Multi-Instance Causal Representation Learning for Instance Label Prediction and Out-of-Distribution Generalization

Weijia Zhang[1*], **Xuanhui Zhang**[2], **Han-Wen Deng**[1], **Min-Ling Zhang**[1]
[1]School of Computer Science and Engineering, Southeast University, Nanjing 210096, China
[2] School of Information Management, Nanjing University, Nanjing 210023, China
zhangwj@seu.edu.cn, zhangxhdo@163.com, {denghw,zhangml}@seu.edu.cn

Abstract

Multi-instance learning (MIL) deals with objects represented as bags of instances and can predict instance labels from bag-level supervision. However, significant performance gaps exist between instance-level MIL algorithms and supervised learners since the instance labels are unavailable in MIL. Most existing MIL algorithms tackle the problem by treating multi-instance bags as harmful ambiguities and predicting instance labels by reducing the supervision inexactness. This work studies MIL from a new perspective by considering bags as auxiliary information, and utilize it to identify instance-level causal representations from bag-level weak supervision. We propose the CausalMIL algorithm, which not only excels at instance label prediction but also provides robustness to distribution change by synergistically integrating MIL with identifiable variational autoencoder. Our approach is based on a practical and general assumption: the prior distribution over the instance latent representations belongs to the non-factorized exponential family conditioning on the multi-instance bags. Experiments on synthetic and real-world datasets demonstrate that our approach significantly outperforms various baselines on instance label prediction and out-of-distribution generalization tasks.

1 Introduction

Supervised learning has achieved great success in many applications. However, an important limitation of existing supervised learning algorithms is that they model complex objects as a single feature vector and thus cannot make fine-grained predictions without the corresponding level of supervision, e.g., localizing the region of interest within an image from coarse-grained image-level supervision. Unfortunately, acquiring fine-grained labels is often not only a tedious task but also prohibitively expensive, especially for applications that require a high level of domain-specific expertise such as drug activity prediction [8] and medical image classification [24].

Multi-Instance Learning (MIL) [8] is a weakly supervised learning paradigm originally proposed for drug activity prediction, where the task is to predict whether a molecule is suitable for binding to a target receptor. Since a molecule can take many low-energy conformations and its suitability for making drugs depends on some specific but unknown conformations, objects in MIL are represented by groups of instances called *bags* where each instance is described by its own feature vector, instead of represented by a single feature vector as in standard supervised learning. Because only molecule-level drug binding suitability is known to human experts, MIL algorithms are only coarsely supervised at the bag-level, while fine-grained instance labels are unknown.

*Corresponding author

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

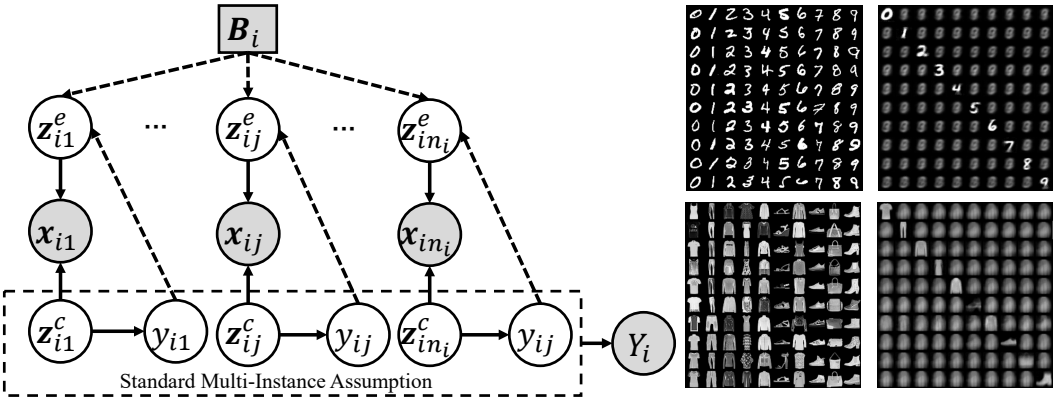

Figure 1: (Left) Our proposed graphical model. Shaded nodes denote observed variables and white nodes represent the latent ones. Dashed lines denote edges that may exist depending on the bags, while solid lines indicate relationships that are invariant with respect to the bags. (Right) Multi-instance bags of MNIST and FashionMNIST images and their reconstructions from $\boldsymbol{z}^c$ (each class per row) inferred by CausalMIL.

Utilizing the bag-level supervision, the prediction tasks of MIL are two-fold: predicting the bag labels, e.g., whether a molecule is suitable for making drugs or a human organ contains cancerous cells; and predicting the instance labels, e.g., which specific molecular conformations are suitable or which particular cells are cancerous. Although theoretical results have demonstrated the feasibility of learning accurate instance concepts from bag labels [9], the empirical performances of weakly supervised instance label predictions are far behind their fully supervised counterparts.

The inexactness in MIL supervision causes the main difficulty of instance label prediction [42]. Under the standard multi-instance assumption [11], the bag labels only inform the learner that instances from negative bags are negative, while the exact labels for instances in the positive bags are unknown. Most previous instance-level MIL algorithms solve the above problem by disambiguating the bag labels, i.e., finding the most positive instance from positive bags [28, 35] or utilizing the attention mechanism to infer how the different instances contribute to the bag label [15, 31, 29].

In this work, we propose to utilize the information hidden in the multi-instance bags for identifying the latent factors that generate the observed instances, and utilize the bag labels for encouraging disentanglement among the latent factors. Specifically, we consider each observed instance as generated from instance-specific causal factors $\boldsymbol{z}^c_{ij}$ and bag-inherited non-causal factors $\boldsymbol{z}^e_{ij}$ as depicted in Figure 1 (left). The instance-specific factors $\boldsymbol{z}^c_{ij}$ are responsible for capturing information that is causal to the instance labels, while the bag-inherited factors $\boldsymbol{z}^e_{ij}$ capture spurious information that may be inherited from the bags and correlates with the instance labels. The bag labels are considered as the effects of their instance labels according to the standard multi-instance assumption.

By identifying and disentangling the latent factors, our model brings two advantages over existing MIL algorithms. Firstly, as the identified latent factors capture semantic meaningful causal representations are of much lower dimensions than the observed instances, their identification significantly reduces the difficulties of inferring instance labels from bag supervision. Secondly, since only the instance-specific factors $\boldsymbol{z}^c_{ij}$ are causes to the instance labels y_{ij}, disentangling $\boldsymbol{z}^c_{ij}$ from $\boldsymbol{z}^e_{ij}$ remove the spurious correlations between $\boldsymbol{B}_i$ and y_{ij}. Therefore, identification and disentanglement of $\boldsymbol{z}^c_{ij}$ is not only beneficial for instance label predictions since $\boldsymbol{z}^e_{ij}$ will vary among bags, but is also useful for out-of-distribution (OOD) generalization since the factors that are the effects of instances labels are excluded from the predictive model [4, 1, 23]. It is worth noting that although $\boldsymbol{B}_i$ and y_{ij} are unconditionally independent, they become conditionally dependent because previous algorithms that do not identify or disentangle the latent factors implicitly condition on either $\boldsymbol{x}_{ij}$ or $\boldsymbol{z}^e_{ij}$ (i.e., the collider or descendants of the collider in a v-structure [26]).

