

Interventional Multi-Instance Learning with Deconfounded Instance-Level Prediction

Tiancheng Lin^{1,2}, Hongteng Xu^{3,4,5}, Canqian Yang^{1,2} and Yi Xu^{1,2*}

¹ Shanghai Key Lab of Digital Media Processing and Transmission, Shanghai Jiao Tong University

² MoE Key Lab of Artificial Intelligence, AI Institute, Shanghai Jiao Tong University

³ Gaoling School of Artificial Intelligence, Renmin University of China

⁴ Beijing Key Laboratory of Big Data Management and Analysis Methods ⁵ JD Explore Academy
 {ltc19940819, charles.young, xuyi}@sjtu.edu.cn, hongtengxu@ruc.edu.cn

Abstract

When applying multi-instance learning (MIL) to make predictions for bags of instances, the prediction accuracy of an instance often depends on not only the instance itself but also its context in the corresponding bag. From the viewpoint of causal inference, such bag contextual prior works as a confounder and may result in model robustness and interpretability issues. Focusing on this problem, we propose a novel interventional multi-instance learning (IMIL) framework to achieve deconfounded instance-level prediction. Unlike traditional likelihood-based strategies, we design an Expectation-Maximization (EM) algorithm based on causal intervention, providing a robust instance selection in the training phase and suppressing the bias caused by the bag contextual prior. Experiments on pathological image analysis demonstrate that our IMIL method substantially reduces false positives and outperforms state-of-the-art MIL methods.

Introduction

In many real-world scenarios, fine-grained labels of data, *e.g.*, pixel-wise annotations of high-resolution images, are often unavailable due to the limitations of human resources, time, and budgets. To mitigate the requirement for high-quality labels, multi-instance learning (MIL) treats multiple instances as a bag and learns an instance-level predictive model from a set of labeled bags (Dietterich, Lathrop, and Lozano-Pérez 1997). Such a paradigm has been widely used in many applications, *e.g.*, image classification (Wu et al. 2015a), object detection (Wan et al. 2019), semantic segmentation (Xu et al. 2019), *etc.* Among them, whole slide pathological image (WSI) classification is a representative example. Each WSI is a bag with a pathological label, and the patches of the WSI are unlabeled instances in the bag. The MIL framework learns an instance-level classifier to indicate the patches corresponding to the lesions.

Although many MIL methods have been proposed and achieved encouraging performance in extensive applications (Hou et al. 2016; Chen et al. 2019; Chikontwe et al. 2020), they often suffer from an issue called “bag contextual prior”. In particular, the bag contextual prior is a kind of

Copyright © 2022, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

* Corresponding author.

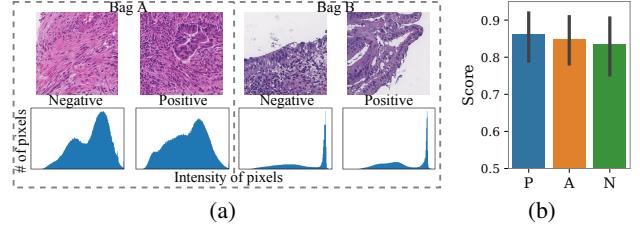


Figure 1: Qualitative and quantitative evidence for bag contextual prior. (a) The patches in two bags and their histograms of pixel intensities. (b) Average scores of positive/all/negative (P/A/N) instances over positive bags and their confidence intervals.

instance-shared information corresponding to bags but irrelevant to their instances, which may be inherited by models like deep neural networks (DNNs) and lead to questionable instance-level prediction. Figure 1(a) illustrates the bag contextual prior in WSI classification. In each bag, its patches (instances) with different labels (positive/negative) often have similar attributes on color and texture. For different bags, on the contrary, their patches with the same labels can be very different in vision. As a result, the “similarity” within each bag and the “difference” across different bags, which harm instance-level prediction, could be wrongly exploited by MIL models. As shown in Figure 1(b), the models predict similar scores for the instances in the same bag. From the viewpoint of causal inference, the above bag contextual prior is a confounder that causes the spurious correlation between instances and labels, making the prediction depend on both the key instance and its useless context. Therefore, a robust and interpretable MIL model should build an efficient mechanism to suppress the bias caused by the bag contextual prior, predicting the classification scores via revealing the actual causality between instances and labels.

To achieve deconfounded instance-level prediction, we propose a novel interventional multi-instance learning (IMIL), in which a structural causal model (SCM) (Pearl, Glymour, and Jewell 2016) analyzes the causalities among the bag contextual prior, instances, and labels. As depicted in Figure 2, our IMIL is an Expectation-Maximization (EM) algorithmic framework with two novel strategies for con-

