Partial Label Causal Representation Learning for Instance-Dependent Supervision and Domain Generalization

Yi-Zhi Wang[1,3], **Weijia Zhang**[2], **Min-Ling Zhang**[1,3*]

[1]School of Computer Science and Engineering, Southeast University, Nanjing 210096, China
[2]School of Information and Physical Sciences, The University of Newcastle, Callaghan, NSW 2308, Australia
[3]Key Laboratory of Computer Network and Information Integration (Southeast University), Ministry of Education, China
wang_yz@seu.edu.cn, weijia.zhang@newcastle.edu.au, zhangml@seu.edu.cn

Abstract

Partial label learning (PLL) addresses situations where each training example is associated with a set of candidate labels, among which only one corresponds to the true class label. As the candidate labels often come from crowdsourced workers, their generation is inherently dependent on the features of the instance. Existing PLL methods primarily aim to resolve these ambiguous labels to enhance classification accuracy, overlooking the opportunity to use this feature dependency for causal representation learning. This focus on accuracy can make PLL systems vulnerable to stylistic variations and shifts in domain. In this paper, we explore the learning of causal representations within an instance-dependent PLL framework, introducing a new approach that uncovers identifiable latent representations. By separating content from style in the identified causal representation, we introduce CausalPLL+, an algorithm for instance-dependent PLL based on causal representation. Our algorithm performs exceptionally well in terms of both classification accuracy and generalization robustness. Qualitative and quantitative experiments on instance-dependent PLL benchmarks and domain generalization tasks verify the effectiveness of our approach.

1 Introduction

Causal Representation Learning (CRL) (Schölkopf et al. 2021) aims to infer compact high-level latent variables from high-dimensional and low-level observations. A core task in CRL is learning identifiable latent representation, i.e., developing representation learning algorithms that can provably identify high-level latent factors such as an object's shape, location, and colour. While problems such as domain shift, out-of-distribution samples, and data bias have long plagued modern statistical learning systems (Liu et al. 2022; Zhu et al. 2025), CRL offers a unique and promising perspective to achieve greater effectiveness in robustness and generalization.

Since previous work has demonstrated that learning identifiable representations is impossible for arbitrary data-generating process in an unsupervised setting (Locatello et al. 2019; Khemakhem et al. 2020), much of the recent efforts in CRL have been diverted to learning causal representation from data with additional structures and supervisions (Khemakhem et al. 2020; Kivva et al. 2022). For example, several recent studies have delved into understanding causal representations with additional information or under specific types of weak supervision signals (Zhang et al. 2022; Brehmer et al. 2022; Yao et al. 2021; Lin et al. 2024).

This paper explores the possibility of identifying causal representation under the Partial Label Learning (PLL) paradigm and its benefits for more stable and robust weakly supervised learning systems. PLL has garnered significant attention over the past decade as a form of weakly supervised learning due to its prevalence in many real-world applications, such as automatic image annotation (Chen, Patel, and Chellappa 2017; Tang, Zhang, and Zhang 2024b, 2023; Yang, Tang, and Zhang 2024), web mining (Luo and Orabona 2010; Scheffer, Decomain, and Wrobel 2001) and multimedia content analysis (Zeng et al. 2013; Cour et al. 2009; Tang et al. 2024; Tang, Zhang, and Zhang 2024a). Unlike standard supervised learning, where the training data contains i.i.d. samples associated with a single class label, learners in PLL are given samples associated with candidate label sets containing the (unknown) ground-truth label and several candidate labels.

In real-world PLL applications, candidate labels are typically provided by crowd-sourced annotators who select several labels that are likely correct. Therefore, the generation process of these candidate labels is closely tied to the characteristics of the instance, a concept adeptly termed Instance-Dependent Partial Label Learning (IDPLL) (Xu et al. 2021). IDPLL is a realistic yet particularly challenging scenario. The candidate labels are related to the sample feature, making it difficult for the model to discern the ground-truth labels from the candidate set. Furthermore, the ambiguity of unknown ground-truth labels makes it difficult for models to learn the core differences between different categories.

Since the generation of instance-dependent candidate labels inherently depends on the instance feature under the IDPLL setting, it is desirable to model the generative relationship between the instance features and their associated candidate labels. For example, consider an image with the ground-truth label 'Husky', which may be provided with false-positive candidate labels such as 'Wolf' and 'Samoyed' due to their visual similarity. Instead of treating the candidate label set as noises and disambiguating

[*]Corresponding author

the ground-truth label from false positive ones, it is advantageous to separate the characteristics specific to each breed from those shared across breeds and external conditions such as background or lighting. This would allow the model to concentrate on disambiguating the core features and reduce the influence of extraneous factors.

CRL provides unique tools for effectively modelling the generative relationship between the instance features and their associated candidate labels. As each candidate label inherently contains style and content information that is closely related to the instance feature, CRL algorithms have the potential to identify their corresponding low-dimensional latent factors. It is worth noting that the goal here is not to learn completely disentangled latent representation, i.e., identifying dimension-wise independent latent factors. Instead, we aim to identify latent representations that block-separated *content* from *style*, as this will not only facilitate the task of PLL classification but also improve the classifier's robustness to distribution shifts.

From the perspective of exploiting instance-dependent candidate labels for causal representation learning, this work proposes a novel generative approach for effectively modelling the generation process of instances and candidate label sets while extracting identifiable and content-style disentangled causal representations. Furthermore, we introduce a prior-based contrastive learning method and a label refinement disambiguation strategy to further improve the model's representation quality and classification performance. The proposed model not only achieves state-of-the-art classification performance on IDPLL benchmarks but also demonstrates robustness to the changes between training and test distributions. Our contributions can be summarized as:

- We introduce a novel VAE framework enabling the model to learn identifiable causal representations from data, achieving disentanglement of content and style.
- Based on this framework, we propose an effective Partial Label Learning algorithm, CausalPLL+, which enables effective disambiguation in instance-dependent scenarios through contrastive learning and label refinement.
- We conduct extensive empirical studies on various datasets and settings, proposing a new, more realistic method for IDPLL data generation. Experimental results demonstrate the effectiveness of CausalPLL+ in IDPLL classification and domain shift scenarios.

