CMVAE: Causal Meta VAE for Unsupervised Meta-Learning

Guodong Qi [1,2], **Huimin Yu** [1,2,3,4]

[1]College of Information Science and Electronic Engineering, Zhejiang University
[2]ZJU-League Research & Development Center
[3]State Key Lab of CAD&CG, Zhejiang University
[4]Zhejiang Provincial Key Laboratory of Information Processing, Communication and Networking
{guodong_qi, yhm2005}@zju.edu.cn

Abstract

Unsupervised meta-learning aims to learn the meta knowledge from unlabeled data and rapidly adapt to novel tasks. However, existing approaches may be misled by the context-bias (e.g. background) from the training data. In this paper, we abstract the unsupervised meta-learning problem into a Structural Causal Model (SCM) and point out that such bias arises due to hidden confounders. To eliminate the confounders, we define the priors are *conditionally* independent, learn the relationships between priors and intervene on them with casual factorization. Furthermore, we propose Causal Meta VAE (CMVAE) that encodes the priors into latent codes in the causal space and learns their relationships simultaneously to achieve the downstream few-shot image classification task. Results on toy datasets and three benchmark datasets demonstrate that our method can remove the context-bias and it outperforms other state-of-the-art unsupervised meta-learning algorithms because of bias-removal. Code is available at https://github.com/GuodongQi/CMVAE.

1 Introduction

Regular meta-learning algorithms such as (Finn, Abbeel, and Levine 2017; Snell, Swersky, and Zemel 2017) aim to learn the meta knowledge to adapt to novel tasks quickly. However, it requires various supervised tasks on large labeled datasets during the meta-training phase. Recently, researchers take great interest in *unsupervised meta-learning* (Hsu, Levine, and Finn 2019; Khodadadeh et al. 2021). Different from regular meta-learning, unsupervised meta-learning contains unsupervised meta-training and supervised meta-test. It aims to learn a learning procedure with unlabeled datasets in the meta-training and solve novel supervised human-crafted tasks in the meta-test.

Previous methods focus on the pseudo-label generation of the task. However, they may ignore the bias. Figure 1a illustrates a binary-classification toy example where the background prior is one of bias. In the training images, the "birds" are always together with the "sky" and the "airplanes" always park on the ground. As a result, the model will take the "sky" as a part of the "bird", and mistakenly recognize the "airplane" test image as a "bird". It is essential to remove the effect of background prior *i.e.*, context-bias.

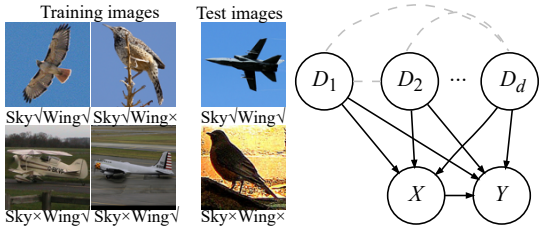


Figure 1: (a) Illustration of context-bias. (b) SCM of unsupervised meta-learning. The dashed line means that the relationship (DAG) need to be learned.

However, discerning the context-bias is challenging, because the priors may not be independent. For example, in the task of Figure 1a, the "wing" and the "sky" prior is not independent statistically[1]. When the "sky" prior is removed, the "wing" prior will be changed, and then the prediction will be affected. In this case, the model will not know whether the "sky" or "wing" prior is the context-bias.

To address the problems, we analyze, discern and remove the context-bias from a causal perspective via three theories, *i.e.*, Structural Causal Model (SCM) (Glymour, Pearl, and Jewell 2016), Common Cause Principle (CCP) (Schölkopf et al. 2021) and Independent Causal Mechanism (ICM) (Schölkopf et al. 2012). Among them, SCM describes the relevant concepts and how they interact with each other. CCP reveals that if two observables are statistically dependent, then there exists a variable such that they are independent conditioned on the variable. ICM states that the conditional distribution of each prior given its causes does not influence the others. In other words, SCM explains how the bias affects predictions. CCP makes it reasonable to assume the priors are *conditionally* independent. For example, in Figure 1a there exists a "flying" prior, which causally affects "sky" and "wing" and makes them independent when conditioned on the prior. ICM allows us to remove one prior (*e.g.*, $p(\text{sky}|\text{flying})$) will not affect another prior (*e.g.*, $p(\text{wing}|\text{flying})$).