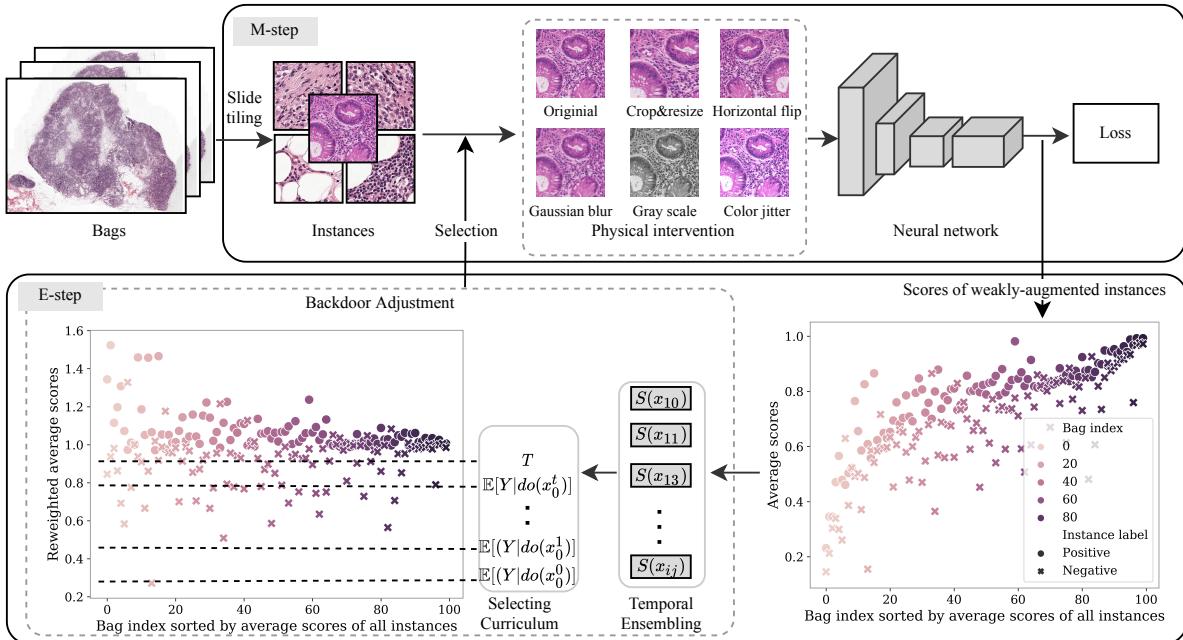


Figure 2: An illustration of the proposed IMIL framework in WSI classification. Physical intervention and backdoor adjustment are adopted for the deconfounded training in the M-step and the instance selection in the E-step, respectively. Here, the instance-level labels are only used for illustration but never applied in the training phase.

founder bias removal and robust instance selection. In the training phase, we initially assign the instances with the label of the bag they belong to, then we alternate the following E-step and M-step until convergence. In the M-step, we apply physical interventions to remove the confounder bias, where various data augmentations are adopted. In the E-step, we approximate the total effect of our model, where we first reweight the scores of instances to get the de-biased prediction and then select instances via both direct causal effect and indirect mediator effect. Note that, different from the existing instance selection criteria (Hou et al. 2016; Chen et al. 2019; Wang et al. 2019a), our model approximates the causal effect without external information.

We compare our IMIL framework with state-of-the-art MIL methods through the lens of causality and analyze their connections and differences in detail. The effectiveness of our IMIL is verified over two public WSI datasets, *i.e.*, DigestPath (Li et al. 2019a) and Camelyon16 (Bejnordi et al. 2017). Experimental results show that our IMIL achieves superior performance in the WSI classification tasks. Significantly, the proposed physical intervention is compatible with all compared MIL methods, bringing consistent performance boosting. Furthermore, we demonstrate the potentials of IMIL to multi-class multi-label MIL problems on the Pascal VOC dataset (Everingham et al. 2015).

Related Work

Bag-level MIL Bag-level MIL either implicitly utilizes bag-to-bag distance/similarity or explicitly trains a bag classifier (Wang et al. 2018), whose novel distance metrics and aggregation operators are designed based on various neu-

ral network architectures (Nazeri, Aminpour, and Ebrahimi 2018; Wang et al. 2019b; Zhao et al. 2020a), pooling strategies (Yan et al. 2018), and attention mechanisms (Ilse, Tomczak, and Welling 2018; Shi et al. 2020). For large-scale MIL scenarios like gigapixel image analysis, we often have to implement the bag-level MIL models by a two-stage strategy (Campanella et al. 2019; Tellez et al. 2019; Li et al. 2019b; Zhao et al. 2020b; Yao et al. 2020; Li, Li, and Eliceiri 2021), training an instance-level feature extractor, and then aggregating instance features as bag-level representations.

Instance-level MIL Instance-level MIL is a natural solution to gigapixel image analysis, where a classifier is trained to produce a *score* for each instance, and the instance scores are aggregated to produce a bag score. The representative method is SimpleMIL, which directly propagates the bag label to its instances (Ray and Craven 2005; Cheplygina et al. 2017). To suppress the noise caused by the instance-level supervision, the work in (Wang et al. 2019a) directly introduces extra cleaner annotations for partial instances by imposing larger weights on them. Alternatively, various modifications have been introduced to SimpleMIL (Hou et al. 2016; Chen et al. 2019; Chikontwe et al. 2020), aiming at using discriminative instances for training. As shown in Figure 2, these approaches are essentially in the EM framework, training model in M-step and selecting instances in E-step.

Causal Inference in Computer Vision Causal inference has been introduced to a growing number of computer vision tasks, including class-incremental learning (Hu et al. 2021), long-tailed classification (Tang, Huang, and Zhang 2020), unsupervised representation learning (Wang et al.

2020; Mitrovic et al. 2020), few-shot learning (Yue et al. 2020), weakly supervised semantic segmentation (Zhang et al. 2020) and so on. It not only offers an interpretation framework to analyze visual problems, but also empowers the models by removing the spurious correlation (Qi et al. 2020), leveraging path-effect analysis (Niu et al. 2021), and generating counterfactual examples (Yue et al. 2021). In MIL problems, StableMIL (Zhang 2019) takes “adding an instance to a bag” as a treatment and observes the bag labels under different treatments, which focuses on bag-level prediction, while our IMIL focuses on instance-level tasks with confounding issues caused by bag contextual prior.

Interventional Multi-Instance Learning Revisit MIL through Causal Inference

Denote $\{X_i, Y_i\}_{i=1}^I$ as a set of coarsely-labeled bags. The bag $X_i = \{x_{ij}\}_{j=1}^{N_i}$ contains N_i unlabeled instances — each instance-level label *i.e.*, $y_{ij} \in \{0, 1\}$ for x_{ij} , is unavailable. For each X_i , its bag-level label Y_i is derived under the standard multiple instance assumption (Dietterich, Lathrop, and Lozano-Pérez 1997), *i.e.*, $Y_i = 1$ when $\exists y_{ij} = 1$ for $j = 1, \dots, N_i$, otherwise, $Y_i = 0$.