2 Related Work

2.1 VAE and Identifiable Causal Representations

Variational Autoencoders (VAEs) are a class of deep generative models that combine amortized variational inference and neural networks to model the generation process by fitting the posterior and likelihood distributions of samples (Kingma and Welling 2013). Specifically, VAEs optimize the evidence lower bound (ELBO) of the likelihood:

$$\mathbb{E}_{\boldsymbol{z}\sim q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})}\left[\ln p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z})\right]-\mathrm{KL}(q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})||p_{\boldsymbol{\theta}}(\boldsymbol{z})). \quad (1)$$

VAEs are inherently related to the field of causal representation learning due to their flexibility in modelling probability graphical models. This has attracted considerable research efforts into their latent variable identifiability. Locatello et al. suggests that it is impossible to learn identifiable representations from the data in completely unsupervised settings. Meanwhile, Khemakhem et al. provided identifiable results under the VAE framework for the first time. It has been shown that latent factors $\boldsymbol{z}$ can be identified by employing a conditionally factorized prior distribution $p_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{u})$ over the latent variables, where $\boldsymbol{u}$ is an additionally observed variable (Khemakhem et al. 2020).

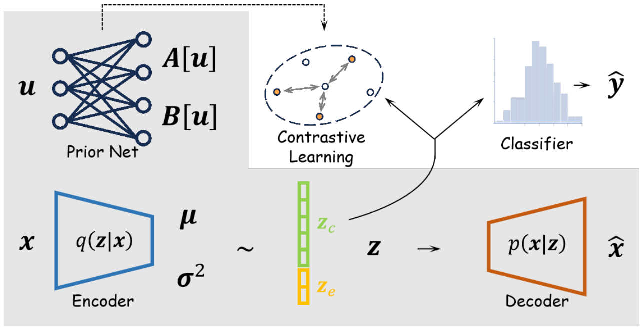

Figure 1: Framework of CausalPLL+. Unless otherwise specified (arrows), components within the shaded area do not exchange information with those outside. The dashed arrow on prior net and contrastive learning module means that CL is conducted between $\boldsymbol{z}_c$ and distribution parameters of all classes in the prior net, rather than $\boldsymbol{A}[\boldsymbol{u}]$ and $\boldsymbol{B}[\boldsymbol{u}]$.

However, most recent efforts in CRL have been focusing on learning identifiable latent representations that are mutually independent and their causal structures (Brehmer et al. 2022; Lin et al. 2024); however, our focus is instead inferencing identifiable representations under the data generation process of IDPLL and exploring their benefits for partial label classification.

2.2 Partial Label Learning

Partial Label Learning (PLL) (Cour et al. 2009) is a subclass of weakly supervised learning (Zhou 2017). In PLL, each training sample is associated with a candidate label set containing an unknown ground-truth label and several false positive labels. Early efforts in PLL have focused on scenarios in which the candidate labels are randomly generated or class-dependent. For instance, Lv et al. (2020) and Wen et al. employ self-training techniques to determine ground-truth labels during training iteratively. Feng et al. (2020) studies consistent classifiers under the assumption of uniformly generated partial labels. Wu, Wang, and Zhang (2022) studies manifold-preserving consistency regularizations in PLL.

As the generation of candidate labels depends on the instance features in real-world PLL applications, IDPLL (Xu et al. 2021) more closely resembles the data generation process of practical situations. Xu et al. (2021) employ variational inference to estimate the latent label distribution. Wu, Wang, and Zhang (2024) perform knowledge distillation and leverage a rectifcation process to obtain reliable representations. However, existing literature (Qiao, Xu, and Geng

2022; Xia et al. 2022; Xu et al. 2023) rarely explores the generative process between examples and candidate label sets, nor does it address spurious features and domain shift issues in IDPLL scenarios. This paper explicitly models the generative relationship between instances and their candidate labels to extract causal representations which decouple content and style, improving classification and robustness.

3 Methodology

Notations Let $\mathcal{X} \subset \mathbb{R}^D$ denote the D-dimensional instance space and $\mathcal{Y} = \{1, 2, \cdots, K\}$ denote the label space with K distinct labels. $\mathcal{Z} \subset \mathbb{R}^M$ is the M-dimensional latent space where $M \ll D$. PLL assumes that the ground-truth label $y \in \mathcal{Y}$ of an instance $\boldsymbol{x} \in \mathcal{X}$ is contained within a candidate label set $\mathcal{S} \subset \mathcal{Y}$. For simplicity, we use the Boolean vector $\boldsymbol{s} \in \{0, 1\}^K$ to represent the partial label corresponding to $\mathcal{S}$. The goal of PLL is to learn a classifier $h : \mathcal{Z} \to \mathcal{Y}$ on a partial label dataset $\mathcal{D} = \{(\boldsymbol{x}_i, \boldsymbol{s}_i) | 1 \leq i \leq N\}$. For the classifier h, we use $h_k(\boldsymbol{z})$ to denote the output of classifier h on label k given input $\boldsymbol{z}$. For the VAE framework, we use $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{K \times M_c}$ to denote the matrix storing the mean and variance of the content prior $p(\boldsymbol{z}_c|\boldsymbol{u})$ for K categories, respectively. $\boldsymbol{u}$ stands for auxiliary variable, which is usually a normalized candidate label vector. For convenience, we denote the mean and variance corresponding to the categories contained in $\boldsymbol{u}$ as $\boldsymbol{A}[\boldsymbol{u}]$ and $\boldsymbol{B}[\boldsymbol{u}]$, respectively. Lastly, we use $\hat{\boldsymbol{s}}$ to represent the refined candidate labels.

3.1 Model Identifiability and Content-Style Disentangled Causal Representation

Identifiability in existing VAE frameworks is often achieved through a conditional prior $p(\boldsymbol{z}|\boldsymbol{u}))$, where $\boldsymbol{u}$ serves as an auxiliary variable. Suppose we could observe instance $\boldsymbol{x} \in \mathbb{R}^D$ and auxiliary variable $\boldsymbol{u} \in \mathbb{R}^K$, and $\boldsymbol{z} \in \mathbb{R}^M$ is a latent variable. The observed instance $\boldsymbol{x}$ can be regarded as generated by $\boldsymbol{z}$ through an arbitrary mixing function $\boldsymbol{f}$:

$$\boldsymbol{x} = \boldsymbol{f}(\boldsymbol{z}) + \boldsymbol{\epsilon}, \tag{2}$$

where $\boldsymbol{\epsilon}$ is a noise variable with probability density function $p(\boldsymbol{\epsilon})$ independent of $\boldsymbol{z}$ or $\boldsymbol{f}$. Hence, we can express the posterior likelihood of the data in the following form:

$$p_{\boldsymbol{f}}(\boldsymbol{x}|\boldsymbol{z}) = p_{\boldsymbol{\epsilon}}(\boldsymbol{x} - \boldsymbol{f}(\boldsymbol{z})). \tag{3}$$