Specially, we build the SCM in Figure 1b. The bias

[1]$P(\text{wing}, \text{sky}) = 1/4$, $P(\text{wing}) = 3/4$, $P(\text{sky}) = 1/2$, we have $P(\text{wing}, \text{sky}) \neq P(\text{wing})P(\text{sky})$, so they are dependent.

emerges because the priors are confounders that cause spurious correlations from the inputs to predictions. To achieve bias-removal, we define the relationships between priors based on CCP, obtain the structure with a learnable directed acyclic graph (DAG), causally factorize the joint distribution of priors based on ICM, and then perform causal intervention (Glymour, Pearl, and Jewell 2016) in sequence.

Furthermore, we design the Causal Meta VAE (CMVAE), which learns the priors and the causal factorization simultaneously. Particularly, we propose the causal intervention formula with the SCM. It leads us to learn the conditionally independent latent codes (priors) as well as the DAG (causal factorization). To make the correspondence between the latent codes and priors, we adopt the VAE-based framework (Kingma and Welling 2014) since VAE has been shown to achieve some useful disentangling performance (Higgins et al. 2016). The "DAG-ness" can be quantified by a regularizer (Zheng et al. 2018). Besides, we introduce the Causal Latent Space (CaLS) and show its addability, which makes it feasible to represent the class-concept codes while keeping the DAG. We also extend one baseline (Lee et al. 2021) into our CMVAE to achieve the downstream few-shot classification with the unsupervised meta-learning settings. The contributions of this paper are as follows:

- We point out the context-bias and the dependent priors in unsupervised meta-learning. We propose to learn the relationship among the priors with a learnable DAG and make the priors causally independent and factorize.
- We design the intervention formula, introduce the CaLS, and propose CMVAE to learn the factors and the factorization for the downstream classification simultaneously.
- Extensive experiments on two toy datasets and three widely used benchmark datasets demonstrate that CMVAE outperforms other state-of-the-art unsupervised meta-learning algorithms. Furthermore, we show that CMVAE can be intervened to generate counterfactual samples with some meaningful explanation.

2 Related Work

Unsupervised Meta-Learning aims to learn the meta-knowledge with unlabeled training data. CACTU (Hsu, Levine, and Finn 2019) and UMTRA (Khodadadeh, Bölöni, and Shah 2019) try to create synthetic labels. GMVAE (Lee et al. 2021) introduces a Mixture of Gaussian priors by performing Expectation-Maximization (EM). However, none of them notices the bias in the few-shot tasks.

Causal Inference helps machines understand how and why causes influence their effects (Glymour, Pearl, and Jewell 2016). Recently, the connection between causality and machine learning (Magliacane et al. 2018; Bengio et al. 2020; Kyono, Zhang, and van der Schaar 2020) or computer vision (Lopez-Paz et al. 2017; Yang et al. 2021b; Wang et al. 2020) have gained increasing interest. Recently, IFSL (Yue et al. 2020) introduces the causality into few-shot learning problem with an SCM. However, CMVAE differs since it explicitly learns and utilizes the causal factorization.

DAG Learning is to estimate the structure of variables. There are three types of methods, the discrete optimization (Scanagatta et al. 2016; Viinikka et al. 2020), the continuous optimization (Zheng et al. 2018, 2020) and the sampling-based methods (Charpentier, Kibler, and Günnemann 2022). CMVAE incorporates recent continuous optimization methods to learn the DAG of the context-priors.