Our identifiability result builds upon the recent identifiable variational autoencoder (iVAE) [17], but extends to a more flexible non-factorized prior distribution conditioning on the bag information instead of the factorized prior conditioning on a single instance assumed in the original iVAE. Furthermore, we propose to utilize a permutation invariant set transformation network for aggregating information

from instances within the same bag to facilitate identifiability. Because the set transformation network universally approximates any set function [39], instances within the same bag are not only treated as sources of supervision inexactness but also effectively contribute to identifiability via conditioning the prior distribution. We summarize our key contribution as follows:

- We propose a novel CausalMIL algorithm that utilizes the bag information with permutation invariant set transformation networks to theoretical guarantee the identifiability of latent factors under non-factorized conditional exponential family priors;
- We propose a general graphical model covering many real-world MIL applications, and show that the causal and non-causal factors of the instance labels can be disentangled from the identified latents based on bag-level supervision.
- Using a variety of datasets, we show that CausalMIL significantly outperforms existing MIL algorithms on instance label prediction tasks, and achieves better out-of-distribution generalization capabilities when compared to supervised invariant algorithms.

2 Preliminaries

2.1 Notations

Let $\mathcal{X} = \mathbb{R}^m$ denote the instance space and $\mathcal{Y} = \{0, 1\}$ denote the label space. The training data contains samples organized in n training bags $\mathcal{B} = \{\mathbf{B}_1, \cdots, \mathbf{B}_i, \cdots, \mathbf{B}_n\}$, where each bag is a set that contains different numbers of n_i instances, i.e., $\mathbf{B}_i = \{\boldsymbol{x}_{i1}, \cdots, \boldsymbol{x}_{ij}, \cdots, \boldsymbol{x}_{in_i}\}$ with $\boldsymbol{x}_{ij} \in \mathcal{X}$. In the rest of this paper, we will also use $\mathbf{B}_i$ to denote the auxiliary information contained in the bag by abuse of notation. The learning algorithm is provided with bags $\mathbf{B}_i$ and their associated bag labels $Y_i \in \mathcal{Y}$ during training. Although there exists instance labels $y_{ij} \in \mathcal{Y}$ for each $\boldsymbol{x}_{ij}$, the instance labels are *unknown* to the learners. We follow the standard multi-instance assumption [11] which assumes that positive bags contain at least one positive instance, and negative bags contain only negative instances. It is worth noting that the assumption only requires bags to be i.i.d., the instances within the same bag can be dependent.

2.2 VAE and Identifiability

We briefly introduce Variational Autoencoder (VAE) and its identifiability results. VAE can be considered as the combination of a generative latent variable model and an associated inference model, where both models are parameterized by neural networks [18]. Specifically, VAE learns the joint distribution $p_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{z}) = p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z})p_{\boldsymbol{\theta}}(\boldsymbol{z})$ where $p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z})$ is the conditional distribution of observing $\boldsymbol{x}$ given $\boldsymbol{z}$, $\boldsymbol{\theta}$ is the set of generative parameters, and $p_{\boldsymbol{\theta}}(\boldsymbol{z}) = \Pi_{i=1}^{d} p_{\boldsymbol{\theta}}(z_i)$ is the factorized prior distribution of the latents. By introducing an inference model $q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})$, the set of parameters $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$ can be jointly optimized through maximizing the evidence lower bound (ELBO) on the marginal likelihood $p_{\boldsymbol{\theta}}(\boldsymbol{x})$:

$$\begin{aligned}\mathcal{L} &= \mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})}[\log p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z})] - D_{\mathrm{KL}}(q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})||p(\boldsymbol{z})) \\ &= \log p_{\boldsymbol{\theta}}(\boldsymbol{x}) - D_{\mathrm{KL}}(q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})||p_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x})) \leq \log p_{\boldsymbol{\theta}}(\boldsymbol{x}),\end{aligned} \tag{1}$$

where D_{KL} denotes the KL-divergence between the approximation and the true posterior, and $\mathcal{L}$ is a lower bound of the marginal likelihood $p_{\boldsymbol{\theta}}(\boldsymbol{x})$ because of the non-negativity of the KL-divergence.

Recently, it has been shown that VAEs with unconditional prior distributions $p_{\boldsymbol{\theta}}(\boldsymbol{z})$ are not identifiable [22], but the latent factors $\boldsymbol{z}$ can be identified with a conditionally factorized prior distribution $p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{u})$ over the latent variables to break the symmetry, where $\boldsymbol{u}$ is an additionally observed variable [17].

3 Methods

3.1 CausalMIL

We now formally describe our graphical model depicted in Figure 1 (left). For any given bag $\boldsymbol{B}_i$, we model its observed instances $\boldsymbol{x}_{ij}$ as generated from unknown latent factors $\boldsymbol{z}_{ij} = (\boldsymbol{z}_{ij}^c, \boldsymbol{z}_{ij}^e) \in \mathbb{R}^d$ with $d \ll m$. We let $\boldsymbol{z}_{ij}^c$ correspond to the causes of the instance label y_{ij}, i.e., the line stroke that

makes a digit represent the Arabic number '1' in Figure 1 (right). In contrast $\boldsymbol{z}_{ij}^{e}$ captures other factors related to the observed $\boldsymbol{x}$ but are not causes of the instance label, i.e., the angle, thickness, and color of the line stroke that are not causal to '1'. The solid lines of the graphical model indicate that the causal relationships are invariant across all bags, and the dashed lines denote that their relationships may vary or even be absent across different bags. The bag label Y_i is determined by the labels of its instances y_{ij} according to the standard multi-instance assumption. To sum up, we posit that the graphical structure in Figure 1 should satisfy the following assumptions:

Assumption 1. *(a) The bag labels are determined by its instances according to the standard multi-instance assumption; (b) the causal graph in Figure 1 is a Directed Acyclic Graph (DAG); (c) given the latent factors $\boldsymbol{z}_{ij}$, the observed instances are independent of the labels and the bags: $\boldsymbol{x}_{ij} \perp\!\!\!\perp y_{ij}, Y_i, \boldsymbol{B}_i | \boldsymbol{z}_{ij}$, i.e., the generation mechanism $p(\boldsymbol{x}_{ij}|\boldsymbol{z}_{ij})$ is invariant across bags; (d) $y_{ij} \perp\!\!\!\perp \boldsymbol{B}_i | \boldsymbol{z}_{ij}^{c}$, i.e., the mechanism between the causal factors $\boldsymbol{z}_{ij}^{c}$ and y_{ij} is invariant across bags.*