Furthermore, let $\boldsymbol{\theta} = (\boldsymbol{f}, \boldsymbol{T}, \boldsymbol{\lambda})$ be the parameters of the following conditional generative model, the data generation process can be expressed as:

$$p_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{z}|\boldsymbol{u}) = p_{\boldsymbol{f}}(\boldsymbol{x}|\boldsymbol{z}) p_{\boldsymbol{T},\boldsymbol{\lambda}}(\boldsymbol{z}|\boldsymbol{u}). \tag{4}$$

Without loss of generality, it is common to assume that the latent prior distribution $p(\boldsymbol{z}|\boldsymbol{u})$ follows the exponential family distribution:

$$p_{\boldsymbol{T},\boldsymbol{\lambda}}(\boldsymbol{z}|\boldsymbol{u}) = \prod_i \frac{Q_i\left(z_i\right)}{Z_i(\boldsymbol{u})} \exp\left[\sum_{j=1}^{k} T_{i,j}\left(z_i\right) \lambda_{i,j}(\boldsymbol{u})\right]. \tag{5}$$

With the generative model specified according to (3)-(5), Khemakhem et al. (2020) have shown that the model parameters $(\boldsymbol{f}, \boldsymbol{T}, \boldsymbol{\lambda})$ can be identified up to an equivalence class induced by component-wise and invertible linear transformations with the following assumptions:

(a) The set $\{\boldsymbol{x} \in \mathcal{X} | \phi_{\boldsymbol{\epsilon}}(\boldsymbol{x}) = 0\}$ has measure zero, where $\phi_{\boldsymbol{\epsilon}}$ is the characteristic function of the density $p_{\boldsymbol{\epsilon}}$ defined in (3).

(b) The mixing function $\boldsymbol{f}$ in (3) is injective.

(c) The sufficient statistics $T_{i,j}$ in (5) are differentiable almost everywhere, and $(T_{i,j})_{1 \leq j \leq k}$ are linearly independent on any subset of $\mathcal{X}$ of measure greater than zero.

(d) There exist $nk + 1$ distinct points $\boldsymbol{u}^0, \ldots, \boldsymbol{u}^{nk}$ such that the matrix

$$L = (\boldsymbol{\lambda}(\boldsymbol{u}_1) - \boldsymbol{\lambda}(\boldsymbol{u}_0), \ldots, \boldsymbol{\lambda}(\boldsymbol{u}_{nk}) - \boldsymbol{\lambda}(\boldsymbol{u}_0))$$

of size $nk \times nk$ is invertible.

According to the above theory, one basis for achieving identifiability is to introduce a prior $p(\boldsymbol{z}|\boldsymbol{u})$ which is conditioned on an auxiliary variable $\boldsymbol{u}$. In weakly supervised classification problems such as semi-supervised learning and multi-instance learning (Zhang et al. 2022), a common approach to learning causal representations is to utilize the class information as an auxiliary variable and map it to prior parameters through a prior network. Unlike other weakly supervised learning methods, the supervision information in PLL does not provide exact class label indices, but rather a set of candidate labels. This means that the weak supervision signals in PLL cannot be directly translated into a specific class priority as in other methods.

A naive approach is straightforwardly mapping the labels in the candidate label set into a set of prior distributions; however, this approach poses several problems. Firstly, as the candidate label sets often exhibit highly imbalanced and long-tailed distributions (Wu, Wang, and Zhang 2024), using them directly would severely impede model learning. Furthermore, as the candidate label sets only contain class labels, i.e., label indexes instead of actual semantics, attempting to fit such simple relationships with flexible variational inference models can lead to the collapsing of the variational posterior, resulting in learning failures, as observed in our preliminary experiments.

To avoid posterior collapsing and effectively infer representations from candidate label sets, we propose to exploit the auxiliary information without directly mapping the candidate label sets into prior parameters. Specifically, assuming that each class's latent corresponds to a Gaussian distribution in the latent representation space, we consider them a Gaussian mixture distribution containing the mixture components of their candidate labels with unknown mixing coefficients. Although the Gaussian mixture is not an exponential family distribution, there exists another Gaussian distribution p^* that minimizes the reverse KL divergence between this distribution and the Gaussian mixture corresponding to the set. This distribution p^* can be considered as the "true" conditional prior corresponding to the current candidate label set. Formally, we have the following proposition:

Proposition. *(Content prior). Let* $p^*(\boldsymbol{z}_c|\boldsymbol{u})$ *denote the ground-truth content prior. Then,* $p^*(\boldsymbol{z}_c|\boldsymbol{u})$ *minimizes the KL divergence* $\mathrm{KL}(p^*(\boldsymbol{z}_c|\boldsymbol{u}) || p(\boldsymbol{z}_c|\boldsymbol{u}))$.

Prior $p^*(\boldsymbol{z}_c|\boldsymbol{u})$ is Gaussian which belongs exponential family, thus identifiability conditions could be satisfied. Moreover, we have the following theorem, which could make optimization more feasible.

Theorem 1. *Suppose we have $p(\boldsymbol{z}_c|\boldsymbol{u})$ as a Gaussian mixture distribution:*

$$p(\boldsymbol{z}_c|\boldsymbol{u}) = \sum_{k=1}^{K} u_k \cdot \varphi(\boldsymbol{z}_c; \boldsymbol{A}_k, \boldsymbol{B}_k), \tag{6}$$

where $\varphi(\boldsymbol{z}_c; \boldsymbol{A}_k, \boldsymbol{B}_k)$ is the density of mixing component $\mathcal{N}(\boldsymbol{A}_k, \boldsymbol{B}_k)$. And we use

$$p^*(\boldsymbol{z}_c|\boldsymbol{u}) = \varphi(\boldsymbol{z}_c; \boldsymbol{\mu}_*, \boldsymbol{\sigma}_*^2\boldsymbol{I}), \tag{7}$$

to denote the distribution which minimizes the KL divergence $\mathrm{KL}(p^*(\boldsymbol{z}_c|\boldsymbol{u})||p(\boldsymbol{z}_c|\boldsymbol{u}))$. Then, minimizing $\mathrm{KL}(q(\boldsymbol{z}_c|\boldsymbol{x})||p(\boldsymbol{z}_c|\boldsymbol{u}))$ is equivalent to minimizing $\mathrm{KL}(q(\boldsymbol{z}_c|\boldsymbol{x})||p^*(\boldsymbol{z}_c|\boldsymbol{u}))$.