3 Proposed Formulation

3.1 Problem Statement

Given an unlabeled dataset $\mathcal{U}$ in the meta-training stage, we aim to learn the knowledge which can be adapted to novel tasks in the meta-test stage. Each task $\mathcal{T}$ is drawn from a few-shot labeled dataset $\mathcal{D}$. The $\mathcal{U}$ and $\mathcal{D}$ are drawn from the same distribution but a different set of classes. Specially, a K-way S-shot classification task $\mathcal{T}$ consists of support data $\mathcal{S} = \{(\mathbf{x}_s, \mathbf{y}_s)\}_{s=0}^{KS}$ with K classes of S few labeled samples and query data $\mathcal{Q} = \{\mathbf{x}_q\}_{q=0}^{Q}$ with Q unlabeled samples. Our goal is to predict the labels of $\mathcal{Q}$ given $\mathcal{S}$.

3.2 Causal Insight

Unsupervised meta-learning methods are confused by the context-bias. To analyze how the bias arises, we formulate the problem into SCM in Figure 1b. In the SCM, 1) $D \rightarrow X$ means that the priors D determine where the object appears in an image, *e.g.*, the context-priors in training images of Figure 1a put the bird object in the sky. 2) $D \rightarrow Y$ denotes that the priors D affect the predictions Y, *e.g.*, the wing and sky priors lead to the bird prediction. 3) $D_1, \cdots, D_d$ are dependent statistically, *e.g.*, the "sky", "wing" and prior are not independent but causally dependent. Their causal relationships need to be determined (dashed lines). 4) $X \rightarrow Y$ is the regular classification process.

From the SCM, we observe that context-priors D confound the effect that input X has on prediction Y, which leads to the bias. Thus, it is critical to eliminate the confounding effects, we then apply causal intervention with the do-operator (Glymour, Pearl, and Jewell 2016) as follows (Details in Supp. 3.1),

$$P(\mathbf{y}|do(\mathbf{x})) = \sum_{\mathrm{d}_1,\cdots,\mathrm{d}_d} P(\mathbf{y}|\mathbf{x}, D_1 = \mathrm{d}_1, \cdots, D_d = \mathrm{d}_d) P(D_1 = \mathrm{d}_1, \cdots, D_d = \mathrm{d}_d) \quad (1)$$

where d_i ranges over all values that variables D_i can take.

Equation 1 informs that intervening on $\mathbf{x}$ calls for the joint distribution of D. Note that $D_1, \cdots, D_d$ may be dependent statistically (*i.e.*, $P(D) \neq \Pi_{i=1}^{d} P(D_i)$). Inspired by CCP (Schölkopf et al. 2021), we assume the common causes are ones of priors. Then finding the common causes suggests discovering the causal relationships among the priors. The causal relationships can be represented by a DAG (dashed lines). For example in Figure 1a, the flying prior is the common cause of sky and wing priors, the DAG is "sky $\leftarrow$ flying $\rightarrow$ wing", and the latter two are independent when conditioned on the flying. Furthermore, based on ICM (Schölkopf et al. 2021), the joint distribution $P(D)$ can be factorized into,

$$P(D) = \prod_{i=1}^{d} P(D_i | \mathrm{PA}(i)) \quad (2)$$

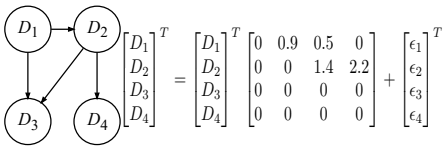


Figure 2: An SEM. [Left]: DAG with 4 nodes. [Right]: A linear equation for Gaussian SEM with noise $\epsilon \sim \mathcal{N}(0, \boldsymbol{I})$.

where $\mathrm{PA}(i)$ denotes the parents of D_i, which can be obtained from the DAG.