Multi-instance learning aims at training a predictive model based on coarsely-labeled bags. As illustrated in Figure 3(a), we can formulate the MIL framework as a causal graph (*a.k.a.*, a structural causal model or SCM) (Pearl, Glymour, and Jewell 2016), denoted as $\mathcal{G} = \{\mathcal{N}, \mathcal{E}\}$. The nodes \mathcal{N} are a set of variables, and the edges \mathcal{E} indicate causal relations in the system, which are shown below:

- $B \rightarrow X$: We denote X as the instance and B as the bag contextual prior. This link reflects the fact that a bag contains multiple instances.
- $B \rightarrow D \leftarrow X$: We denote D as the contextual information shared by the instances in the same bag (*a.k.a.*, instance-shared representation), derived based on the bag contextual prior B . This contextual information is naturally encoded by MIL models as manifold bases (Arora et al. 2019), semantic topics (Bau et al. 2017), typical patterns (Zhang, Wu, and Zhu 2018), *etc.*
- $X \rightarrow Y \leftarrow D$: We denote Y as the classification score determined by X via a direct effect $X \rightarrow Y$ and an indirect effect $D \rightarrow Y$. $X \rightarrow Y$ is obvious, which means the MIL model outputs Y given X . On the other hand, $D \rightarrow Y$ indicates that the bag contextual prior affects the instance labels. Note that $D \rightarrow Y$ always exists in MIL models. Specifically, if $D \not\rightarrow Y$ in Figure 3(a), the only path that transfers knowledge from B to Y : $B \rightarrow X \rightarrow Y$ is blocked by conditioning on X (d-separation (Pearl, Glymour, and Jewell 2016)), then instance labels are no longer related to bags, which conflicts with the setting of MIL.

Again, take WSI classification as an example: 1) $B \rightarrow X$: a WSI contains multiple patches, and the patches are of either different tissue types (*e.g.* mitosis, cellular, neuronal and gland) or diagnoses (*e.g.* cancer/non-cancer) (Chan et al. 2019). 2) $B \rightarrow D \leftarrow X$: The patches in a bag share some underlying information or features, *e.g.*, the

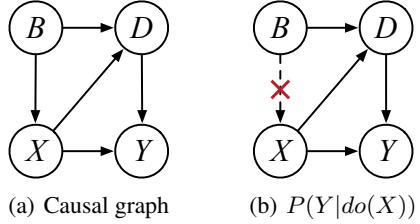


Figure 3: An illustration of the SCM model for MIL.

global low-level features of the bag like colors and textures. Accordingly, D corresponds to such instance-shared information. 3) $X \rightarrow Y \leftarrow D$: A MIL model classifies the patches based on both instance-specific and instance-shared representations. Besides WSI classification, other MIL problems like temporal action localization (Narayan et al. 2019) (videos as bags and frames as instances) and weakly-supervised semantic segmentation (Xu et al. 2019) (images as bags and objects as instances) can also be interpreted by the SCM in Figure 3(a).

In our SCM graph, B confounds X and Y via the backdoor path $X \leftarrow B \rightarrow D \rightarrow Y$ (Pearl 1995), *i.e.*, predicting all instances in a bag to be the same even if some instances are irrelevant to the prediction. On the other hand, $X \rightarrow D \rightarrow Y$ is a mediation path (Pearl 2013), which is the key mechanism of MIL. Accordingly, the instance-shared information D works as a mediator, which encodes the dependencies of instances. Take the objects in an indoor scene as an example. An “indoor” bag tends to contain instances of “TV” rather than “wild animal”, where D contains the “indoor” semantics, which could narrow the search space for instance prediction Y by filtering out those that belong to the “outdoor” scene. As detailed in the following sections, the main difference between our IMIL and the existing methods lies in the operations of confounder and mediator.

Proposed Learning Method

An ideal MIL model should capture the true causality between X and Y . However, from the SCM in Figure 3(a), the conventional correlation of $P(Y|X)$ fails to do so, because the likelihood of Y given X is not only due to X *per se*, but also the spurious correlation caused by the confounder B . Therefore, to pursue the true causality between X and Y , we seek to use the causal intervention $P(Y|do(X))$ instead of the likelihood $P(Y|X)$ for MIL objective. Here, the $do(\cdot)$ operation is defined as forcibly assigning a specific value to a variable, corresponding to applying random controlled trials (Pearl and Mackenzie 2018). Accordingly, we implement our IMIL framework by the following Expectation-Maximization (EM) algorithm, achieving deconfounded training and discriminative instance selection. In the following content, we denote variable as a capital letter and denote value as lowercase.

M-step: Deconfounded training In M-step, the model is optimized under the physical intervention, which aims to “cut-off” the undesired confounding effect, as shown in Fig-

ure 3(b). Instead of enumerating all possible instances in each bag, which is impossible in practice, we adopt strong data augmentations to mimic the random controlled trial. In particular, the bag context prior could be instantiated as structural patterns, geometric arrangements, color distributions, *etc.* Applying data augmentations by spatial and appearance transformations can enhance the diversity of the instances in the same bag. Accordingly, the augmented instances imitate the instances coming from different bags with various contextual priors, which achieves the $do(\cdot)$ operation exactly. In this work, we take the set of data augmentations used in MoCo v2 (Chen et al. 2020) as our default setting,¹ which includes resizing, cropping, horizontally flipping, color jittering, random grayscale conversion. Besides, we also consider random rotation, which has been widely used in image recognition tasks (Wan et al. 2013). As demonstrated in the following experiments, such deconfounded training brings significant improvements to all compared methods, which is a practical, generic, and implementation-friendly solution. It should be noted that leveraging the domain knowledge specialized by tasks may help design more effective data augmentation methods, which is left as our future work.

E-step: Discriminative instance selection via total effect
After applying the deconfounded training above in the M-step, we further select discriminative instances in the E-step, suppressing the confounding bias imposed by those non-discriminative instances in the next iteration. In principle, we introduce the Total Effect (TE) defined below as our criterion for instance selection:

$$TE(Y) = \mathbb{E}[Y|do(X = x)] - \mathbb{E}[Y|do(X = x_0)] \\ = \underbrace{P(1|do(X = x))}_{\text{For binary MIL classification}} - P(1|do(X = x_0)), \quad (1)$$

which measures the expected effect on the prediction Y as the instance X changes from x_0 to x . Here, x_0 is the reference instance, whose effect $\mathbb{E}[Y|do(X = x_0)]$ defines the boundary separating discriminative instances from non-discriminative ones. Accordingly, we select the instances whose effects are larger than that of x_0 . Obviously, the estimation of the TE consists of approximating the causal intervention $P(Y|do(X))$ and setting the reference effect $\mathbb{E}[Y|do(X = x_0)]$. In this work, we apply the backdoor adjustment (Pearl, Glymour, and Jewell 2016) to achieve the causal intervention $P(Y|do(X))$ and propose a progressive curriculum to set the reference effect adaptively.