We now discuss the practicality of Assumption 1. Firstly, the standard multi-instance assumption commonly used in a wide range of MIL applications such as medical image analysis [20], fine-grained sentiment analysis [3], and sound event detection [36]. Secondly, modeling the causal structures as DAGs are common in causal discovery [26, 27]. To the best of our knowledge, most of the causality-based machine learning literature assumes the causal structure to be acyclic. Thirdly, it also makes sense that generating mechanism from latent factors to observed instances $p(\boldsymbol{x}_{ij}|\boldsymbol{z}_{ij})$ is invariant since causal mechanisms are considered to be stable across environments [30]. Otherwise, it would be impossible to infer $\boldsymbol{z}$ from $\boldsymbol{x}$ for any test instances. Lastly, the fourth assumption is not only commonly adopted in invariant and causal representation learning literature [4, 1, 23], but also suitable for MIL applications. For example, when diagnosing whether a cell from a tissue is cancerous, the causal mechanism of diagnosing a cell should be invariant across bags, while the non-causal factors caused by the patient or equipment differences may often change [40].

From now on we will drop the subscript to avoid cluttering the notations. In the following discussion we will show that when the underlying data generating mechanism satisfies Assumption 1, the latent variables $\boldsymbol{z}$ can be identified up to permutation and affine transformations if the conditional prior distribution $p(\boldsymbol{z}|\boldsymbol{B}, y)$ belongs to a general exponential family distribution:

Assumption 2. *The prior distribution of the latent factors $p(\boldsymbol{z}|\boldsymbol{B}, y)$ follows a general exponential family with its parameter specified by an arbitrary function $\boldsymbol{\lambda}(\boldsymbol{B}, y)$ and sufficient statistics $\boldsymbol{T}(\boldsymbol{z}) = [\boldsymbol{T}_f(\boldsymbol{z})^T, \boldsymbol{T}_{NN}(\boldsymbol{z})^T]^T$. Here the sufficient statistics $\boldsymbol{T}(\boldsymbol{z})$ is defined by the concatenation of $\boldsymbol{T}_f(\boldsymbol{z}) = [\boldsymbol{T}_1(z_1)^T, \cdots, \boldsymbol{T}_d(z_d)^T]^T$ from a factorized exponential family and the outputs of a neural networks $\boldsymbol{T}_{NN}(\boldsymbol{z})$ with universal approximation power. Then, the probability density can be written as:*

$$p_{\mathbf{T},\boldsymbol{\lambda}}(\boldsymbol{z}|\boldsymbol{B}, y) = \frac{\mathcal{Q}(\boldsymbol{z})}{\mathcal{C}(\boldsymbol{B}, y)} \exp[\boldsymbol{T}(\boldsymbol{z})^T \boldsymbol{\lambda}(\boldsymbol{B}, y)], \tag{2}$$

where $\mathcal{Q}(\boldsymbol{z})$ is the base measure and $\mathcal{C}(\boldsymbol{B}, y)$ is the normalizing constant.

The above assumption is inspired by the recent advancement of iVAE [17]. However, there are two differences. Firstly, because Figure 1 allows for connections among the components $\boldsymbol{z}_{ij}$ within $\boldsymbol{z}$ as long as the generative model remains a valid DAG, our non-factorized conditional prior assumption is more general than the factorized distribution assumed by iVAE. In Equation 2, we utilize a neural networks $\boldsymbol{T}_{NN}(\boldsymbol{z}^c)$ for capturing the interactions among components of $\boldsymbol{z}$. Secondly, our additionally observed auxiliary information is provided by multi-instance bag $\boldsymbol{B}$ which may consist arbitrary number of instances instead of a single scalar or vector-valued instance $\boldsymbol{u}$ as used by iVAE. Therefore, it is natural to ask (1) if identifiability still holds for non-factorized priors, and (2) how to aggregate information from bags of varying sizes for breaking the symmetry of the prior distribution?

To answer the above questions, we first describe our generative model and discuss its identifiability. Formally, let us consider the following generative models under Assumption 2:

$$p_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{z}|\boldsymbol{B}, y) = p_{\boldsymbol{f}}(\boldsymbol{x}|\boldsymbol{z}) p_{\mathbf{T},\boldsymbol{\lambda}}(\boldsymbol{z}|\boldsymbol{B}, y), \tag{3}$$

$$p_{\boldsymbol{f}}(\boldsymbol{x}|\boldsymbol{z}) = p_{\boldsymbol{\varepsilon}}(\boldsymbol{x} - f(\boldsymbol{z})), \tag{4}$$

where Equation 3 describes the generative process of $\boldsymbol{x}$ given the underlying latent factors $\boldsymbol{z}$, along with the bag context $\boldsymbol{B}$ and the instance label y. Equation 4 implies that the observed representation $\boldsymbol{x}$ is an additive noise function, i.e., $\boldsymbol{x} = f(\boldsymbol{z}) + \boldsymbol{\varepsilon}$ where $\boldsymbol{\varepsilon}$ is independent of $\boldsymbol{f}$ or $\boldsymbol{z}$.

Following standard VAE and iVAE derivation, the evidence lower bound for each instance $\boldsymbol{x}$ in bag $\boldsymbol{B}$ of the above generative model can be written as:

$$\mathcal{L}_{\text{ELBO}} = \mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{B},y)}[\log p_{\boldsymbol{f}}(\boldsymbol{x}|\boldsymbol{z}) + \log p_{\boldsymbol{T},\boldsymbol{\lambda}}(\boldsymbol{z}|\boldsymbol{B},y) - \log q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{B},y)] + \log p(\boldsymbol{B}). \quad (5)$$

Note that the ELBO in Equation 5 contains an additional term of the bag prior $\log p(\boldsymbol{B})$ that do not affect identifiability but improves the estimation for the conditional prior [25]. Our main identifiability results can now be described as:

Theorem 1. *Assume we observe instances from multi-instance bags sampled according to Equation 2-4. Furthermore, assume that the following holds:*

i. The set $\{\boldsymbol{x} \in \mathcal{X} : \varphi_{\boldsymbol{\varepsilon}}(\boldsymbol{x} = 0\}$ has measure zero, where $\varphi_{\boldsymbol{\varepsilon}}$ is the characteristic function of the density $p_{\boldsymbol{\varepsilon}}$ defined in Equation 4.

ii. The function $\boldsymbol{f}$ is injective and all of its second-order cross partial derivatives exist.

iii. The sufficient statistics $T_{\boldsymbol{f}}$ are twice differentiable.

iv. There exist $k+1$ distinct bags $(\boldsymbol{B}^0, y^0), \ldots, (\boldsymbol{B}^d, y^d)$ such that the $k \times k$ matrix E is invertible, where k is the dimension of $\boldsymbol{T}$ and E is defined as:

$$E = (\boldsymbol{\lambda}(\boldsymbol{B}^1, y^1) - \boldsymbol{\lambda}(\boldsymbol{B}^0, y^0), \ldots, \boldsymbol{\lambda}(\boldsymbol{B}^k, y^k) - \boldsymbol{\lambda}(\boldsymbol{B}^0, y^0)) \quad (6)$$

Then, the parameters $\boldsymbol{\theta} = (\boldsymbol{f}, \boldsymbol{T}, \boldsymbol{\lambda})$ are identifiable up to an equivalence class induced by permutation and componentwise transformations.