To discover the DAG, we utilize Gaussian Structural Equation Model (SEM) (Pearl et al. 2000). Figure 2 shows a linear-Gaussian SEM. Formally, given the variables D, there exist functions h_i and $h_j : \mathbb{R}^d \to \mathbb{R}$ such that

$$D_i = h_i(D) + U_i, \quad D_j = h_j(D) + U_j \tag{3}$$

where U_i and U_j are independent Gaussian noises, and h_i and h_j are regarded as structural functions. The relationship between h_i and $\mathrm{PA}()$ is that $h_i(\mathrm{d}_1, \cdots, \mathrm{d}_d)$ does not depend on d_k if $D_k \notin \mathrm{PA}(i)$.

The DAG can be learned by maximum likelihood estimation $\mathbb{E}[D_i|h_i(D)]$ and $\mathbb{E}[D_j|h_j(D)]$ over D. Its "DAGness" can be enforced using a trace exponential regularizer such as NoTears penalization (Zheng et al. 2018). Insufficient penalization weight may not ensure the "DAGness" and weaken the effect of bias-removal, but default weight works for most scenarios. If the causal graph is Non-DAG graph, a solution is to learn such mixed graphs with score-based methods (Bernstein et al. 2020). It is compatible with our method.

3.3 Adjustment Formulation

This section offer an adjustment formulation for Equation 1. Specially, given the DAG function $h = \{h_i\}_{i=1}^d$, the distribution $P(D)$ is approximated by $P(D_1 = \mathrm{d}_1, \cdots, D_d = \mathrm{d}_d) \approx P(D = \mathbf{d}|h(D = \mathbf{d}))$, where $\mathbf{d} = [\mathrm{d}_1|\cdots|\mathrm{d}_d] \in \mathbb{R}^{1\times d}$. Also, given the input data $\mathbf{x}$, we assume its latent codes $\mathbf{z} \in \mathbb{R}^{1\times d}$ via VAE (Kingma and Welling 2014). Since VAE has been shown to achieve some useful disentangling (Higgins et al. 2016), we perform the dimensional-wise product to make each latent code represents one prior, *i.e.*, $\mathbf{z} \leftarrow \mathbf{z} \otimes \mathrm{d}$. Then we have $P(D = \mathbf{d}|h(D = \mathbf{d})) = P(Z = \mathbf{z}|h(Z = \mathbf{z}))$. Finally, the adjustment formulation yields,

$$p(\mathbf{y}|do(\mathbf{x})) = \mathbb{E}_{\underbrace{p(\mathbf{z}|\mathbf{x})}_{\text{Sampling}}} \mathbb{E}_{\underbrace{p(\mathbf{z}|h(\mathbf{z}))}_{\text{Adjusting}}} p(\mathbf{y}|\mathbf{z}) \tag{4}$$

Equation 4 reveals that the causal intervention can be accomplished by the sampling term $p(\mathbf{z}|\mathbf{x})$ and the adjusting term $p(\mathbf{z}|h(\mathbf{z}))$ with the DAG function h. Note that the adjusting term is short for two steps: 1) Draw $\mathbf{e} \sim p(\mathbf{e}|\mathbf{z})$; 2) Make $\mathbf{z} - \mathbf{e} \sim \mathcal{N}(0, \boldsymbol{I})$, which is a constraint that forces h to follow the DAG in $\mathbf{z}$. Thus, we call it adjusting.

While variables $\mathbf{z}$ and function h may be non-identifiable due to non-conditional additionally observed variables (*e.g.*, DAG label) (Khemakhem et al. 2020), we can choose suitable inductive biases to recover a certain structure in the real world (Locatello et al. 2019; Träuble et al. 2021). Besides, the formulation is also sufficient for classification based on the two causal principles. Empirical results in Section 5.4 also reveal some meaningful explanation.