At the t -th E-step, given a bag $\{x_{ij}\}_{j=1}^{N_i}$, we can calculate the score of each x_{ij} by current model, denoted as $s^{(t)}(x_{ij})$. For MIL classification tasks, the score is often derived by a sigmoid or softmax operation. Based on the scores, we approximate $\mathbb{E}[Y|do(X = x)]$ by an energy-based model (Tang, Huang, and Zhang 2020):

$$\mathbb{E}[Y|do(X = x)] \propto \frac{S(x_{ij})}{\sum_j S(x_{ij})} = E(x_{ij}). \quad (2)$$

¹We call this setting “strong” augmentation in this work. The corresponding “weak” augmentation used in the following step just includes resizing, cropping, and flipping.

This model is different from conventional softmax directly derived by the scores. Firstly, the $S(x_{ij})$ represents the exponential moving average scores derived by the Temporal Ensembling method (Laine and Aila 2016), which estimates the unnormalized effect of $do(X = x_{ij})$. It is calculated as follows:

$$S(x_{ij}) \leftarrow mS(x_{ij}) + (1 - m)s^{(t)}(x_{ij}). \quad (3)$$

This mechanism can be explained as applying momentary interval sampling multiple times, where $S(x_{ij})$ is the ensemble of scores and m is a momentum coefficient. Applying $S(x_{ij})$ rather than $s^{(t)}(x_{ij})$ helps to enhance the robustness of instance selection (Laine and Aila 2016). Secondly, the denominator of $E(x_{ij})$ is the average score of all instances in a bag. It works as a propensity score (Austin 2011), balancing the observational bias of instances, as shown in the scatter plots of Figure 2. Note that our implement is in the form of inverse probability weighting (Pearl, Glymour, and Jewell 2016) since the confounder B is unobserved.

For the reference effect, we initialize $\mathbb{E}[Y|do(X = x_0)]$ as 0 and increase it gradually with the increase of iterations. In particular, we calculate $E(x_{ij})$ for each instance and remain the discriminative instances that correspond to the top $\lfloor R^{(t)} K \rfloor$ largest values in $\{E(x_{ij})\}$, where $R^{(t)} = (1 - \tau t)$, τ is the decay ratio, and $K = \sum_{i=1}^I N_i$ is the number of instances. Accordingly, the reference effect at the t -th E-step, *i.e.*, $\mathbb{E}[Y|do(X = x_0^{(t)})]$, is

$$\max\{\epsilon : \sum_{x_{ij}} \mathbb{I}(E(x_{ij}) \geq \epsilon) = \lfloor R^{(t)} K \rfloor\}, \quad (4)$$

where $\mathbb{I}(\cdot)$ is an indicator function generating 1 if the statement is true. The $R^{(t)}$ stops updating when the average reweighted score of the τK smallest selected instances exceeds a predefined threshold T , *i.e.*,

$$\min_{\mathcal{X} \subset \mathcal{X}_t} \left\{ \frac{1}{\tau K} \sum_{x \in \mathcal{X}} E(x) : |\mathcal{X}| = \tau K \right\} \geq T. \quad (5)$$

where \mathcal{X}_t is the set of selected instances. It is noted that the higher the score is, the more likely newly filtered instances are to be discriminative. Such a procedure can effectively prevent MIL models from over-fitting (Wei et al. 2020). Plugging Eq. (2) and Eq. (4) into Eq. (1), we approximate the TE for each instance x at the t -th E-step as $TE(Y) = E(x) - \mathbb{E}[Y|do(X = x_0^{(t)})]$.

Justifications

Implementations of Causal Intervention While causal intervention is agnostic to methods, datasets, and backbones in theory, their implementations are often task-specific in practice. Focusing on the EM framework of MIL, we implement the causal intervention at the M-step and that at the E-step by the physical intervention and the backdoor adjustment method, respectively. The physical intervention increases the diversity of data, which is beneficial for the M-step to avoid the over-fitting problem. However, it removes the confounding bias by introducing new randomness (Sohn et al. 2020) rather than selecting instances. On the contrary, the backdoor adjustment method approximates

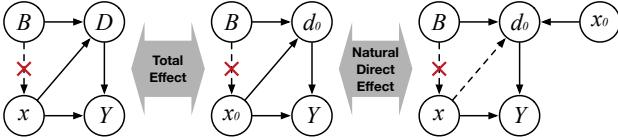


Figure 4: An illustration of the Total Effect and the Natural Direct Effect. Here, d_0 is the instance-shared information corresponding to the reference instance x_0 .

MIL	Simple	Patch	RCE	Top- k	Semi	IMIL
$do(D)$	-	✓	✓	✓	-	-
$do(X)$	-	-	-	-	✓	✓
Effect	-	NDE	NDE	NDE	TE	TE

Table 1: The differences among various MIL methods.

TE to select discriminative instances. As shown in Figure 4, approximating TE helps to remove the undesired confounder bias (*i.e.*, cut off the backdoor path $X \leftarrow B \rightarrow D \rightarrow Y$) and keep the mediator effect (*i.e.*, reserve the mediator path $X \rightarrow D \rightarrow Y$), which is more suitable for the E-step.

Multi-class Multi-label MIL Problems Our IMIL can be easily applied to multi-class multi-label MIL problems, as long as we take each class as a classical (binary) MIL problem. In such situations, the concept of “discriminative” varies from class, and the non-discriminative instance in one class may provide reliable supervision as a negative instance for another class. Therefore, we may select instances in a relatively conservative manner. Specifically, if one instance is not selected for one class, we only reduce the significance of its loss in this class while maintaining its significance for the other classes. This procedure can be understood as a more fine-grained soft version of our selection method.