Although the above discussion addresses the latent identifiability in PLL by effectively leveraging the weakly supervised information provided in the candidate label set, another unique hurdle exists: not all information contained in the latent factors is necessary for weakly supervised classification. To see this, consider partitioning the latent factors into the ones that capture content $\boldsymbol{z}_c$ and style information $\boldsymbol{z}_e$, respectively. On the one hand, the content latent factors capture the core characteristics of each class shared across all instances. On the other hand, the style latent factors correspond to information not causally related to class, e.g., background and lighting variations that are inconsistent across different instances of the same class. Formally,

Assumption 1. *(Content-invariance). For $\boldsymbol{x}, \tilde{\boldsymbol{x}} \in \mathcal{X}$ and $y = \tilde{y}$, the conditional density of the latents $p_{\tilde{\boldsymbol{z}}|\boldsymbol{z}}$ satisfies:*

$$p_{\tilde{\boldsymbol{z}}|\boldsymbol{z}}(\tilde{\boldsymbol{z}}|\boldsymbol{z}) = \delta(\tilde{\boldsymbol{z}}_c - \boldsymbol{z}_c) p_{\tilde{\boldsymbol{z}}_{\boldsymbol{e}}|\boldsymbol{z}_e}(\tilde{\boldsymbol{z}}_e|\boldsymbol{z}_e),$$

where $\boldsymbol{z} = (\boldsymbol{z}_c, \boldsymbol{z}_e)$, and δ is the Dirac delta function. In other words, $\tilde{\boldsymbol{z}}_c = \boldsymbol{z}_{\boldsymbol{c}}$ almost everywhere.

Assumption 2. *(Style-variation). Let $\mathcal{A}$ be a set containing subsets of styles $A \subseteq \{1, \cdots, n_s\}$ and let p_A be a probability distribution on $\mathcal{A}$. The style conditional distributions should satisfy:*

$$A \sim p_A, p_A(A) > 0 \textit{ for all } A \subseteq \{1, \cdots, n_s\} \textit{ and } A \neq \emptyset,$$

$$p_{\tilde{\boldsymbol{z}}_e|\boldsymbol{z}_e, A}(\tilde{\boldsymbol{z}}_e|\boldsymbol{z}_e, A) = \delta(\tilde{\boldsymbol{z}}^e_{A^c} - \boldsymbol{z}^e_{A^c}) p_{\tilde{z}^e_A|\boldsymbol{z}^e_A}(\tilde{z}^e_A|\boldsymbol{z}^e_A).$$

Loosely speaking, the first assumption asserts that the content within each category should remain constant, while the second assumption specifies that certain style factors should change. Importantly, the second assumption is flexible as it does not require all styles to change.

However, VAEs in previous work uniformly extract all factors possibly needed for reconstruction. The vanilla VAE's KL divergence requires the posterior distribution to fit a standard normal distribution as closely as possible, inevitably hindering clear separations between different categories. This also explains why using features extracted from VAEs for classification often yields dissatisfactory results in practical applications. In contrast, models like iVAE incorporate learnable priors during training, adjusting the KL divergence to bring the posterior distribution closer to the learned conditional priors. However, such settings default all latent factors to impact sample classification, indiscriminately utilizing non-causal features in the data, thereby inevitably suffering severe damage during distribution shift.

The above two cases not only affect the performance and robustness of the model, but also deviate from the original intention of causal representation learning. Summarizing the above two cases, we can see that if you want to acquire high-quality features, style and content may be handled differently. Therefore, an idea of this paper is born. In the following methods, we will divide the latent embedding into two parts and treat them separately according to their characteristics. Specifically, we divide latent code $\boldsymbol{z}$ into $\boldsymbol{z}_c$ and $\boldsymbol{z}_e$, i.e. $\boldsymbol{z} = (\boldsymbol{z}_c^T, \boldsymbol{z}_e^T)^T$. Where $\boldsymbol{z}_c$ follows a conditional prior $p(\boldsymbol{z}_c|\boldsymbol{u})$ regulated by the auxiliary variable $\boldsymbol{u}$, and the prior of $\boldsymbol{z}_e$ is a standard normal distribution. At the same time, because the components of $\boldsymbol{z}$ are independent of each other, $\boldsymbol{z}$ still obeys an exponential family distribution as a whole.

3.2 Overall Framework

Figure 1 provides a concise overview of the model's structure. It consists primarily of five components: the encoder $q(\boldsymbol{z}|\boldsymbol{x})$, the decoder $p(\boldsymbol{x}|\boldsymbol{z})$, the prior network $p(\boldsymbol{z}_c|\boldsymbol{u})$, the classifier $q(\boldsymbol{y}|\boldsymbol{z}_c)$, and the contrastive learning module. $\boldsymbol{x}$ and the auxiliary variable $\boldsymbol{u}$ are fed into the encoder and prior network, respectively. Subsequently, we sample from the posterior distribution using reparameterization to obtain the latent code $\boldsymbol{z}$. As mentioned earlier, the latent embedding $\boldsymbol{z}$ can be divided into two parts: the content embedding $\boldsymbol{z}_c$, which encodes category-related information whose prior following conditional distribution $p(\boldsymbol{z}_c|\boldsymbol{u})$, and the style representation $\boldsymbol{z}_e$, independent of class, with its prior $p(\boldsymbol{z}_e)$ following a standard normal distribution $\mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$.

Since the content and style components of $\boldsymbol{z}$ are independent, the KL divergence can be expressed as:

$$\mathrm{KL}(q(\boldsymbol{z}_c|\boldsymbol{x})\|p(\boldsymbol{z}_c|\boldsymbol{u})) + \mathrm{KL}(q(\boldsymbol{z}_e|\boldsymbol{x})\|p(\boldsymbol{z}_e)), \tag{8}$$

where parameters for the conditional distribution $p(\boldsymbol{z}_c|\boldsymbol{u})$ are generated by the prior network. In practice, the prior network is implemented as a single-layer linear mapping similar to word embedding. This not only avoids issues of pattern collapse but also enhances the model's interpretability. It is worth noting that instead of directly using the candidate label set $\boldsymbol{s}$ as the auxiliary information, CausalPLL+ integrates the learning of latent representations with the refinement of the candidate label set. This approach addresses two problems inherent in IDPLL. Firstly, discerning the true class label from the candidate label set is naturally integrated with the inference of the representations. Secondly, refining the candidate label set further improves the auxiliary information for inferencing the representation. A more detailed elaboration of the integrated label refinement process is discussed in Section 3.3.