Though (Yang et al. 2021a; Kim et al. 2021) have studied learning causality with VAE, their generative process is "noises $\to$ causal codes $\to$ images"and needs additional observations to learn the distributution of codes, which is impractical and limited. While our generative is "noise $\to$ images" and make "noise = causal codes". It is as flexible as vanilla VAE. Compared to Deconfounder (Wang and Blei 2019), our causal structure on the latent confounders is defined and to be learned by DAG learning methods.

3.4 Causal Latent Space

To achieve the downstream task such as clustering and classification, we introduce the causal latent space (CaLS) and study the computation of weighted sum in this space. Particularly, we assume the distribution of the causally independent codes $\mathbf{z} \in \mathbb{R}^{1\times d}$ is Gaussian [2]

$$\mathbf{z} \sim \mathcal{N}(h(\mathbf{z}), \boldsymbol{I}) \tag{5}$$

We refer to the latent codes in CaLS as causal codes, and the causal codes follow the same DAG. Then, the weighted sum latent codes can be obtained by the following proposition.

Proposition 1. *Assume there are n causal codes $\mathbf{Z} \in \mathbb{R}^{n\times d}$ shared same h that represents the DAG, an assignment $\boldsymbol{w} \in \mathbb{R}^{n\times 1}$ satisfying $\boldsymbol{w}^T\mathbf{1} = 1$ and the weighted sum $\overline{\mathbf{z}} = \boldsymbol{w}^T\mathbf{Z}$. Then*

$$\overline{\mathbf{z}} \sim \mathcal{N}(h(\overline{\mathbf{z}}), \boldsymbol{w}^T\boldsymbol{w}\boldsymbol{I}) \tag{6}$$

whenever h is linear or non-linear function. Proof is available in Supp. 3.2.

Proposition 1 shows that whatever the function h, the causal relationships of $\overline{\mathbf{z}}$ by weighted sum over $\mathbf{z}$ will remain unchanged as h can express the DAG structure.

4 Causal Meta VAE

To demonstrate the effectiveness of pipeline, we extent the baseline (Lee et al. 2021) into our CMVAE. It includes the Causal Mixture of Gaussian (CMoG), unsupervised meta-training and meta-test methods with novel causal Expectation Maximization. The following notation subscript is used: $\mathbf{z}_{[i]} \in \mathbb{R}^{1\times d}$ for i-th observation of $\mathbf{Z}$, and $\mathbf{z}_j \in \mathbb{R}^{n\times 1}$ for j-th dimension of $\mathbf{Z}$. Figure 3 shows the graphical model of CMVAE.

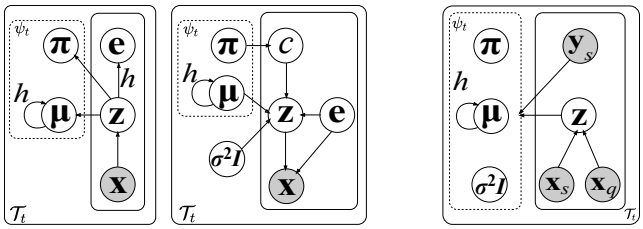
4.1 Causal Mixture of Gaussians

The Causal Mixture of Gaussians (CMoG) is an extension of MoG distribution in the CaLS based on proposition 1,

$$\begin{aligned} c \sim \mathrm{Cat}(\boldsymbol{\pi}), \quad \mathbf{z}|c \sim \mathcal{N}(\boldsymbol{\mu}_{[k]}, \boldsymbol{\sigma}^2_{[k]}\boldsymbol{I}), \\ \boldsymbol{\mu}_{[k]} \sim \mathcal{N}(h(\boldsymbol{\mu}_{[k]}), s_k^2\boldsymbol{I}) \end{aligned} \tag{7}$$

where $\boldsymbol{\pi}$ is K dimensional weights, $(\boldsymbol{\mu}_{[k]}, \boldsymbol{\sigma}^2_{[k]})$ are mean and diagonal covariance of the k-th mixture modality, and

[2] Actually we assume the error term $\boldsymbol{\epsilon} = \mathbf{z} - h(\mathbf{z})$ is Gaussian and ignore this error to focus on $\mathbf{z}$ and the corresponding space.