It is interesting to observe that condition (iv) of Theorem 1 is easier to satisfy in MIL than standard supervised learning. Intuitively, the condition requires that the auxiliary information should "break the symmetry" in the representation space the model could learn. An intuitive analogy is inferring an object's shape from its shadow: if we only observe one shadow of the object, it's difficult to know its shape; however, if we observe multiple objects under similar lighting conditions (from a bag of instances), we may identify the lighting. If we observe an object under different lighting conditions (from many bags of instances), we may identify the underlying shape. Another analogy is to consider the bags as H&E stained histopathology images and the instances as image patches: in one bag, we observe cells under the same staining process, which provides information regarding the staining procedure; in different bags, we observe cancerous cells under different staining procedures, which makes it possible to infer the causal representations of cancerous cells. In standard supervised learning, instances are independent with no auxiliary information given. In MIL, instances within bags are naturally dependent and organized in bags, which may then provide the necessary auxiliary information for latent identifiability. We provide the proof for Theorem 1 in the Appendices.

Technically, condition (iv) can be restated as requiring the vectors $(\boldsymbol{\lambda}(\boldsymbol{B}^1, y^1) - \boldsymbol{\lambda}(\boldsymbol{B}^0, y^0), \ldots, \boldsymbol{\lambda}(\boldsymbol{B}^k, y^k) - \boldsymbol{\lambda}(\boldsymbol{B}^0, y^0))$ to be independent. Therefore, the instance label y becomes unnecessary if there exist $k+1$ distinct bags such that $E = (\boldsymbol{\lambda}(\boldsymbol{B}^1) - \boldsymbol{\lambda}(\boldsymbol{B}^0), \ldots, \boldsymbol{\lambda}(\boldsymbol{B}^k) - \boldsymbol{\lambda}(\boldsymbol{B}^0))$ of size $k \times k$ is invertible, which is especially attractive in MIL where the instance labels are not available but the bags are abundant.

Two practical issues need to be addressed before applying the identifiability theorems to solve MIL problems. The first hurdle lies in how to characterize the auxiliary information for conditioning the prior distribution of $\boldsymbol{z}$. This is different from standard iVAE where the auxiliary information is readily provided in a single-instance $\boldsymbol{u}$ since the auxiliary information in MIL are provided in the form of instances sampled from bags. To make things worse, the auxiliary information also cannot be derived from the bag labels or the multi-instance assumption. This is because they only tell us the bag labels and the relationship between instances labels and bag labels, instead of the generative process of the bags and their instances.

To solve the above problem, we propose to utilize a trainable permutation invariant function parameterized by deep neural networks. Formally, we define:

$$\boldsymbol{B}_i = net(\{\boldsymbol{x}_{i1}, \cdots, \boldsymbol{x}_{in_i}\}) = \rho[pool(\{\phi(\boldsymbol{x}_{i1}), \cdots, \phi(\boldsymbol{x}_{in_i})\})], \quad (7)$$

where ρ and ϕ are arbitrary continuous functions implemented through neural networks, $pool$ is the sum operator over the instances $\boldsymbol{x}$ transformed by the neural network ϕ, and finally ρ is another neural

network that operates on the pooled transformations. An important property of the transformation defined in Equation 7 is that it is capable to express any permutation invariant function that operates on set inputs [39]. Since a multi-instance bags are both permutation invariant and contain instances of varying sizes [15], it is adequate to model its bag information as a set function of the instances.

The second missing piece of the puzzle is how to separate $\boldsymbol{z}^c$ from $\boldsymbol{z}^e$ in the identified latent factors $\boldsymbol{z}$, since we would only require $\boldsymbol{z}^c$ for the downstream tasks. To solve this problem, we take a different approach from a typical VAE-based disentanglement algorithm where their goal is to ensure the components of $\boldsymbol{z}$, i.e., z_p for $p = 1, \cdots, d$, are independent and correspond to meaningful semantics [13]. Although component-wise disentanglement is desirable for causal representation learning in general, due to the fact that multi-instance bags are only weakly supervised at the bag level, we argue that a more pragmatic approach is to utilize the bag labels for separating the causal factors $\boldsymbol{z}^c$ of the instance labels from the non-causal ones $\boldsymbol{z}^e$ without requiring additional information.

We now revisit the graphical model in Figure 1 to discuss the benefit of not pursuing component-wise disentanglement and using only $\boldsymbol{z}^c$ for instance classification. The instance labels y are correlated with both $\boldsymbol{z}^c$ and $\boldsymbol{z}^e$, and thus the entirety of $\boldsymbol{z}$ is predictive of y. Therefore, for supervised learning tasks that assume i.i.d. training and test data, using $\boldsymbol{z}$ is preferable for prediction. However, in MIL our goal is to predict the instance labels from unseen test bags. Since the bag-inherited factor $\boldsymbol{z}^e$ may be different among training and test bags, using $\boldsymbol{z}^e$ would negatively affect prediction. Furthermore, since instance classifiers in MIL must infer instance-level labels from bag-level weak supervision, excluding spurious correlations and including only the low dimensional causal representation also reduces the difficulty in learning the instance classifier. Another important benefit of separating $\boldsymbol{z}^c$ from $\boldsymbol{z}^e$ is that it also promotes OOD generalization. By considering bags as environments and the non-causal factors $\boldsymbol{z}^e$ as the effects of environment-induced biases, using $\boldsymbol{z}^c$ and excluding $\boldsymbol{z}^e$ improves prediction robustness since $p(y|\boldsymbol{z}^c)$ is invariant across bags as specified by Assumption 1(d), which also coincides with the goal of invariant causal prediction algorithms such as [4, 1, 23].