The evidence lower bound (ELBO) of the model can be

Table 1: Accuracy (mean±std) comparisons on FashionMNIST, Kuzushiji-MNIST, SVHN and CIFAR10 with instance-dependent partial labels on different ambiguity levels.

Dataset	Method	$\tau = 16$	$\tau = 32$	$\tau = 64$
FashionMNIST	CausalPLL+	**94.49** $\pm$ **0.37**%	**93.60** $\pm$ **0.10**%	**92.75** $\pm$ **0.18**%
	PLCR	93.28 $\pm$ 0.24%	92.46 $\pm$ 0.13%	90.72 $\pm$ 0.15%
	VALEN	88.36 $\pm$ 0.20%	87.25 $\pm$ 0.19%	85.67 $\pm$ 0.24%
	LWS	88.50 $\pm$ 0.19%	84.84 $\pm$ 0.51%	81.23 $\pm$ 2.07%
	PRODEN	87.32 $\pm$ 0.19%	86.34 $\pm$ 0.08%	85.15 $\pm$ 0.24%
	RC	89.56 $\pm$ 0.18%	89.05 $\pm$ 0.12%	87.65 $\pm$ 0.10%
	CC	89.31 $\pm$ 0.07%	88.46 $\pm$ 0.03%	87.11 $\pm$ 0.11%
	Fully Supervised		95.54 $\pm$ 0.07%	
KMNIST	CausalPLL+	**98.49** $\pm$ **0.08**%	**97.89** $\pm$ **0.14**%	**96.96** $\pm$ **0.10**%
	PLCR	97.84 $\pm$ 0.04%	96.03 $\pm$ 0.60%	91.43 $\pm$ 0.58%
	VALEN	86.08 $\pm$ 0.37%	82.23 $\pm$ 0.36%	77.18 $\pm$ 0.56%
	LWS	88.94 $\pm$ 0.17%	86.37 $\pm$ 0.89%	83.16 $\pm$ 0.46%
	PRODEN	88.50 $\pm$ 0.24%	86.27 $\pm$ 0.33%	82.92 $\pm$ 0.45%
	RC	91.41 $\pm$ 0.07%	89.63 $\pm$ 0.06%	87.15 $\pm$ 0.11%
	CC	91.77 $\pm$ 0.08%	89.81 $\pm$ 0.12%	86.40 $\pm$ 0.15%
	Fully Supervised		99.03 $\pm$ 0.04%	
SVHN	CausalPLL+	**97.50** $\pm$ **0.21**%	**97.05** $\pm$ **0.37**%	**96.56** $\pm$ **0.26**%
	PLCR	97.15 $\pm$ 0.09%	96.59 $\pm$ 0.15%	95.97 $\pm$ 0.18%
	VALEN	96.58 $\pm$ 0.20%	96.02 $\pm$ 0.39%	95.27 $\pm$ 0.37%
	LWS	96.24 $\pm$ 0.08%	95.87 $\pm$ 0.09%	94.79 $\pm$ 0.18%
	PRODEN	96.18 $\pm$ 0.17%	95.31 $\pm$ 0.22%	94.83 $\pm$ 0.25%
	RC	95.68 $\pm$ 0.24%	95.38 $\pm$ 0.13%	94.77 $\pm$ 0.16%
	CC	95.39 $\pm$ 0.26%	94.75 $\pm$ 0.47%	93.58 $\pm$ 0.38%
	Fully Supervised		98.09 $\pm$ 0.06%	
CIFAR10	CausalPLL+	95.91 $\pm$ 0.28%	**94.04** $\pm$ **0.26**%	**89.66** $\pm$ **0.32**%
	PLCR	**96.28** $\pm$ **0.09**%	93.97 $\pm$ 0.07%	88.82 $\pm$ 0.11%
	VALEN	89.63 $\pm$ 0.34%	86.35 $\pm$ 0.32%	78.28 $\pm$ 0.41%
	LWS	85.38 $\pm$ 0.21%	81.47 $\pm$ 0.20%	74.10 $\pm$ 0.25%
	PRODEN	93.84 $\pm$ 0.48%	90.07 $\pm$ 0.49%	86.36 $\pm$ 0.53%
	RC	86.33 $\pm$ 0.11%	81.19 $\pm$ 0.11%	74.93 $\pm$ 0.21%
	CC	86.26 $\pm$ 0.10%	82.73 $\pm$ 0.23%	76.48 $\pm$ 0.12%
	Fully Supervised		97.67 $\pm$ 0.13%	

expressed as:

$$\begin{aligned}\mathcal{L}_{\text{ELBO}} &= \mathbb{E}_{\boldsymbol{z}\sim q(\boldsymbol{z}|\boldsymbol{x})}\left[\ln p(\boldsymbol{x}|\boldsymbol{z})\right] \\ &-\text{KL}(q(\boldsymbol{z}_c|\boldsymbol{x})\|p(\boldsymbol{z}_c|\boldsymbol{u}=\hat{\boldsymbol{s}})) \\ &-\text{KL}(q(\boldsymbol{z}_e|\boldsymbol{x})\|p(\boldsymbol{z}_e)).\end{aligned} \tag{9}$$

As the only exact supervision information in PLL is that non-candidate labels are not ground truth, we utilize an "only negatives matter" loss function on the the content representation $\boldsymbol{z}_c$ for classification. In this part, we also use the refined candidate vectors $\hat{\boldsymbol{s}}$. Specifically,

$$\mathcal{L}_{\text{err}} = \sum_{k=1}^{K}(1-\hat{s}_k)\cdot\ln(1-h_k(\boldsymbol{z}_c)). \tag{10}$$

While performing reconstruction and classification, we also introduce a novel contrastive learning module based on the latent space. Specifically:

$$\mathcal{L}_{\text{CL}} = \frac{-1}{|\mathcal{S}|}\sum_{i\in\mathcal{S}}\ln\frac{\exp(\boldsymbol{z}\cdot\tilde{\boldsymbol{\mu}}_i)}{\sum_{j\in\mathcal{Y}}\exp(\boldsymbol{z}\cdot\tilde{\boldsymbol{\mu}}_j)}, \tag{11}$$

where $\tilde{\boldsymbol{\mu}}_i$ is the i-th mean vector in content prior mean matrix $\boldsymbol{A}$. Finally, the loss function of CausalPLL+ is:

$$\mathcal{L} = \lambda_{\text{ELBO}}\cdot\mathcal{L}_{\text{ELBO}} + \lambda_{\text{CL}}\cdot\mathcal{L}_{\text{CL}} + \mathcal{L}_{\text{err}}. \tag{12}$$

3.3 Candidate Label Refinement

Traditional PLL disambiguation methods mostly solely utilize the discriminative information provided by supervision. However, by using variational generative models, we can also reconstruct the data generation process. In the CausalPLL+ framework, more precise auxiliary information enables the model to learn better priors, and these improved priors, in turn, help the model achieve more accurate classification, which leads to even more veracious auxiliary information. This iterative process ensures that by the end of training, the model not only effectively models the generative process and priors but also achieves outstanding discriminative performance. Therefore, we proposed a candidate label refinement strategy to gradually eliminate labels that are more likely to be wrong. Specifically, for each sample, we maintain a vector and perform momentum updates using the unnormalized prediction scores from the classifier.