(a) (b)

Figure 3: Graphical model of CMVAE. (a) Unsupervised meta-training. CMoG prior $\psi_t = \{\boldsymbol{\pi}, \boldsymbol{\mu}\}$. [Left] Variational posterior $q_\phi(\mathbf{z}|\mathbf{x}, \mathcal{T}_t)$, $q_\phi(\mathbf{e}|\mathbf{z}, \mathbf{x})$. ψ_t is learned by causal-EM. [Right] Generative model $p_\theta(\mathbf{x}|\mathbf{z}, \mathbf{e})$, $p(\mathbf{z}|\mathbf{e})$. (b)Meta-test by semi-supervised causal-EM.

the scalar s_k^2 is a scaling parameter. Here we take the diagonal covariance $\boldsymbol{\sigma}_{[k]}^2 \boldsymbol{I}$ instead of $\boldsymbol{\Sigma}_k$ since the relationships between dimensions can be mined by learning the DAG. From another perspective, Eq. 7 can be seen as a regularization to make the modality causally independent. we refer to it as causal modality.

4.2 Unsupervised Meta-training

We now describe unsupervised meta-training in causal latent space based on VAE (Kingma and Welling 2014). Given a meta-training task $\mathcal{T}_t = \{\mathbf{x}^i \in \mathcal{U}\}_{i=1}^M$, the goals are to optimize the variational lower bound of the data marginal likelihood of task $\mathcal{T}_t$ using an variational posterior. Specifically, for the unsupervised meta-learning where labels are unknown, we define the variational posterior $q_\phi(\mathbf{z}|\mathbf{x}, \mathcal{T}_t)$ and the task-specific CMoG priors $p_{\psi_t^*}(\mathbf{z})$. For learning the causal structure, let $\mathbf{e}$ be sampled from the causal latent space, where function h is applied to $\mathbf{z}$, *i.e.*, $\mathbf{e}|\mathbf{z} \sim \mathcal{N}(h(\mathbf{z}), \boldsymbol{I})$. For posterior network, we use a factorization $q_\phi(\mathbf{e}, \mathbf{z}|\mathbf{x}, \mathcal{T}_t) = q_\phi(\mathbf{e}|\mathbf{z}, \mathbf{x}, \mathcal{T}_t) q_\phi(\mathbf{z}|\mathbf{x}, \mathcal{T}_t)$, sampling $\mathbf{z}$ given $\mathbf{x} \in \mathcal{T}_t$ first, then conditionally sampling $\mathbf{e}$ based on these values. It leads to the evidence lower bound (ELBO) (Details in Supp. 3.3),

$$\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x},\mathcal{T}_t)}[\mathbb{E}_{q_\phi(\mathbf{e}|\mathbf{z},\mathbf{x})}[\log p_\theta(\mathbf{x}|\mathbf{z},\mathbf{e}) - \log \frac{q_\phi(\mathbf{e}|\mathbf{z},\mathbf{x})}{p(\mathbf{e}|\mathbf{z})}] + \log p_{\psi_t^*}(\mathbf{z}) - \log q_\phi(\mathbf{z}|\mathbf{x}, \mathcal{T}_t)] \quad (8)$$

where $x \in \mathcal{T}_t$. The ELBO can be approximated by Monte Carlo estimation. We then describe these variational posteriors and priors in detail.

