathbf{E})^T}{\sqrt{D}}\right)(W_v\mathbf{E}) \tag{3}$$

where W_q, W_k, and W_v are three different linear projections to map CAF sequences into a common subspace for similarity measure.
Uncertainty-guided Modulation. To obtain the final prediction, each CAF is fed into a classifier, then logits of the class-token sample are modulated by all samples to obtain the final logit for that class. Specifically, for patch-token samples in $\mathbf{E}'$, its classifier shares the same learnable weights with $f^k(\cdot)$ to generate the prediction $\mathbf{O}_p$. The class token prediction $\mathbf{O}_c$ is calculated by an identity mapping of from class-token sample in $\mathbf{E}'$. The variance along the $L+1$ dimension of $\mathbf{O} = [\mathbf{O}_p, \mathbf{O}_c] \in \mathbb{R}^{C\times D\times(L+1)}$ that consists of multiple outputs from multilevel CAFs is calculated. Such variance is denoted by a tensor $\mathbf{u} \in \mathbb{R}^{C\times D}$ and reflects the uncertainty for every class logit across multiple samples. It is then converted into certainty by using $1 - tanh(\mathbf{u})$ to modulate logits from $\mathbf{O}_c$:

$$\mathbf{O}'_c = \mathbf{O}_c \otimes (1 - tanh(\mathbf{u})) \tag{4}$$

where $\mathbf{O}'_c$ is the final prediction after causal intervention, which is then put into the Binary Cross-Entropy Loss with sigmoid to guide the training process.
Summary of Advantage: PCI novelly implements backdoor adjustment at sample-level and logit-level by incorporating attention-based re-weighting and uncertainty-guided modulation, respectively.

3 Experiment

3.1 Dataset and Data Preprocessing

An endoscopy dataset consisting of 4574 endoscopy images from 3840 patients at $*$ Hospital was collected with approval from the Institutional Review Board

with a waiver of informed consent. 40.66% of the images represents gastric atrophy (GA), while 59.34% depicts non-GA cases. The distribution across different gastric sites is 1069:1145:370:704:395:891 refers to C1, C2, C3, O1, O2, and O3. Training, validation, and test sets are divided into 3658:455:461, respectively. **Data preprocessing** is performed on our datasets. In detail, the combination of random rotation, random translation, and random scaling is used first for image augmentation. The augmented images are resized to 256×256 and cropped to 224×224 centered on the image.

3.2 Implementation details

ResNet-50+ViT-B_16 [18] is used as a backbone, which integrates ResNet-50 and a transformer that contains 12 transformer blocks with 12 multi-attention heads and an embedding dimension of 768. Adam optimizer with an initial learning rate of $1e-4$ is used for a batch size of 24. Our model was trained for 150 epochs, and the best performance on validation set was selected for testing. All our codes were implemented in Pytorch on an NVIDIA Tesla A40 GPU. Evaluation metrics included mean average precision (mAP), average per-class precision (CP), recall (CR), and F1 score (CF1), and average overall precision (OP), recall (OR), and F1 score (OF1). Our dataset and code are available on GitHub [6].

3.3 Comparison with State-of-the-art Methods

We compare our methods with recent SOTA multi-label classification-based methods: TA-DCL [23], TS-Former [27], C-Tran [9], and Q2L [12]). In addition, two causal inference-based methods, (CCD [11] and IDA [10]), are also included for comparison. TA-DCL is a triplet attention network designed to learn label embeddings. In our task, we only use intra-pool contrastive and denote it as "TA-DCL" in Table 1. TS-Former is a two-stream transformer learning framework that incorporates spatial and semantic features via a multi-shot attention mechanism. C-Tran consists of a transformer encoder trained to predict target labels based on an input set of masked labels and visual features. Q2L leverages transformer decoders to query the existence of a class label. CCD presents a novel causal context debiasing module to pursue the direct effect of an instance, while IDA adopts two attention layers with multiple sampling interventions to compensate attention against the confounder context.

As shown in Table 1, our method consistently outperforms others under fair experimental settings. Notably, because only two labels are required for each input in this task, C-Tran achieves the highest CP of 95.34% with masked labels. IDA obtains the highest OP of 94.92%, highlighting the effectiveness of attention layers with multiple sampling interventions. Both CCD and IDA incorporate causal inference and achieve the second-best performance across all metrics, implying the efficacy of causal inference approach. Our method advances causal inference by incorporating multilevel feature and progressive causal intervention,

[6] https://github.com/rabbittsui/Multilevel-Causal

Table 1. Experiments results demonstrated the effectiveness of our method.