Table 2: Accuracy (mean±std) comparisons on MNIST→MNIST-M, MNIST→SVHN, SVHN→USPS with instance-dependent partial labels on different ambiguity levels.

Dataset	Method	$\tau = 16$	$\tau = 32$	$\tau = 64$
MNIST→MNIST-M	CausalPLL+	97.85 ± 0.11%	**96.58 ± 0.10**%	**94.67 ± 0.18**%
	PLCR	97.63 ± 0.08%	95.34 ± 0.09%	94.59 ± 0.11%
	PiCO	**98.64 ± 0.07**%	78.63 ± 1.60%	57.52 ± 4.17%
	LWS	96.93 ± 0.09%	95.61 ± 0.15%	92.21 ± 0.18%
	RC	96.77 ± 0.10%	96.47 ± 0.10%	93.59 ± 0.09%
	CC	97.08 ± 0.05%	96.15 ± 0.10%	94.45 ± 0.07%
MNIST→SVHN	CausalPLL+	94.13 ± 0.15%	**93.11 ± 0.12**%	**91.04 ± 0.19**%
	PLCR	93.95 ± 0.10%	92.50 ± 0.09%	87.39 ± 0.18%
	PiCO	**95.57 ± 0.08**%	84.67 ± 0.15%	62.30 ± 0.21%
	LWS	93.80 ± 0.12%	67.52 ± 0.71%	42.18 ± 1.66%
	RC	94.16 ± 0.08%	91.75 ± 0.10%	83.65 ± 0.15%
	CC	93.81 ± 0.14%	91.81 ± 0.15%	86.39 ± 0.29%
USPS→SVHN	CausalPLL+	**86.72 ± 0.16**%	**81.65 ± 0.17**%	**73.15 ± 0.25**%
	PLCR	83.70 ± 0.17%	77.32 ± 0.23%	69.23 ± 0.27%
	PiCO	80.17 ± 0.18%	56.24 ± 1.14%	36.61 ± 7.23%
	LWS	78.25 ± 0.15%	28.05 ± 4.78%	26.68 ± 5.25%
	RC	71.17 ± 0.20%	53.34 ± 1.28%	43.92 ± 3.16%
	CC	80.23 ± 0.06%	56.97 ± 1.34%	43.99 ± 1.27%

This process can be expressed as:

$$\boldsymbol{\gamma}_{t+1} = (1 - m) \cdot \boldsymbol{\gamma}_t + m \cdot \hat{\boldsymbol{y}}, \tag{13}$$

where m is the momentum factor and $\boldsymbol{\gamma}$ is the average of past model predictions. The refined candidate vector could be expressed as:

$$\hat{\boldsymbol{s}} = \mathrm{softmax}(\frac{\boldsymbol{\gamma} - \boldsymbol{s} \cdot \text{int\_max}}{T}), \tag{14}$$

where T is the temperature. With this mechanism, we can progressively eliminate the least scoring classes from the current candidate label set.

4 Experiments

The experiments in this paper are primarily divided into three parts. Section 4.2 focuses on the model's classification performance in IDPLL tasks. Section 4.3 investigates the model's generalization ability in the presence of style variation. Finally, in Section 4.4, we examine the nature of representations extracted by CausalPLL+ and observe the different impacts of content embeddings and style embeddings on image generation.

4.1 Experiment Setup

Datasets For IDPLL classification tasks, experiments were conducted on four well-known benchmarks: Fashion-MNIST (Xiao, Rasul, and Vollgraf 2017), Kuzushiji-MNIST (Clanuwat et al. 2018), SVHN (Netzer et al. 2011), and CIFAR10 (Krizhevsky, Hinton et al. 2009).

Regarding domain generalization issues, we utilized three sets of classic datasets in this domain, mixing them at different ratios in training and testing sets. These three pairs include MNIST→MNIST-M, MNIST→SVHN, and SVHN→USPS. Among them, the MNIST-M dataset (Ganin and Lempitsky 2015) is obtained by blending digits from the original set over patches randomly extracted from colour photos from BSDS500 (Arbelaez et al. 2010). The mixing ratio of these three pairs was 80%-20% in the training set and 20%-80% in the test set. More details on implementation can be found in the supplementary material.

Data Generation Method In previous partial label learning research, it is a common practice to manually corrupt the existing fully-supervised datasets into partially labelled versions. However, existing data generation methods in IDPLL may suffer from an underconfidence problem, causing synthetic data to diverge from real-world situations. For example, for a sample $\boldsymbol{y} = (0, 0, 1)^T$ which has a very confident prediction $\hat{\boldsymbol{y}} = (.01, .01, .98)^T$, the corresponding $\boldsymbol{s}$ would be $(1, 1, 1)^T$, which is very unconfident. And the contradiction appeared. Moreover, current research in IDPLL lacks an approach similar to those in classical PLL that adjusts the level of ambiguity in weakly supervised data. Therefore, we proposed a novel data generation method that closely approximates real-world scenarios while allowing control over the ambiguity of supervised information. In brief, τ represents the level of ambiguity in the candidate label set, where a higher τ indicates greater ambiguity. Details of this data generation mechanism will be provided in the appendix.