Variational Posterior. The task-conditioned variational posterior $q_\phi(\mathbf{z}|\mathbf{x}, \mathcal{T}_t)$ is to encode the dependency into the latent space between data in current task. Following (Lee et al. 2021), we take task $\mathcal{T}_t$ as inputs and denote,

$$H = \mathrm{TE}(F(\mathbf{x})), \mathbf{x} \in \mathcal{T}_t, \quad \boldsymbol{\mu} = W_{\boldsymbol{\mu}} H + b_{\boldsymbol{\mu}}, \\ \boldsymbol{\sigma}^2 = \exp(W_{\boldsymbol{\sigma}^2} H + b_{\boldsymbol{\sigma}^2}), q_\phi(\mathbf{z}|\mathbf{x}, \mathcal{T}_t) = \mathcal{N}(\mathbf{z}|\boldsymbol{\mu}, \boldsymbol{\sigma}^2) \quad (9)$$

where $\mathrm{TE}(\cdot)$ is multi-head self-attention mechanism (Vaswani et al. 2017), F is a convolutional neural network (or an identity function). To learn the causal structure, we apply the function h to the latent space and then sample $\mathbf{e}$ from the obtained causal latent space,

$$q_\phi(\mathbf{e}|\mathbf{z}, \mathbf{x}) = \mathcal{N}(\mathbf{e}|h(\mathbf{z}), \boldsymbol{I}), \quad \mathbf{z} \sim q_\phi(\mathbf{z}|\mathbf{x}, \mathcal{T}_t) \quad (10)$$

Causally Conditional Prior. Ideally if the DAG h represents the true causal structure, the conditional prior $p(\mathbf{e}|\mathbf{z})$ can be obtained by replacing the unknown h,

$$p(\mathbf{e}|\mathbf{z}) = \mathcal{N}(0, \boldsymbol{I}) + h(\mathbf{z}) = \mathcal{N}(\mathbf{e}|\mathbf{z}, 2\boldsymbol{I}) \quad (11)$$

Task-specific Prior. The task-specific causal multi-modal prior is modeled via CMoG and formally factorized as:

$$p_{\psi_t}(\mathbf{z}) = \sum_{c=0}^{K} p_{\psi_t}(\mathbf{z}|c) p_{\psi_t}(c), \quad p_{\psi_t}(c) = \mathrm{Cat}(c|\boldsymbol{\pi}), \\ p_{\psi_t}(\mathbf{z}|c) = \mathcal{N}(\mathbf{z}|\boldsymbol{\mu}_{[k]}, \boldsymbol{\sigma}_{[k]}^2 \boldsymbol{I}) \mathcal{N}(\boldsymbol{\mu}_{[k]}|h(\boldsymbol{\mu}_{[k]}), s_k^2 \boldsymbol{I}) \quad (12)$$

where the task-specific parameters ψ_t is defined as $\psi_t = \{\boldsymbol{\pi}, \boldsymbol{\mu}_{[k]}, \boldsymbol{\sigma}_{[k]}^2 \boldsymbol{I}, s_k^2\}$. Maximizing ELBO in Eq. 8 results in locally maximizing the following maximum causal posterior (MCP) problem:

$$\psi_t^* = \underset{\psi_t}{\mathrm{argmax}} \sum \log p(\psi_t|\mathbf{z}) \quad (13)$$

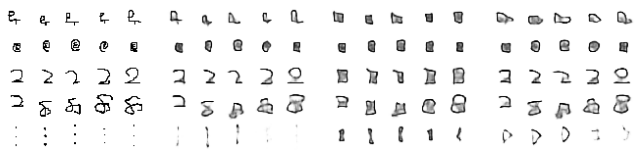
Without losing the DAG structure, the derived EM equations in closed forms are referred to as *causal-EM* (Derivations in Supp. 3.4),