Connections with Existing MIL Methods It should be noted that our IMIL framework provides a new way to analyze existing state-of-the-art MIL methods. Specifically, Table 1 makes a comparison for various MIL methods from the viewpoint of causal intervention. Essentially, most existing MIL methods can be categorized into three classes according to their instance selection strategies. The method in the first class is the SimpleMIL (Cheplygina et al. 2017), which simply uses all instances without causal intervention; The methods in the second class select instances by calculating the Natural Direct Effect (NDE) shown in Figure 4. In particular, $NDE(Y) = \mathbb{E}[Y_{d_0}|do(X = x)] - \mathbb{E}[Y_{d_0}|do(X = x_0)]$, where Y_{d_0} is the counterfactual output achieved under the condition of $do(X = x_0)$. Because the NDE completely removes the entire effect of D , which may lose some information beneficial for the learning problem, the MIL methods in this class often require some external information as compensations. The representative methods include Top- k MIL (Chikontwe et al. 2020), RCEMIL (Chen et al. 2019) and PatchCNN (Hou et al. 2016). These methods select instances by comparing the scores of instances from the same bag, which is actually an intervention on D that forces the mediator-specific effect to be the same. The x_0

Dataset	Class	Instance Size	Train		Test
			# of bags	# of instances	# of instances
DigestPath	Malignant	512	250	10,133	N/A
	Normal		410	33,110	N/A
Camelyon	Metastases	256	40	64,430	60,545
	Normal		40	617,056	60,545

Table 2: Summary of the WSI datasets. N/A: not available.

is decided by the heuristic setting for Top- k MIL, the statistical information for RCEMIL, and the bag-specific threshold for PatchCNN. The methods in the third class select instances by approximating the Total effect. Among them, the SemiMIL in (Wang et al. 2019a) selects instances in a partially deconfounded manner, where the instances with extra annotations are adjusted by assigning a larger weight, while the rest is still confounded. Our proposed IMIL also uses TE for instance selection. It uses the backdoor adjustment to separate the confounding effect from the mediator effect. This method can be taken as a plug-in component to the existing MIL methods. Additionally, our IMIL is free of external information because TE retains the mediator effect of D , which we believe is informative, and hence external information is unnecessary in our method.

Experiments

Pathological Image Analysis

To demonstrate the effectiveness of our IMIL method, we apply it to the WSI classification problem and compare it with state-of-the-art MIL methods. In particular, we consider the classification of colonoscopy tissues (malignant v.s. normal) and the classification of lymph node sections (metastases v.s. normal). The datasets used in our experiments are DigestPath (Li et al. 2019a)² and Camelyon16 (Bejnordi et al. 2017)³, both of which have bag-level and instance-level labels for each image and its patches, respectively. Specifically, we apply Otsu’s method (Otsu 1979) to remove the background of the images and extract non-overlapped patches from the foreground regions. The statistics of these two datasets are summarized in Table 2.

The competitors of our method include an **Oracle** model with full supervision at the instance-level and the state-of-the-art MIL methods like the **SimpleMIL** in (Cheplygina et al. 2017), the **PatchCNN** in (Hou et al. 2016), the **RCEMIL** trained by RCE loss (Chen et al. 2019), the **SemiMIL** trained with both bag-level and partial instance-level labels, and the **Top- k MIL** trained by Top- k selection (Chikontwe et al. 2020). For PatchCNN, SemiMIL and RCEMIL, some external information should be provided: the specific threshold for each bag, partial instance-level labels, and two statistical values for the re-weighting scheme. For other methods, including ours, only the bag-level labels are used for training.

²<https://digestpath2019.grand-challenge.org/>

³<https://camelyon16.grand-challenge.org/>

Method	DigestPath					Camelyon				
	AUC	ACC	F1	REC	PRE	AUC	ACC	F1	REC	PRE
Using extra labels	Oracle	92.36	88.51	74.80	71.53	80.63	81.35	62.05	39.48	24.75
	+ CI	95.52 +3.16	90.66 +2.15	79.83 +5.03	76.85+5.32	84.99 +4.36	86.23 +4.88	62.35+0.30	39.79+0.31	24.88+0.13
	RCEMIL	87.50	82.75	63.90	65.21	63.83	79.52	70.73	62.95	49.73
PatchCNN	+ CI	91.97+4.47	87.46+4.71	73.63+9.73	73.12+7.91	74.68+10.85	85.48+5.96	76.09+5.36	69.55+6.60	54.60+4.87
	+ CI*	91.65-0.32	85.58-1.88	71.13-2.50	69.14-3.98	74.90+0.22	84.69-0.79	76.98+0.89	71.29+1.74	57.15+2.55
	SemiMIL	91.09	83.46	70.26	81.85	62.86	78.61	73.24	67.61	55.84
No extra labels	+ CI	94.53+3.44	88.28+4.82	77.92+7.66	85.69+3.84	72.05+9.19	85.42+6.81	80.78 +7.54	77.11 +9.50	64.75+8.91
	+ CI*	91.84-2.69	82.13-6.15	69.92-8.00	88.29 +2.60	57.57-14.48	71.06-14.36	67.81-12.97	65.63-11.48	61.47-3.28
	SimpleMIL	91.94	87.68	75.43	78.28	74.02	72.2	67.26	66.63	65.38
Top-kMIL	+ CI	94.40+2.46	90.10+2.42	79.90+4.47	80.83+2.55	80.07+6.05	77.49+5.29	72.99+5.73	71.35+4.72	67.28 +1.90
	+ CI*	92.81-1.59	85.88-4.22	73.98-5.92	85.28+4.45	65.76-14.31	73.24-4.25	69.89-3.10	67.77-3.58	63.30-3.98
	IMIL (Ours)	88.16	81.09	65.81	77.83	57.65	74.19	69.85	67.33	62.13
	+ CI	92.06+3.90	86.84 +5.75	73.74+7.93	78.77+0.94	69.6 +11.95	80.18 +5.99	75.34 +5.39	71.35 +4.02	61.43-0.70
										85.10+11.63

Table 3: Numerical results (%) on dataset of DigestPath and Camelyon16. ‘+CI’ means that causal intervention is applied in the M-step. ‘*’ means the external information is noised.

For a fair comparison, all the methods use ResNet-18 (He et al. 2016) as their backbone models. Adam optimizer is used with an initial learning rate of 0.001, and the batch size is set to 64. We run 50 epochs in total and decay the learning rate with the cosine decay schedule (Loshchilov and Hutter 2016). For our method, the hyperparameters are $m = 0.5$, $\tau = 0.05$ and $T = 0.95$ by default. We evaluate the instance-level performance of each method based on 5-fold cross-validation, and the measurements include Area Under Curve (AUC), accuracy (ACC), F1-score, recall (REC) and precision (PRE). The numerical comparisons for various methods on the two datasets are shown in Table 3. We can find that the causal intervention consistently improves the methods in all settings, indicating that causal intervention is agnostic to methods and datasets.