If the instance labels are known, then our goal of disentanglement can be straightforwardly achieved by utilizing the PC algorithm [33] to learn a Markov equivalent class of DAGs and find the direct causes as done in [23]. Unfortunately, this is not the case for MIL where only bag labels are known. To solve this problem, we propose to utilize the bag label Y with the ELBO in Equation 5 using a bag-wise maximum operator. Specifically, we optimize for the following target

$$\begin{aligned}\mathcal{L}_{\text{CausalMIL}} = &\log p_{\boldsymbol{f}}(\boldsymbol{x}^*|\boldsymbol{z}^*) + \alpha \log p_{\boldsymbol{\omega}}(Y|\boldsymbol{z}^*) - \text{KL}[q_{\boldsymbol{\phi}}(\boldsymbol{z}^*|\boldsymbol{x}^*, \boldsymbol{B})||p_{\boldsymbol{T},\boldsymbol{\lambda}}(\boldsymbol{z}^*|\boldsymbol{B})] \\ &+ \log p_{\boldsymbol{\vartheta}}(\boldsymbol{B}|\boldsymbol{z}) - \text{KL}[q_{\boldsymbol{\psi}}(\boldsymbol{z}|\boldsymbol{B})||p(\boldsymbol{B})],\end{aligned} \tag{8}$$

where $\boldsymbol{z}^* = \arg\max_{\boldsymbol{B}} p_{\boldsymbol{\omega}}(Y|\boldsymbol{z})$, and $p_{\boldsymbol{\omega}}(Y|\boldsymbol{z})$ is a *linear* classifier with α as its weighting hyper-parameter. It is also worth noting that the above ELBO is written with respect to a multi-instance bag instead of the instances. Furthermore, the last two terms can be obtained by deriving an ELBO for $\log p_{\boldsymbol{\vartheta}}(\boldsymbol{B})$ in Equation 5 [25].

In other words, the reconstruction, classification, and conditional KL terms in the ELBO are optimized for *only one instance per bag*, and the instance is chosen using a maximum operator to satisfy the standard multi-instance assumption. The formulation of Equation 8 is effective for learning $\boldsymbol{z}^c$ while ignoring $\boldsymbol{z}^e$ because of the following two reasons. Firstly, as the ELBO is optimized per mini-batches consisted of many multi-instance bags using the reparameterization trick [18], the optimization process will force the linear classifier $p_{\boldsymbol{\omega}}(y|\boldsymbol{z})$ to predict using the $\boldsymbol{z}^c$ and ignore $\boldsymbol{z}^e$; otherwise, the linearity of $p_{\boldsymbol{\omega}}(y|\boldsymbol{z})$ will not be able to predict the instance labels since $\boldsymbol{z}^e$ are different for different bags in the mini-batch. Moreover, because the reconstruction probabilities are evaluated at only one latent factor for each bag, the encoder $q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x}, \boldsymbol{B})$ is encouraged to encode only $\boldsymbol{z}^c$ since it carries the content causal information for reconstructing the positive $\boldsymbol{x}$ and remains invariant across different bags in the mini-batch. To see this, consider the opposite scenario where the reconstruction losses are evaluated for every instance in the bag, the encoder would also encode $\boldsymbol{z}^e$ since it remains the same within that bag. Therefore, CausalMIL is targeted to learn $\boldsymbol{z}^c$ that are causal to the positive instances of a bag, and ignores both the non-causal factors $\boldsymbol{z}^e$ and the causal factors for negative instances, as illustrated by the reconstructions in Figure 1 (right) and Figure 2.

It is worth noting that there exist two types of approximate inference in CausalMIL. The first one, common to almost all VAE-based algorithms, is the amortization gaps between the ELBOs and the marginal log-likelihoods caused by amortizing the variational parameters over the entire dataset, instead of optimizing for each training sample individually. The second one, specific to CausalMIL, is the approximation gap of only inferring the log-likelihood of one instances per each bag. However, as

Table 1: Results of instance label prediction performances on MNIST, FashionMNIST, and KuzushijiMNIST bags. The reported means $\pm$ std are the macro averaged over 10 one-vs-rest datasets, respectively.

	MNIST Bags		FashionMNIST Bags		KuzushijiMNIST Bags	
	F-score	AUC-PR	F-score	AUC-PR	F-score	AUC-PR
mi-NET	0.595±.078	0.702±.165	0.251±.290	0.329±.450	0.502±.106	0.616±.164
Attn-MIL	0.712±.147	0.776±.222	0.398±.243	0.534±.144	0.638±.145	0.654±.209
KSA-MIL	0.775±.092	0.845±.122	0.545±.288	0.617±.360	0.707±.084	0.782±.116
MIVAE	0.901±.035	0.921±.056	0.701±.257	0.733±.271	0.779±.147	0.838±.102
CausalMIL	**0.966±.018**	**0.981±.018**	**0.748±.206**	**0.822±.168**	**0.833±0.109**	**0.934±.075**

corroborated by our experiment results, these approximation behaviors does not prevent the success of CausalMIL in instance label prediction and OOD generalization.

4 Experiments

In this section, we first evaluate the instance label prediction performances of CausalMIL against MIL baselines including mi-Net [35], Attention-based MIL (Attn-MIL) [15], Kernel Self-Attention-based MIL (KSA-MIL) [29], Multi-Instance Variational Autoencoder (MIVAE) [40]. Then, we evaluate the out-of-distribution generalization ability of CausalMIL by comparing with supervised learning algorithms including ERM and its variants, Invariant Risk Minimization (IRM) [4], IRM with Game Theory (IRM GAME) [1], Invariant Causal Representation Learning (iCaRL) [23].

We implemented CausalMIL using PyTorch and conducted most of the experiments with a single NVIDIA RTX3090 GPU. Detailed settings, parameters, and data for reproducing the results are provided in the Appendices. The implementation code is publicly available at https://github.com/WeijiaZhang24/CausalMIL.

4.1 Instance Label Prediction

4.1.1 Datasets

Our quantitative evaluation first utilizes 30 multi-instance classification datasets generated from MNIST [19], FashionMNIST [38], and KuzushijiMNIST [7] datasets. The bag construction procedure is similar to the multi-instance bags generation from the 20 Newsgroup corpus [43]. Specifically, we create the multi-instance MNIST-bags such that each bag contains a number of images where its bag size is drawn from a Gaussian distribution with fixed mean and variance. The bag is positive if it contains a target digit, e.g., '4', and negative if otherwise. Using 10 different digits as targets, we obtain 10 multi-instance MNIST-bag datasets. Similarly, we construct 10 FashionMNIST-bags and 10 KuzushijiMNIST-bags. The mean bag size to 50 with a standard deviation of 10 and each positive bag contains about 10% of positive instances.

We also evaluate CausalMIL on a hematoxylin and eosin (H&E) stained Colon Cancer histopathology task [32]. Histopathology from whole-slide images is an important application of MIL since supervised predictions require pathologists to provide pixel-level annotations which is extremely expensive and time-consuming. The dataset contains 100 images obtained from tissues of either normal or malignant patient tissues. For each image bag, the instances are generated as patches 27×27 pixels using markings of major nuclei for each cell. A total amount of 22,444 instances (~220 instances per bag) are provided with ground truth instance labels, i.e. whether the cell is epithelial. A bag is labeled positive if it contains at least one epithelial patch and negative if otherwise.