Models	Venue	CP	CR	CF1	OP	OR	OF1	mAP
IDA	ICLR'2023	94.53	90.75	92.11	**94.92**	94.47	94.70	96.91
TA-DCL*	MedIA'2023	87.02	86.05	86.06	87.10	90.68	88.85	92.58
TS-Former	ACM MM'2022	92.11	90.07	90.91	94.46	94.31	94.39	96.51
CCD	CVPR'2022	92.81	92.67	92.47	93.37	94.73	94.54	97.13
C-Tran	CVPR'2021	**95.34**	73.98	79.43	93.41	85.15	89.09	93.56
Q2L	arXiv'2021	85.88	**93.80**	89.23	89.39	94.47	91.86	96.39
Ours		94.22	91.89	**92.65**	94.23	**95.42**	**94.82**	**97.56**

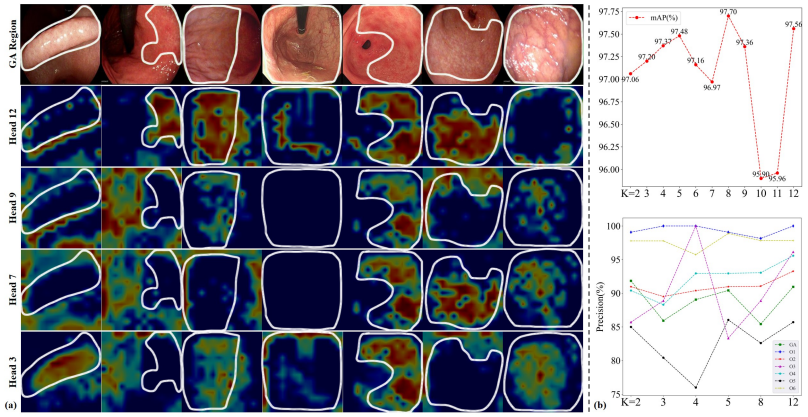

Fig. 3. (a) Some activation maps from different transformer encoder blocks are consistent with GA regoins (white annotation), while some are background. (b) The influence of sampling numbers K on mAP and precision for each class demonstrates that the best performance is achieved when using all multilevel features.

achieving the highest mAP of 97.56%, CF1 of 92.65%, OR of 95.42%, and OF1 of 94.82%. These results demonstrate the effectiveness of our approach in enhancing robust class-aware features for different labels.

To intuitively demonstrate the effectiveness of our method, for each multilevel sampling, we visualize activation maps of each classifier $f^k(\cdot)$ that correctly predicts GA using GradCAM [5]. To study the explainability of activation map, region of GA is also given in Fig. 3(a). It is observed that attention map in different layers attends to different things: some for the GA region, and some for the background. By assigning different attention weights to these regions, accurate activation maps can be enhanced to facilitate the learning of CAF.

3.4 Ablation study

We conduct ablation studies on our multilevel causal model from 2 aspects: 1) components analysis of our method, and 2) varying the number of multilevel

Table 2. Ablation study demonstrated the effectiveness of key components.

Backbone	Sample	Attention	Modulate	CP	CR	CF1	OP	OR	OF1	mAP
✓	×	×	×	86.57	87.50	86.90	88.13	91.47	89.77	93.66
✓	✓	×	×	93.31	89.29	90.81	**94.39**	93.05	93.72	96.86
	✓	✓	×	93.09	83.79	86.31	92.12	88.63	90.34	96.58
✓	✓	×	✓	**95.10**	89.83	91.91	94.31	94.31	94.31	**97.93**
✓	✓	✓	×	92.85	**92.01**	92.25	94.21	95.10	94.69	97.27
✓	✓	✓	✓	94.22	91.89	**92.65**	94.23	**95.42**	**94.82**	97.56

features.

Effects of Key Components. To evaluate the impact of our multilevel sampling, attention, and modulation on causal intervention, ablation experiments are conducted by splitting and reconstructing our model with different components according to the default setting. "Backbone", "Sample", "Attention", and "Modulate" in Table 2 indicate multi-class token transformer, multilevel sampling, self-attention, and uncertainty-guided modulation, respectively. The symbols "✓" and "×" indicate the application or removal of components in our method. A blank "Backbone" in the third row denotes applying a single class token in Transformer. It is observed that using only class-token or patch-token gets limited improvement due to their respective limitation. However, combining them together improves the performance, demonstrating the effectiveness of our multilevel feature sampling. Furthermore, we investigate the effectiveness of attention and modulation on causal intervention in the last four rows of Table 2. It is obvious that either attention or modulation improves the performance across all metrics, while uncertainty-guided modulation yields the highest mAP. Performance is further improved when both attention and modulation are used, indicating the complementary of our key components.

Effects of Sampling Levels. We further investigate the effects of multilevel features by varying numbers of intermediate features in the transformer. For a fair comparison, we gradually incorporate features of shallower blocks of transformer into causal intervention, and the results of mAP and precision for each class are shown in Fig. 3(b). It is observed that mAP fluctuates with the incorporation of features from different transformer blocks, especially dropping features from the first three transformer blocks achieves the lowest. We conclude that different blocks capture different spatial regions, as demonstrated in Fig. 3(a). Therefore, such differences between focused regions can generate different causal-effected CAFs, leading to different predictions for each label.