For the methods using external information, the finer granularity the supervision used, the better performance is achieved. However, their performance is also highly dependent on the quality of external labels, which decreases dramatically given noisy external labels (denoted by *). Note that the results of the oracle model should indicate the upper bound of obtainable performance, while its low recall on Camelyon16 is mainly due to the highly imbalanced ratio of the data. Among the methods without external information, our IMIL obtains superior performance. It consistently improves SimpleMIL on most measurements by non-trivial margins. Especially, IMIL remarkably promotes precision by 15.3% and 16.33% on DigestPath and Camelyon16, respectively, indicating a substantial reduction over false positives. Although the method of Top- k MIL performs well on DigestPath, it receives extremely low recall (15.28% without causal intervention) on Camelyon16 because the setting of k is sensitive to the change of bag size. Additionally, the top- k selection may degenerate into max-max selection criteria, which tends to have relatively low recall and high specificity (Xu et al. 2019). On the contrary, our method can adaptively select discriminative instances instead of setting a fixed number.

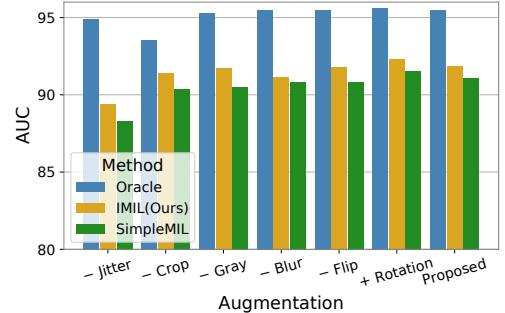


Figure 5: AUC under different composition of data augmentations, where $-/+$ means one individual data augmentation is removed/added.

Further Analysis

Data Augmentations To quantitatively assess the contributions of different data augmentations, we remove/add individual data augmentation in the M-step on three methods (*i.e.*, Oracle, SimpleMIL, and our IMIL). For these methods, removing/adding a significant augmentation method is expected to harm/improve their performance. Figure 5 indicates that “color jittering” is the most important for IMIL and SimpleMIL — the instances in the same bag share both labels and staining conditions, thus jittering could prevent their models from exploiting this spurious correlation. For the oracle model, the “color jittering” is unimportant since the instances are fully-supervised, and the staining condition will cause less confounding bias. Instead, the models may over-fit the co-occurrence of tissues, thus “random resizing and cropping” is necessary for removing these biases. For other augmentations, “grayscale conversion” and “Gaussian blurring” achieve the causal intervention at the cost of hurting the characteristics (the HE staining and the resolution) of WSIs, while “flipping” and “rotation” bring limited improvements because WSIs have no dominant orientation.

Parameters	Settings	AUC	Acc	F1	REC	PRE
Momentum	$m = 0$	92.18	86.69	73.55	79.25	68.80
	$m = 0.25$	91.97	86.72	73.37	78.16	69.25
	$m = 0.5$	92.06	86.84	73.74	78.77	69.60
	$m = 0.75$	92.47	87.32	74.68	79.32	70.70
	$m = 0.9$	91.60	85.56	72.88	79.75	68.07
Threshold	$T = 0.9$	91.67	86.69	72.66	76.34	69.55
	$T = 0.95$	92.06	86.84	73.74	78.77	69.60
	$T = 1$	92.23	87.35	74.34	77.94	71.46
	$T = 1.05$	91.28	86.79	72.23	74.12	70.49
Step	$\tau = 0.025$	91.99	86.90	73.88	78.02	70.46
	$\tau = 0.05$	92.06	86.84	73.74	78.77	69.60
	$\tau = 0.075$	92.46	87.43	74.56	78.20	71.46

Table 4: The robustness of our IMIL to its hyperparameters.

Robustness to Hyperparameters For those important hyperparameters of our method, we change one of them in a range and fix the others to their default values. The performance of our IMIL under different configurations is shown in Table 4. For the momentum used in Temporal Ensembling, the best performance is achieved when $m = 0.75$. This result reflects that a relatively large momentum is beneficial for robust instance selection. For the threshold used in selecting curriculum, we set T around 1 for the assumption that scores of positive and negative instances can be separated by average scores in most case, as shown in Figure 2. Intuitively, higher threshold will keep less instances selected and result in lower recall ($T = 1.05$). For the step used in selecting curriculum, a small τ results in a slow process of instance selection, while a large one may be too aggressive. Overall, our IMIL is robust to the hyperparameters.

Qualitative Results We present qualitative results of IMIL in two aspects: the selecting curriculum procedure and the patch classification on new WSIs. In Figure 6(a), discriminative instances are gradually selected, validating our initial assertions that the average scores of bags naturally can serve as propensity scores. Though the achieved supervision is not perfectly clean, it is worth noting that we do not need any external supervision. In Figure 6(b), the heatmaps indicate the regions of high tumor probability. Notably, our IMIL can accurately distinguish tumors from normal tissues and considerably reduce false positives compared to SimpleMIL, suggesting the effectiveness of the proposed method.

Bag-Level Classification on Pascal VOC

To demonstrate the universality of our IMIL method, besides WSI classification problems, we further test it on the PASCAL VOC 07 dataset (Everingham et al. 2015). This dataset contains 9963 natural images of 20 categories, which is challenging as the appearances of objects are diverse. Following (Wu et al. 2015b), we take each image as a bag and adopt region proposal methods, *e.g.*, Region Proposal Network (RPN) (Ren et al. 2015), to generate instances. Since the instance-level labels are unavailable in VOC, we only deploy the methods free of extra information and equip them with three MIL pooling operators, *i.e.*, max pooling, mean pooling and voting, for bag classification. For all the meth-

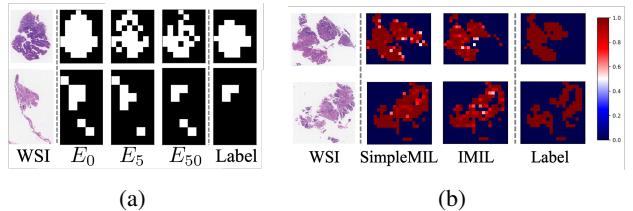


Figure 6: (a) The procedure of selecting curriculum: From left to right, WSI, selected instances at epoch 0/5/50, denoted by E_0 , E_5 and E_{50} , and ground-truth of instance label. (b) The results of instance classification: From left to right, WSI, result of SimpleMIL, result of IMIL, ground-truth of instance label.