4.1.2 Quantitative Results

For quantitative evaluation, we use the area under the Precision-Recall curve (AUC-PR), instead of area under the ROC curve due to class imbalance at the instance level. For example, although the original MNIST datasets are balanced, only ~5% of the instances are positive in the MNIST-bags. We also report the f-scores of the compared methods. For the methods that directly output instance prediction such as mi-NET, MIVAE, and CausalMIL, the results are obtained from thresholding their predicted scores. For methods based on the attention mechanism, such as Attn-MIL and KSA-MIL, the results are obtained by first normalizing the instance-to-bag attention weights to $[0, 1]$ and

Table 2: Instance label prediction performances of the Colon Cancer dataset. Experiments are repeated for 5 times and the average metric±standard deviation of 5-fold cross validations are reported.

Method	F-score	AUC-PR
mi-Net	0.392±.017	0.491±.028
Attn-MIL	0.466±.037	0.536±.014
KSA-MIL	0.510±.029	0.578±.025
MIVAE	0.675±.033	0.747±.032
CausalMIL	**0.833±.024**	**0.878±.046**

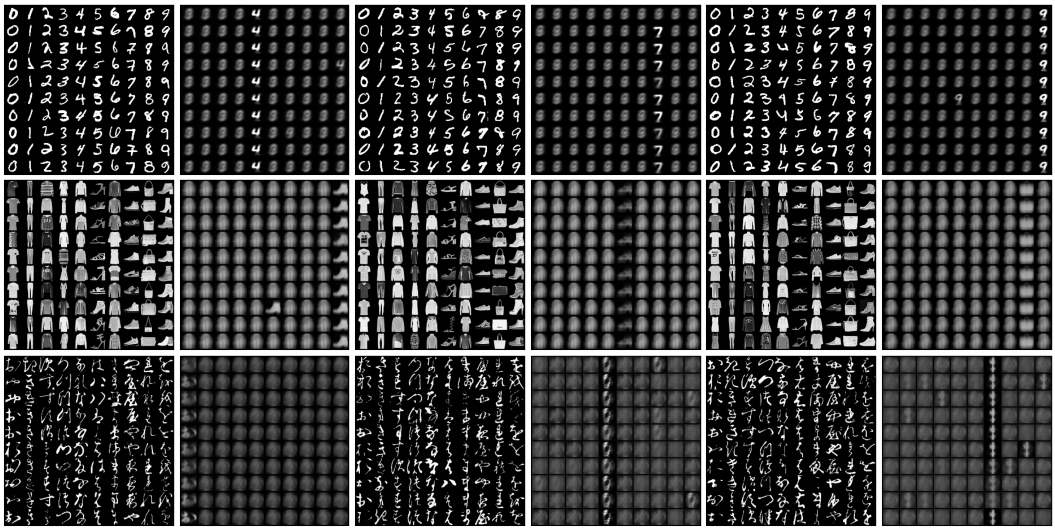

Figure 2: MNIST digits and CausalMIL reconstructions for digits '4', '7', '9'. Figures are best viewed when zoomed. FashionMNIST objects and CausalMIL reconstructions for 'dress', 'sandal', and 'handbag'. More results can be found in the Appendices. Figures best viewed when zoomed.

applying the threshold. As there is usually no instance label available in real application scenarios, we use 0.5 uniformly as the classification threshold.

Table 1 shows the quantitative results on MNIST, FashionMNIST, and KuzushijiMNIST bags. CausalMIL performs significantly better than the compared methods on both metrics. Note that the large standard deviations for FashionMNIST-bags are because some fashion objects are much more difficult to classify (e.g., distinguishing shirts from pullovers and coats). Furthermore, we can see that attention-based MIL algorithms exhibit significant performance variations even for MNIST-bags. A possible reason is that when the data satisfy the standard multi-instance assumption, learning a weighted average of the individual instance contributions is not efficient. Among the compared algorithms, CausalMIL achieves the best performances with the lowest standard deviations.

Table 2 shows the instance prediction performances of the compared algorithms on the Colon Cancer dataset. The result trends are similar to previous ones. For discriminative deep learning-based algorithms, attention-pooling (Attn-MIL) performs better than max-pooling (mi-Net), while incorporating the self-attention mechanism (KSA-MIL) further improves the performances. A recent VAE-based generative MIL algorithm (MIVAE) performs better than the discriminative algorithms but still significantly worse than CausalMIL because it does not learn meaningful representations and does not provide identifiability guarantee.

4.1.3 Qualitative Results

We now qualitatively evaluate the inferred latent factors of CausalMIL by examining the reconstructions of decoder $p_{\boldsymbol{f}}(\boldsymbol{x}|\boldsymbol{z})$ using the inferred latent factors in Figure 2. For example, in the first row we show the original images and reconstructions of MNIST-bags using '4', '7', and '9' as positive instances. In the reconstructions of '4' we can see that CausalMIL learns the semantic causal characteristics of the positive instances while ignoring the others. The '4's are reconstructed with a

Table 3: Classification performances on distributionally biased ColoredMNIST and ColoredFashionMNIST datasets in terms of accuracy $\pm$ standard deviation. CausalMIL (ours) is weakly supervised whereas the other approaches are supervised.

	ColoredMNIST		ColoredFashionMNIST	
	Train	Test	Train	Test
ERM	0.849±.002	0.105±.007	0.832±.010	0.225±.007
ERM1	0.848±.002	0.109±.005	0.813±.014	0.333±.089
ERM2	0.850±.002	0.101±.002	0.844±.019	0.132±.008
ROBUST MIN MAX	0.843±.004	0.152±.025	0.828±.001	0.292±.086
IRM	0.593±.044	0.628±.096	0.398±.243	0.534±.194
IRM GAME	0.634±.011	0.599±.027	0.690±.101	0.702±.015
iCaRL	0.706±.008	0.688±.007	0.545±.288	0.617±.360
MIVAE	0.810±.005	0.156±.003	0.823±.003	0.284±.120
CausalMIL	**0.948±.003**	**0.925±.004**	**0.911±.006**	**0.866±.009**

'U' shape at the top and a short line at the bottom, while other digits are reconstructed to noise. The 'U' shape of '4' is clear because it is crucial for the positive instance; however, the line at the bottom is blurry since it can be written in different angles and the angles are irrelevant. The reconstructions from FashionMNIST-bags and KuzushijiMNIST-bags datasets also corroborate the above results.

4.2 Out-of-Distribution Generalization

To further validate whether CausalMIL identifies the causal representations $\boldsymbol{z}^c$, we evaluate CausalMIL on out-of-distribution (OOD) generalization tasks and compare it supervised algorithms designed specifically for OOD generalization [4, 1, 23].