4.2 IDPLL Classification

In this section, we evaluated the classification performance of CausalPLL+ across varying levels of ambiguity in IDPLL tasks. As shown in Table 1, CausalPLL+ achieved superior performance across most levels of ambiguity on three benchmarks. Moreover, its performance notably outperformed other baseline models in situations with higher ambiguity levels. The reason for this is that when the degree of ambiguity is large, the model would be seriously

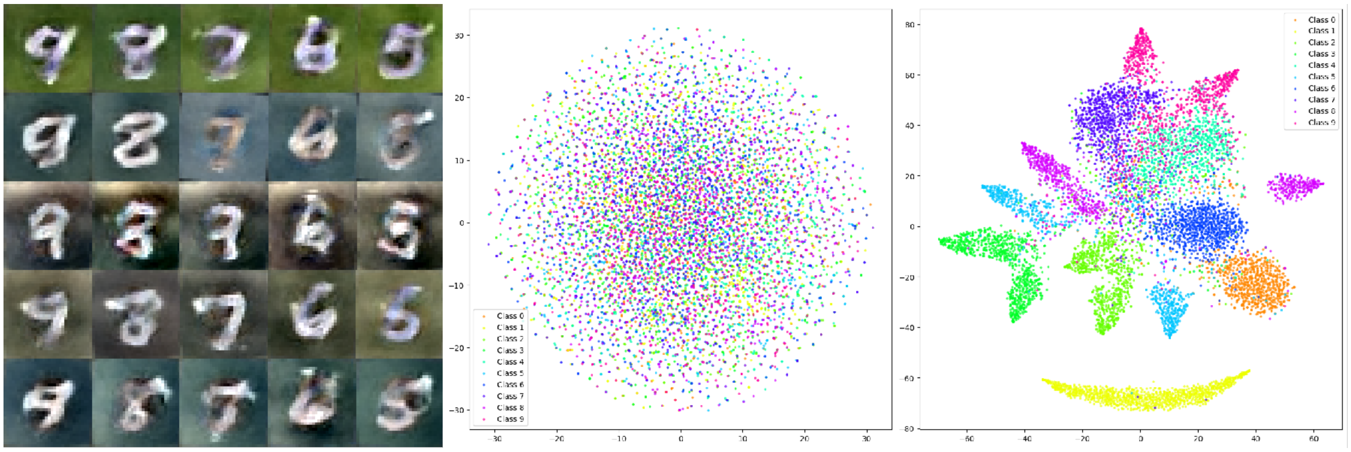

Figure 2: (Left) Controlled image generation by sampling from the latent space. Images in one row share the same style but have different contents. (Middle and Right) The t-SNE visualization for the latent space of *style* versus *content*.

disturbed by too many candidate labels. Therefore, the label refinement mechanism can eliminate more false candidates, thus contributing to the performance improvement of the model. The experiments demonstrate that CausalPLL+ excels not only in conventional IDPLL classification tasks but also highlights its effectiveness and superiority, suggesting its versatility and broader applicability as an algorithm.

4.3 Domain Shift

We now study the model's performance under variations of styles and domain shifts. We selected three pairs of datasets, exhibiting an increasing level of domain shift. MNIST→MNIST-M involves mild changes in background and color, while MNIST→SVHN and USPS→SVHN introduce significant variations in camera angles, digit styles, and background complexities. As the changes in distributions intensify, the results show that these domain shifts have a substantial impact on the model's performance, with more severe shifts leading to a notable decline in accuracy.

Across all three benchmarks, CausalPLL+ outperformed the baselines in most instances. The performance degradation of the compared baselines is most noticeable on USPS→SVHN, as most perform worse than a random guess. This demonstrates that the method's representation decoupling mechanism effectively mitigates the impact of domain shifts, enhancing the model's robustness against distribution shifts.

4.4 Quantitative Results and Visualization

In this section, we observe the impacts of z_c and z_e on image generation, while also studying their distinct properties in the representation space.

Figure 2 (Left) showcases the model's generation results on the MNIST→MNIST-M dataset. These images are not reconstructions of real samples but direct samples from the latent space. Specifically, each row in the figure represents different z_c values. We obtain mean and variance parameters for five classes from the prior network, reparameterizing to derive five distinct z_c values. Meanwhile, each column represents different z_e, sampled directly from a standard normal distribution. From these generated images, it's evident that z_c primarily influences image categories, while z_e affects style elements such as color and background, with less impact on the image's content or category. The distinct roles of z_c and z_e validate our method's effective content-style decoupling. This decoupling not only enhances the model's robustness against style variations and domain shifts but also demonstrates potential for controlled image generation. Figure 2 (Middle and Right) is the t-SNE visualization for the latent space of style and content. The samples in the right figure exhibit a clear separation while those in the middle figure are completely mixed, which confirms that the content embeddings could effectively capture the class-specific features while the style embeddings successfully maintain class-irrelevant.

5 Conclusion and Discussion

In this paper, we investigate latent representation identifiability within the PLL paradigm and propose a novel framework, CausalPLL+, that addresses challenges in IDPLL classification, as well as domain shift and style variation problems that have plagued related algorithms. We introduce a novel prior network that enhances model interpretability without compromising performance, bridging the gap between identifiability theory and practical PLL applications. Furthermore, we bifurcated the latent embedding into two branches, explicitly decoupling content from style. This enhancement equips the model with greater robustness against style variations and domain shifts. Additionally, we proposed a contrastive learning approach under the PLL paradigm to effectively leverage the learned prior from the model. Lastly, we introduced a label refinement disambiguation strategy that reduces vagueness in supervision by progressively eliminating erroneous labels. This method is particularly effective when dealing with highly ambiguous candidate label sets. Extensive empirical studies confirm the effectiveness of the proposed method.

References

Arbelaez, P.; Maire, M.; Fowlkes, C.; and Malik, J. 2010. Contour detection and hierarchical image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 33(5): 898–916.

Brehmer, J.; De Haan, P.; Lippe, P.; and Cohen, T. S. 2022. Weakly Supervised Causal Representation Learning. *Advances in Neural Information Processing Systems*, 35: 38319–38331.

Chen, C.-H.; Patel, V. M.; and Chellappa, R. 2017. Learning from ambiguously labeled face images. *IEEE transactions on pattern analysis and machine intelligence*, 40(7): 1653–1667.

Clanuwat, T.; Bober-Irizar, M.; Kitamoto, A.; Lamb, A.; Yamamoto, K.; and Ha, D. 2018. Deep learning for classical japanese literature. *arXiv preprint arXiv:1812.01718*.

Cour, T.; Sapp, B.; Jordan, C.; and Taskar, B. 2009. Learning from ambiguously labeled images. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 919–926. IEEE.

Feng, L.; Lv, J.; Han, B.; Xu, M.; Niu, G.; Geng, X.; An, B.; and Sugiyama, M. 2020. Provably Consistent Partial-Label Learning. In Larochelle, H.; Ranzato, M.; Hadsell, R.; Balcan, M.; and Lin, H., eds., *Advances in Neural Information Processing Systems*, volume 33, 10948–10960. Curran Associates, Inc.

Ganin, Y.; and Lempitsky, V. 2015. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*, 1180–1189. PMLR.