MIL	Aggregator		
	max	mean	voting
Simple	63.89	68.2	65.46
	+CI	74.8+10.91	77.97+9.77
Top- k	64.77	67.97	65.13
	+CI	73.48+8.71	77.13+9.16
IMIL	65	68.65	65.66
	+CI	75.78 +10.78	78.34 +9.69

Table 5: The mean average precision over Pascal VOC 07.

ods, we reports its mean average precision (mAP) on the *test set* in Table 5. The proposed causal intervention steadily improves all compared methods, where our IMIL achieves the best performance on bag-level prediction, demonstrating the proposed framework’s stability and effectiveness.

Conclusion

We present a novel interventional multi-instance learning (IMIL) method to suppress the negative influence of bag contextual prior to instance-level prediction. In particular, we propose a causal graph for MIL and equip the EM-based MIL paradigm with causal intervention, combining the training process with data augmentations and adaptive instance selection. Experimental results show that our IMIL achieves promising performance on various computer vision tasks. In the future, we plan to design more task-specific data augmentation methods to improve the physical intervention strategy. Moreover, a systematic comparison will be considered for various causal intervention methods in MIL tasks.

Acknowledgments

Dr. Yi Xu was supported in part by NSFC 62171282, Shanghai Municipal Science and Technology Major Project (2021SHZDZX0102), 111 project BP0719010, and SJTU Science and Technology Innovation Special Fund ZH2018ZDA17. Dr. Hongteng Xu was supported in part by NSFC 62106271 and Intelligent Social Governance Platform, Major Innovation & Planning Interdisciplinary Platform for the “Double-First Class” Initiative, RUC. He also thanks the support of Public Policy and Decision-making Research Lab of RUC and JD Explore Academy. Authors thank Zengchao Xu for inspiring discussions.

References

- Arora, S.; Cohen, N.; Hu, W.; and Luo, Y. 2019. Implicit regularization in deep matrix factorization. *Advances in Neural Information Processing Systems*, 32: 7413–7424.
- Austin, P. C. 2011. An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate behavioral research*, 46(3): 399–424.
- Bau, D.; Zhou, B.; Khosla, A.; Oliva, A.; and Torralba, A. 2017. Network dissection: Quantifying interpretability of deep visual representations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 6541–6549.
- Bejnordi, B. E.; Veta, M.; Van Diest, P. J.; Van Ginneken, B.; Karssemeijer, N.; Litjens, G.; Van Der Laak, J. A.; Hermsen, M.; Manson, Q. F.; Balkenhol, M.; et al. 2017. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *Jama*, 318(22): 2199–2210.
- Campanella, G.; Hanna, M. G.; Geneslaw, L.; Miraflor, A.; Silva, V. W. K.; Busam, K. J.; Brogi, E.; Reuter, V. E.; Klimstra, D. S.; and Fuchs, T. J. 2019. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nature medicine*, 25(8): 1301–1309.
- Chan, L.; Hosseini, M. S.; Rowsell, C.; Plataniotis, K. N.; and Damaskinos, S. 2019. Histosegnet: Semantic segmentation of histological tissue type in whole slide images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 10662–10671.
- Chen, H.; Han, X.; Fan, X.; Lou, X.; Liu, H.; Huang, J.; and Yao, J. 2019. Rectified Cross-Entropy and Upper Transition Loss for Weakly Supervised Whole Slide Image Classifier. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 351–359. Springer.
- Chen, X.; Fan, H.; Girshick, R.; and He, K. 2020. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*.
- Cheplygina, V.; Sørensen, L.; Tax, D. M. J.; de Bruijne, M.; and Loog, M. 2017. Label Stability in Multiple Instance Learning. *arXiv:1703.04986*.
- Chikontwe, P.; Kim, M.; Nam, S. J.; Go, H.; and Park, S. H. 2020. Multiple Instance Learning with Center Embeddings for Histopathology Classification. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 519–528. Springer.
- Dietterich, T. G.; Lathrop, R. H.; and Lozano-Pérez, T. 1997. Solving the multiple instance problem with axis-parallel rectangles. *Artificial intelligence*, 89(1-2): 31–71.
- Everingham, M.; Eslami, S. A.; Van Gool, L.; Williams, C. K.; Winn, J.; and Zisserman, A. 2015. The pascal visual object classes challenge: A retrospective. *International journal of computer vision*, 111(1): 98–136.
- He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.
- Hou, L.; Samaras, D.; Kurc, T. M.; Gao, Y.; Davis, J. E.; and Saltz, J. H. 2016. Patch-based convolutional neural network for whole slide tissue image classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2424–2433.
- Hu, X.; Tang, K.; Miao, C.; Hua, X.-S.; and Zhang, H. 2021. Distilling Causal Effect of Data in Class-Incremental Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3957–3966.
- Ilse, M.; Tomczak, J. M.; and Welling, M. 2018. Attention-based deep multiple instance learning. *arXiv preprint arXiv:1802.04712*.
- Laine, S.; and Aila, T. 2016. Temporal ensembling for semi-supervised learning. *arXiv preprint arXiv:1610.02242*.
- Li, B.; Li, Y.; and Eliceiri, K. W. 2021. Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14318–14328.
- Li, J.; Yang, S.; Huang, X.; Da, Q.; Yang, X.; Hu, Z.; Duan, Q.; Wang, C.; and Li, H. 2019a. Signet ring cell detection with a semi-supervised learning framework. In *International Conference on Information Processing in Medical Imaging*, 842–854. Springer.
- Li, S.; Liu, Y.; Sui, X.; Chen, C.; Tjio, G.; Ting, D. S. W.; and Goh, R. S. M. 2019b. Multi-instance multi-scale CNN for medical image classification. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 531–539. Springer.
- Loshchilov, I.; and Hutter, F. 2016. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*.
- Mitrovic, J.; McWilliams, B.; Walker, J.; Buesing, L.; and Blundell, C. 2020. Representation learning via invariant causal mechanisms. *arXiv preprint arXiv:2010.07922*.
- Narayan, S.; Cholakkal, H.; Khan, F. S.; and Shao, L. 2019. 3c-net: Category count and center loss for weakly-supervised action localization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 8679–8687.
- Nazeri, K.; Aminpour, A.; and Ebrahimi, M. 2018. Two-stage convolutional neural network for breast cancer histology image classification. In *International Conference Image Analysis and Recognition*, 717–726. Springer.
- Niu, Y.; Tang, K.; Zhang, H.; Lu, Z.; Hua, X.-S.; and Wen, J.-R. 2021. Counterfactual vqa: A cause-effect look at language bias. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12700–12710.
- Otsu, N. 1979. A threshold selection method from gray-level histograms. *IEEE transactions on systems, man, and cybernetics*, 9(1): 62–66.
- Pearl, J. 1995. Causal diagrams for empirical research. *Biometrika*, 82(4): 669–688.
- Pearl, J. 2013. Direct and indirect effects. *arXiv preprint arXiv:1301.2300*.