We utilize two widely-adopted biased classification tasks, ColoredMNIST and ColoredFashionMNIST that are commonly used in the invariant risk minimization literature [4]. The datasets are constructed following the same procedure as in [4], where the task is to classify whether a digit is less than 5. The instance labels are first binarized and randomly flipped with 25% probability. Then, the images are colored by green with probabilities p_e which vary across environments, while the rest $1 - p_e$ of the images are colored by red. There are three environments (two training, one test) where $p_e = 0.1, 0.2,$ and 0.9, respectively. Therefore, simply predicting colors instead of digits will lead to high accuracy in the training environments, but the correlation is reversed in the test environment. For CausalMIL, we construct multi-instance bags based on the flipped instance labels utilizing instances from both training environments. Only the bag labels are provided to CausalMIL during training, while the instances from the test environment are used for testing.

We compare the instance label prediction performances of CausalMIL against two categories of *supervised algorithms*. The first category includes standard Empirical Risk Minimization using CNN trained with both (ERM) and each of the training environment (ERM1 and ERM2), and ROBUST MIN MAX which minimizes the maximum loss across different environments. The second category are specifically designed OOD generalization algorithms including Invariant Risk Minimization(IRM) [4], IRM GAME [1], and iCaRL [23]. Considering that most of the baseline results come from iCaRL [23], we set CausalMIL to use the same convolutional structure as those adopted in [23] for fair comparison.

The results are reported in Table 3. We can see that CausalMIL performs significantly better than all compared supervised methods. There are two reasons behind this. The first reason is that CausalMIL effectively forces the biases into $\boldsymbol{z}^e$ by disentangling the causal from non-causal factors, and uses only $\boldsymbol{z}^c$ for prediction. Although previous supervised OOD generalization algorithms, i.e., IRM and iCaRL, aim to identify and only use the causes of y to improve robustness, they mainly rely on the instance-level labels for identifiability and utilize the scarcely available environments for robustness. Unfortunately, as the instance-level labels are biased from the training environments, their approaches are less effective. On the contrary, in CausalMIL we rely on the abundance of bag information for identification. The second reason is that the formulation of MIL is naturally more robust to instance-level label noises as positive bags naturally consist of both positive and negative instances. Furthermore, previous theoretical results have also shown that generative MIL algorithm [9] is effective even when negative bags are falsely labeled as positive.

5 Related Work

Most MIL algorithms can be categorized into two groups according to whether they work at the bag space or at the instance space. Bag space algorithms [43, 37, 41, 10] work by embedding multi-instance bags into a single feature vector representation and then solving the single-instance learning problem in the embedded space, and therefore are capable of predicting instance labels. Instance space MIL algorithms aim to directly separate the positive instances from the negative ones [2, 16, 12, 34]. These algorithms can be used to predict instance labels, although not all of them are explicitly designed for the task.However, these algorithms only work with pre-computed features.

Recently, several deep learning-based MIL algorithms have been proposed by utilizing permutation-invariant pooling operations, and can be used for instance label prediction. For example, [35] used a max-pooling layer with a fully connected neural network; [15] introduced the attention mechanism as a permutation-invariant MIL pooling operation and used the attention weights for instance label prediction; [29] proposed to integrate self-attention with the attention mechanism for capturing the non-i.i.d. information among instances; [21] utilized Gumbel reparametrization and proposed an algorithm for the generalized MIL assumption. Unfortunately, none of the above methods learn semantically meaningful representations.

Unsupervised learning of meaningful representations with VAE-based models has attracted much attention [13, 6]. Since it has been shown that unsupervised disentanglement is theoretically impossible without inductive biases [22], various methods have been proposed for learning disentangled representations using different forms of weak supervisions. One form of weakly-supervised disentanglement learning that is closely related to MIL is group-based VAE [5, 14]. Similar to MIL where instances are organized into bags, group-based algorithms require the objects to be divided into groups: those within the same group share the same *content* but have different *styles* that are independent of the contents, while those among different groups should have different contents.

Recently, MIVAE [40] is proposed based on connections between MIL and group-based disentangled representation learning which utilizes VAE for capturing the dependencies among instances as the shared content factors. However, MIVAE also does not learn semantically meaningful latents, nor does it provide identifiability on what latents it actually learns because of its unconditional prior. Please refer to the Appendices for a more detailed comparison between CausalMIL and MIVE.

6 Conclusion

In this work, we proposed CausalMIL, an algorithm that learns semantically meaningful causal representations from multi-instance bags and the corresponding bag-level supervision. The learned representations not only significantly improve the performances of traditional MIL tasks such as instance label prediction, but also exhibits notable performances in supervised learning tasks such as OOD generalization. Qualitative and quantative evaluation results show that CausalMIL performs significantly better than existing state-of-the-art deep MIL algorithms on semi-synthetic and real-world datasets.

There are some limitations of our method and many future directions worth investigating. For example, CausalMIL is only designed for the standard multi-instance assumption. It would be interesting to see if the results extend to other assumptions in multi-instance learning. A possible way for such extension is to utilize the Gumbel reparameterization as discussed in [21]. Another possible direction would be developing an extension to learning from positive unlabeled (PU) examples, another weakly supervised learning problem that is closely related to MIL.

Acknowledgments and Disclosure of Funding

The authors wish to thank the anonymous reviewers for their constructive comments and suggestions. This work was supported by the National Science Foundation of China (62206047,72204110). We thank the Big Data Center of Southeast University for providing the facility support on the numerical calculations in this paper.

References

[1] Kartik Ahuja, Karthikeyan Shanmugam, Kush Varshney, and Amit Dhurandhar. Invariant risk minimization games. In *Proceedings of the 37th International Conference on Machine Learning*, PMLR 119, pages 145–155, 2020.

[2] Stuart Andrews, Ioannis Tsochantaridis, and Thomas Hofmann. Support vector machines for multiple-instance learning. In *Advances in Neural Information Processing Systems 15*, pages 577–584, 2002.

[3] Stefanos Angelidis and Mirella Lapata. Multiple instance learning networks for fine-grained sentiment analysis. *Transactions of the Association for Computational Linguistics*, (6):17–34, 2018.

[4] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint*, arXiv: 1907.02893, 2019.

[5] Diane Bouchacourt, Ryota Tomioka, and Sebastian Nowozin. Multi-level variational autoencoder: Learning disentangled representations from grouped observations. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*, pages 2095–2102, 2018.

[6] Ricky T. Q. Chen, Xuechen Li, Roger Grosse, and David Duvenaud. Isolating sources of disentanglement in variational autoencoders. In *Advances in Neural Information Processing Systems 31*, page 2615–2625, 2018.

[7] Tarin Clanuwat, Mikel Bober-Irizar, Asanobu Kitamoto, Alex Lamb, Kazuaki Yamamoto, and David Ha. Deep learning for classical Japanese literature. *arXiv preprint*, arXiv:1812.01718, 2018.