4 Conclusion

To the best of our knowledge, our multilevel causal model is the first to achieve multi-label gastric atrophy diagnosis. We formulate the SCM to illustrate the confounding effect of strong visual interference in endoscopic images and propose a causal intervention strategy. Our model effectively captures multilevel

confounder contexts by a dual token path sampling module based on a transformer. To mitigate interference, a progressive causal intervention strategy involving sample-level reweighting and logit-level modulation is proposed to reinforce the correct causal relationship between images and labels. Comparative analyses with SOTA methods and ablation studies conducted on our collected GA-related endoscopy dataset demonstrate the effectiveness of our model.

Acknowledgments. This study was funded in part by Key R&D Program of Shandong Province under Grant 2021CXGC010506, in part by the Joint fund for Smart Computing of Shandong Natural Science Foundation under Grant ZR2020LZH013, in part by the Major Scientific and Technological Innovation Project in Shandong Province under Grant 2022CXGC010504, in part by New Universities 20 items Funding Project of Jinan under Grant 2021GXRC108, in part by Shandong Provincial Natural Science Foundation under Grant ZR2022LZH007, and in part by Qingdao Key Technology Research and Industrialization-Future Industry Cultivation Special Project under Grant 22-3-4-xxgg-5-nsh; in part by Qingdao Science and Technology Benefiting the People Demonstration Project 24-1-8-cspz-20-nsh.

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

References

1. Chahal, D., Byrne, M.F.: A primer on artificial intelligence and its application to endoscopy. Gastrointestinal endoscopy **92**(4), 813–820 (2020)
2. Chen, T., Xu, M., Hui, X., Wu, H., Lin, L.: Learning semantic-specific graph representation for multi-label image recognition. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 522–531 (2019)
3. Fallah, H., Bruno, E., Bellot, P., Murisasco, E.: Exploiting label dependencies for multi-label document classification using transformers. In: Proceedings of the ACM Symposium on Document Engineering 2023. pp. 1–4 (2023)
4. Gao, B.B., Zhou, H.Y.: Learning to discover multi-class attentional regions for multi-label image recognition. IEEE Transactions on Image Processing **30**, 5920–5932 (2021)
5. Gildenblat, J., contributors: Pytorch library for cam methods. https://github.com/jacobgil/pytorch-grad-cam (2021)
6. 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)
7. Kimura, K., Takemoto, T.: An endoscopic recognition of the atrophic border and its significance in chronic gastritis. Endoscopy **1**(03), 87–97 (1969)
8. Klang, E., Soroush, A., Nadkarni, G.N., Sharif, K., Lahat, A.: Deep learning and gastric cancer: Systematic review of ai-assisted endoscopy. Diagnostics **13**(24), 3613 (2023)
9. 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. pp. 16478–16488 (2021)
10. 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)

11. 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)
12. 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)
13. Liu, W., Wang, H., Shen, X., Tsang, I.W.: The emerging trends of multi-label learning. IEEE transactions on pattern analysis and machine intelligence **44**(11), 7955–7974 (2021)
14. Nega Tarekegn, A., Ullah, M., Alaya Cheikh, F.: Deep learning for multi-label learning: A comprehensive survey. arXiv e-prints pp. arXiv–2401 (2024)
15. Pearl, J.: Causal inference in statistics: An overview (2009)
16. Pearl, J., Glymour, M., Jewell, N.P.: Causal inference in statistics: A primer. John Wiley & Sons (2016)
17. Tang, P., Yan, X., Nan, Y., Xiang, S., Krammer, S., Lasser, T.: Fusionm4net: A multi-stage multi-modal learning algorithm for multi-label skin lesion classification. Medical Image Analysis **76**, 102307 (2022)
18. 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)
19. Wang, Y., He, D., Li, F., Long, X., Zhou, Z., Ma, J., Wen, S.: Multi-label classification with label graph superimposing. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 34, pp. 12265–12272 (2020)
20. Wang, Z., Chen, T., Li, G., Xu, R., Lin, L.: Multi-label image recognition by recurrently discovering attentional regions. In: Proceedings of the IEEE international conference on computer vision. pp. 464–472 (2017)
21. Yang, J., Ou, Y., Chen, Z., Liao, J., Sun, W., Luo, Y., Luo, C.: A benchmark dataset of endoscopic images and novel deep learning method to detect intestinal metaplasia and gastritis atrophy. IEEE Journal of Biomedical and Health Informatics **27**(1), 7–16 (2022)
22. Zhang, J., Zhao, Q., Adeli, E., Pfefferbaum, A., Sullivan, E.V., Paul, R., Valcour, V., Pohl, K.M.: Multi-label, multi-domain learning identifies compounding effects of hiv and cognitive impairment. Medical Image Analysis **75**, 102246 (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, H., Rai, P., Du, L., Buntine, W.: Bayesian multi-label learning with sparse features and labels, and label co-occurrences. In: International Conference on Artificial Intelligence and Statistics. pp. 1943–1951. PMLR (2018)
25. Zhu, F., Li, H., Ouyang, W., Yu, N., Wang, X.: Learning spatial regularization with image-level supervisions for multi-label image classification. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5513–5522 (2017)
26. Zhu, X., Xiong, Y., Dai, J., Yuan, L., Wei, Y.: Deep feature flow for video recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2349–2358 (2017)
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)