Khemakhem, I.; Kingma, D.; Monti, R.; and Hyvarinen, A. 2020. Variational autoencoders and nonlinear ica: A unifying framework. In *International Conference on Artificial Intelligence and Statistics*, 2207–2217. PMLR.

Kingma, D. P.; and Welling, M. 2013. Auto-Encoding Variational Bayes. *arXiv preprint arXiv:1312.6114*.

Kivva, B.; Rajendran, G.; Ravikumar, P.; and Aragam, B. 2022. Identifiability of Deep Generative Models without Auxiliary Information. *Advances in Neural Information Processing Systems*, 35: 15687–15701.

Krizhevsky, A.; Hinton, G.; et al. 2009. Learning multiple layers of features from tiny images.

Lin, Y.; Yao, Y.; Shi, X.; Gong, M.; Shen, X.; Xu, D.; and Liu, T. 2024. Cs-isolate: Extracting hard confident examples by content and style isolation. *Advances in Neural Information Processing Systems*, 36.

Liu, Y.; Wei, Y.-S.; Yan, H.; Li, G.-B.; and Lin, L. 2022. Causal reasoning meets visual representation learning: A prospective study. *Machine Intelligence Research*, 19(6): 485–511.

Locatello, F.; Bauer, S.; Lucic, M.; Raetsch, G.; Gelly, S.; Schölkopf, B.; and Bachem, O. 2019. Challenging common assumptions in the unsupervised learning of disentangled representations. In *international conference on machine learning*, 4114–4124. PMLR.

Luo, J.; and Orabona, F. 2010. Learning from candidate labeling sets. *Advances in neural information processing systems*, 23.

Lv, J.; Xu, M.; Feng, L.; Niu, G.; Geng, X.; and Sugiyama, M. 2020. Progressive Identification of True Labels for Partial-Label Learning. In *International Conference on Machine Learning*, 6500–6510. PMLR.

Netzer, Y.; Wang, T.; Coates, A.; Bissacco, A.; Wu, B.; Ng, A. Y.; et al. 2011. Reading digits in natural images with unsupervised feature learning. In *NIPS workshop on deep learning and unsupervised feature learning*, volume 2011, 4. Granada.

Qiao, C.; Xu, N.; and Geng, X. 2022. Decompositional generation process for instance-dependent partial label learning. *arXiv preprint arXiv:2204.03845*.

Scheffer, T.; Decomain, C.; and Wrobel, S. 2001. Mining the web with active hidden Markov models. In *Proceedings 2001 IEEE International Conference on Data Mining*, 645–646. IEEE.

Schölkopf, B.; Locatello, F.; Bauer, S.; Ke, N. R.; Kalchbrenner, N.; Goyal, A.; and Bengio, Y. 2021. Toward Causal Representation Learning. *Proceedings of the IEEE*, 109(5): 612–634.

Tang, W.; Yang, Y.-F.; Wang, Z.; Zhang, W.; and Zhang, M.-L. 2024. Multi-Instance Partial-Label Learning with Margin Adjustment. In *Advances in Neural Information Processing Systems 37, Vancouver, Canada*.

Tang, W.; Zhang, W.; and Zhang, M.-L. 2023. Disambiguated Attention Embedding for Multi-Instance Partial-Label Learning. 56756–56771.

Tang, W.; Zhang, W.; and Zhang, M.-L. 2024a. Exploiting Conjugate Label Information for Multi-Instance Partial-Label Learning. In *Proceedings of the 33rd International Joint Conference on Artificial Intelligence, Jeju, South Korea*, 4973–4981.

Tang, W.; Zhang, W.; and Zhang, M.-L. 2024b. Multi-Instance Partial-Label Learning: Towards Exploiting Dual Inexact Supervision. *Science China Information Sciences*, 67(3): 132103:1–132103:14.

Wen, H.; Cui, J.; Hang, H.; Liu, J.; Wang, Y.; and Lin, Z. 2021. Leveraged weighted loss for partial label learning. In *International conference on machine learning*, 11091–11100. PMLR.

Wu, D.-D.; Wang, D.-B.; and Zhang, M.-L. 2022. Revisiting Consistency Regularization for Deep Partial Label Learning. In *International Conference on Machine Learning*, 24212–24225. PMLR.

Wu, D.-D.; Wang, D.-B.; and Zhang, M.-L. 2024. Distilling Reliable Knowledge for Instance-Dependent Partial Label Learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 15888–15896.

Xia, S.; Lv, J.; Xu, N.; and Geng, X. 2022. Ambiguity-Induced Contrastive Learning for Instance-Dependent Partial Label Learning. In *IJCAI*, 3615–3621.

Xiao, H.; Rasul, K.; and Vollgraf, R. 2017. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*.

Xu, N.; Liu, B.; Lv, J.; Qiao, C.; and Geng, X. 2023. Progressive purification for instance-dependent partial label learning. In *International Conference on Machine Learning*, 38551–38565. PMLR.

Xu, N.; Qiao, C.; Geng, X.; and Zhang, M.-L. 2021. Instance-dependent partial label learning. *Advances in Neural Information Processing Systems*, 34: 27119–27130.

Yang, Y.-F.; Tang, W.; and Zhang, M.-L. 2024. ProMIPL: A Probabilistic Generative Model for Multi-Instance Partial-Label Learning. In *Proceedings of the 24th IEEE International Conference on Data Mining, Abu Dhabi, UAE*, 1–10.

Yao, Y.; Liu, T.; Gong, M.; Han, B.; Niu, G.; and Zhang, K. 2021. Instance-Dependent Label-Noise Learning under a Structural Causal Model. *Advances in Neural Information Processing Systems*, 34: 4409–4420.

Zeng, Z.; Xiao, S.; Jia, K.; Chan, T.-H.; Gao, S.; Xu, D.; and Ma, Y. 2013. Learning by associating ambiguously labeled images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 708–715.

Zhang, W.; Zhang, X.; Hanwen, D.; and Zhang, M.-L. 2022. Multi-Instance Causal Representation Learning for Instance Label Prediction and out-of-Distribution Generalization. *Advances in Neural Information Processing Systems*, 35: 34940–34953.

Zhou, Z.-H. 2017. A brief introduction to weakly supervised learning. *National Science Review*, 5(1): 44–53.

Zhu, Z.; Tian, H.; Chen, X.; Zhang, K.; and Yu, Y. 2025. Offline model-based reinforcement learning with causal structured world models. *Frontiers of Computer Science*, 19(4): 194347.