[8] Thomas G. Dietterich, Richard H. Lathrop, and Tomás Lozano-Pérez. Solving the multiple instance problem with axis-parallel rectangles. *Artificial Intelligence*, 89(1-2):31–71, 1997.

[9] Gary Doran and Soumya Ray. Multiple-instance learning from distributions. *Journal of Machine Learning Research*, 17(128):1–50, 2016.

[10] Lei Feng, Senlin Shu, Yuzhou Cao, Lue Tao, Hongxin Wei, Tao Xiang, Bo An, and Gang Niu. Multiple-instance learning from similar and dissimilar bags. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, page 374–382, 2021.

[11] James Foulds and Eibe Frank. A review of multi-instance learning assumptions. *The Knowledge Engineering Review*, 25(1):1–25, 2010.

[12] Manuel Haussmann, Fred A. Hamprecht, and Melih Kandemir. Variational bayesian multiple instance learning with gaussian processes. In *2017 IEEE Conference on Computer Vision and Pattern Recognition*, pages 6570–6579, 2017.

[13] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-VAE: Learning basic visual concepts with a constrained variational framework. In *Proceedings of the 5th International Conference on Learning Representations*. https://openreview.net/forum?id=Sy2fzU9gl, 2017.

[14] Haruo Hosoya. Group-based learning of disentangled representationswith generalizability for novel contents. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, pages 2506–2513, 2019.

[15] Maximilian Ilse, Jakub Tomczak, and Max Welling. Attention-based deep multiple instance learning. In *Proceedings of the 35th International Conference on Machine Learning*, PMLR 80, pages 2127–2136, 2018.

[16] Melih Kandemir and Fred A. Hamprecht. Instance label prediction by dirichlet process multiple instance learning. In *Proceedings of the 30th Conference on Uncertainty in Artificial Intelligence*, pages 380–389, 2014.

[17] Ilyes Khemakhem, Diederik P. Kingma, Ricardo Pio Monti, and Aapo Hyvärinen. Variational autoencoders and nonlinear ICA: A unifying framework. In *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistcs*, PMLR 108, pages 2007–2017, 2020.

[18] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. In *Proceedings of the 2nd International Conference on Learning Representations*. https://openreview.net/forum?id=33X9fd2-9FyZd, 2014.

[19] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[20] Bin Li, Yin Li, and Kevin W Eliceiri. 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*, pages 14318–14328, 2021.

[21] Xin-Chun Li, De-Chuan Zhan, Jia-Qi Yang, and Yi Shi. Deep multiple instance selection. *Science China Information Sciences*, 64(3), 2021.

[22] Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Raetsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *Proceedings of the 36th International Conference on Machine Learning*, PMLR 97, pages 4114–4124, 2019.

[23] Chaochao Lu, Yuhuai Wu, José Miguel Hernández-Lobato, and Bernhard Schölkopf. Invariant causal representation learning for out-of-distribution generalization. In *International Conference on Learning Representations*. https://openreview.net/forum?id=-e4EXDWXnSn, 2022.

[24] Ming Y. Lu, Tiffany Y. Chen, Drew F.K. Williamson, Melissa Zhao, Maha Shady, Jana Lipkova, and Faisal Mahmood. AI-based pathology predicts origins for cancers of unknown primary. *Nature*, 594:106–110, 6 2021.

[25] Graziano Mita, Maurizio Filippone, and Pietro Michiardi. An identifiable double vae for disentangled representations. In *Proceedings of the 38th International Conference on Machine Learning, PMLR 139*, pages 7769–7779, 2021.

[26] Judea Pearl. *Causality*. Cambridge University Press, 2009.

[27] Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(5):947–1012, 2016.

[28] Jan Ramon and Luc De Raedt. Multi-instance neural networks. In *Proceedings of the ICML-2000 Workshop on Attribute-value and Relational Learning*, pages 53–60, 2000.

[29] Dawid Rymarczyk, Adriana Borowa, Jacek Tabor, and Bartosz Zielinsk. Kernel self-attention for weakly-supervised image classification using deepmultiple instance learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1721–1730, 2021.

[30] Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021.

[31] Xiaoshuang Shi, Fuyong Xing, Yuanpu Xie, Zizhao Zhang, Lei Cui, and Lin Yang. Loss-based attention for deep multiple instance learning. *Proceedings of the 34th AAAI Conference on Artificial Intelligence*, pages 5742–5749, 2020.

[32] Korsuk Sirinukunwattana, Shan E Ahmed Raza, Yee-Wah Tsang, David R. J. Snead, Ian A. Cree, and Nasir M. Rajpoot. Locality sensitive deep learning for detection and classification of nuclei in routine colon cancer histology images. *IEEE Transactions on Medical Imaging*, 35(5):1196–1206, 2016.

[33] Peter Spirtes, Clark Glymour, and Richard Scheines. *Causation, Prediction, and Search*. The MIT Press, 2000.

[34] Fulton Wang and Ali Pinar. The multiple instance learning gaussian process probit model. In *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics, PMLR 130*, pages 3034–3042, 2021.

[35] Xinggang Wang, Yongluan Yan, Peng Tang, Xiang Bai, and Wenyu Liu. Revisiting multiple instance neural networks. *Pattern Recognition*, 74:15–24, 2018.

[36] Yun Wang, Juncheng Li, and Florian Metze. A comparison of five multiple instance learning pooling functions for sound event detection with weak labeling. In *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 31–35, 2019.

[37] Xiu-Shen Wei, Jianxin Wu, and Zhi-Hua Zhou. Scalable algorithms for multi-instance learning. *IEEE Transactions on Neural Networks and Learning Systems*, 28(4):975–987, 2017.

[38] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-MNIST: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint*, arXiv:1708.07747, 2017.

[39] Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabas Poczos, Russ R Salakhutdinov, and Alexander J Smola. Deep sets. In *Advances in Neural Information Processing Systems 30*, 2017.

[40] Weijia Zhang. Non-i.i.d. multi-instance learning for predicting instance and bag labels using variational autoencoder. In *Proceedings of the 30th International Joint Conference on Artificial Intelligence*, pages 3377–3383, 2021.

[41] Weijia Zhang, Lin Liu, and Jiuyong Li. Robust multi-instance learning with stable instances. In *Proceedings of the 24th European Conference on Artificial Intelligence*, pages 1682–1689, 2020.

[42] Zhi-Hua Zhou. A brief introduction to weakly supervised learning. *National Science Review*, 5(1):44–53, 2017.

[43] Zhi-Hua Zhou, Yu-Yin Sun, and Yu-Feng Li. Multi-instance learning by treating instances as non-i.i.d. samples. In *Proceedings of the 26th International Conference on Machine Learning*, pages 1249–1256, 2009.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [N/A]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [N/A]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes] In the Appendix.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes]

(d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]