- Pearl, J.; Glymour, M.; and Jewell, N. P. 2016. *Causal inference in statistics: A primer*. John Wiley & Sons.
- Pearl, J.; and Mackenzie, D. 2018. *The book of why: the new science of cause and effect*. Basic books.
- Qi, J.; Niu, Y.; Huang, J.; and Zhang, H. 2020. Two causal principles for improving visual dialog. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10860–10869.
- Ray, S.; and Craven, M. 2005. Supervised versus multiple instance learning: An empirical comparison. In *Proceedings of the 22nd international conference on Machine learning*, 697–704.
- Ren, S.; He, K.; Girshick, R.; and Sun, J. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28: 91–99.
- Shi, X.; Xing, F.; Xie, Y.; Zhang, Z.; Cui, L.; and Yang, L. 2020. Loss-based attention for deep multiple instance learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 5742–5749.
- Sohn, K.; Berthelot, D.; Li, C.-L.; Zhang, Z.; Carlini, N.; Cubuk, E. D.; Kurakin, A.; Zhang, H.; and Raffel, C. 2020. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *arXiv preprint arXiv:2001.07685*.
- Tang, K.; Huang, J.; and Zhang, H. 2020. Long-tailed classification by keeping the good and removing the bad momentum causal effect. *arXiv preprint arXiv:2009.12991*.
- Tellez, D.; Litjens, G.; van der Laak, J.; and Ciompi, F. 2019. Neural image compression for gigapixel histopathology image analysis. *IEEE transactions on pattern analysis and machine intelligence*.
- Wan, F.; Liu, C.; Ke, W.; Ji, X.; Jiao, J.; and Ye, Q. 2019. C-mil: Continuation multiple instance learning for weakly supervised object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2199–2208.
- Wan, L.; Zeiler, M.; Zhang, S.; Le Cun, Y.; and Fergus, R. 2013. Regularization of neural networks using drop-connect. In *International conference on machine learning*, 1058–1066. PMLR.
- Wang, T.; Huang, J.; Zhang, H.; and Sun, Q. 2020. Visual commonsense r-cnn. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10760–10770.
- Wang, X.; Chen, H.; Gan, C.; Lin, H.; Dou, Q.; Tsougenis, E.; Huang, Q.; Cai, M.; and Heng, P.-A. 2019a. Weakly supervised deep learning for whole slide lung cancer image analysis. *IEEE transactions on cybernetics*, 50(9): 3950–3962.
- Wang, X.; Yan, Y.; Tang, P.; Bai, X.; and Liu, W. 2018. Revisiting multiple instance neural networks. *Pattern Recognition*, 74: 15–24.
- Wang, X.; Yan, Y.; Tang, P.; Liu, W.; and Guo, X. 2019b. Bag similarity network for deep multi-instance learning. *Information Sciences*, 504: 578–588.
- Wei, H.; Feng, L.; Chen, X.; and An, B. 2020. Combating noisy labels by agreement: A joint training method with co-regularization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13726–13735.
- Wu, J.; Yu, Y.; Huang, C.; and Yu, K. 2015a. Deep multiple instance learning for image classification and auto-annotation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3460–3469.
- Wu, J.; Yu, Y.; Huang, C.; and Yu, K. 2015b. Deep multiple instance learning for image classification and auto-annotation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3460–3469.
- Xu, G.; Song, Z.; Sun, Z.; Ku, C.; Yang, Z.; Liu, C.; Wang, S.; Ma, J.; and Xu, W. 2019. Camel: A weakly supervised learning framework for histopathology image segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 10682–10691.
- Yan, Y.; Wang, X.; Guo, X.; Fang, J.; Liu, W.; and Huang, J. 2018. Deep multi-instance learning with dynamic pooling. In *Asian Conference on Machine Learning*, 662–677. PMLR.
- Yao, J.; Zhu, X.; Jonnagaddala, J.; Hawkins, N.; and Huang, J. 2020. Whole slide images based cancer survival prediction using attention guided deep multiple instance learning networks. *Medical Image Analysis*, 65: 101789.
- Yue, Z.; Wang, T.; Sun, Q.; Hua, X.-S.; and Zhang, H. 2021. Counterfactual zero-shot and open-set visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15404–15414.
- Yue, Z.; Zhang, H.; Sun, Q.; and Hua, X.-S. 2020. Interventional few-shot learning. *arXiv preprint arXiv:2009.13000*.
- Zhang, D.; Zhang, H.; Tang, J.; Hua, X.; and Sun, Q. 2020. Causal intervention for weakly-supervised semantic segmentation. *arXiv preprint arXiv:2009.12547*.
- Zhang, Q.; Wu, Y. N.; and Zhu, S.-C. 2018. Interpretable convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 8827–8836.
- Zhang, W. 2019. Stable multi-instance learning via causal inference. *arXiv: Learning*.
- Zhao, Y.; Yang, F.; Fang, Y.; Liu, H.; Zhou, N.; Zhang, J.; Sun, J.; Yang, S.; Menze, B.; Fan, X.; et al. 2020a. Predicting lymph node metastasis using histopathological images based on multiple instance learning with deep graph convolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4837–4846.
- Zhao, Y.; Yang, F.; Fang, Y.; Liu, H.; Zhou, N.; Zhang, J.; Sun, J.; Yang, S.; Menze, B.; Fan, X.; et al. 2020b. Predicting lymph node metastasis using histopathological images based on multiple instance learning with deep graph convolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4837–4846.