$$\textbf{E: } \omega_{ik} = \frac{\alpha_k \mathcal{N}(\mathbf{z}_{[i]}|\boldsymbol{\mu}_{[k]}, \boldsymbol{I}) \mathcal{N}(\boldsymbol{\mu}_{[k]}|h(\boldsymbol{\mu}_{[k]}), \gamma^2 \boldsymbol{I})}{\sum_k \alpha_k \mathcal{N}(\mathbf{z}_{[i]}|\boldsymbol{\mu}_{[k]}, \boldsymbol{I}) \mathcal{N}(\boldsymbol{\mu}_{[k]}|h(\boldsymbol{\mu}_{[k]}), \gamma^2 \boldsymbol{I})} \\ \textbf{M: } \boldsymbol{\mu}_{[k]} = \frac{\sum_{i=1}^M \omega_{ik} \mathbf{z}_{[i]} (\boldsymbol{I} + \epsilon(\gamma^{-1}\boldsymbol{I})\epsilon^T(\gamma^{-1}\boldsymbol{I}))^{-1}}{\sum_{i=1}^M \omega_{ik}} \quad (14)$$

where $\epsilon(\mathbf{z}) = \mathbf{z} - h(\mathbf{z})$ and $\alpha_k = \frac{\sum_{i=1}^M \omega_{ik}}{\sum_{k=1}^K \sum_{i=1}^M \omega_{ik}}$. It can also be simplified using the inverse covariance matrix and we want to show that the term $\epsilon(\gamma^{-1}\boldsymbol{I})$ allows that j' propagates its information to j if $j' \in \mathrm{PA}(j)$, then intervenes and refines $\boldsymbol{\mu}_{[k]}$. Following the assumption of VAE, the covariance of Gaussian distribution is set to $\boldsymbol{I}$. We also observe that setting s_k^2 to a fixed hyper-parameter γ^2 results in better convergence. The α_k is initialized as $\frac{1}{K}$, and $\boldsymbol{\mu}_{[k]}$ is initialized as: $\boldsymbol{\mu}_{[k]} = \frac{\sum_i^K \mathbf{z}_{[i]} (\boldsymbol{I} + \epsilon(\gamma^{-1}\boldsymbol{I})\epsilon^T(\gamma^{-1}\boldsymbol{I}))^{-1}}{K}$ where $\{\mathbf{z}_{[i]}\}_{i=1}^K$ are randomly selected. By performing a few causal-EM steps iteratively, the MCP converges and task-specific parameters ψ_t^* is obtained.

Algorithm 1: Unsupervised Causal Meta-training

Input: An unlabeled dataset $\mathcal{U}$, causal-EM steps `step`.
Initialized parameterized q_ϕ, p_θ.
while *not converged* **do**
  Generate unlabeled task $\mathcal{T}_t = \{\mathbf{x}_u | \mathbf{x}_u \in \mathcal{U}\}$
  Draw $\mathbf{z} \sim q_\phi(\mathbf{z}|\mathbf{x}, \mathcal{T}_t)$, $\mathbf{e} \sim q_\phi(\mathbf{e}|\mathbf{z}, \mathbf{x})$ in Eq. 9, 10
  Compute ψ_t^* in Eq. 14 with `step` causal-EM
  Compute loss $\mathcal{L}$ in Eq. 16 and update ϕ, θ, h
end while



(a) (b) (c) (d)

Figure 4: Visualization on Omniglot. (a, b) The samples and generated samples for each mode at supervised meta-test step of CMVAE. Each row stands for each modality obtained by EM. (c, d) Counterfactual samples by intervention on causes and effects, respectively. The larger the change, the better the intervention, the more we can show that our method has learned the causes and effects.

4.3 Training Objective

DAG Loss. DAG loss is to ensure the "DAGness". We consider two types. 1) Linear SEM, $h(\mathbf{z}) = \mathbf{z}\mathbf{A}$, where $\mathbf{A} \in \mathbb{R}^{d\times d}$. 2) Nonlinear SEM, we model it with a multi-layer perceptron (MLP), $h_i(\mathbf{z}) = \sigma(\sigma(\sigma(\mathbf{z}\mathbf{W}_i^1)\cdots)\mathbf{W}_i^l)$, and define $[\mathbf{A}]_{mi} = \|m\,\text{th-row}(\mathbf{W}_i^1)\|_2$ where $\|\cdot\|_2$ is ℓ_2 norm. Then the DAG loss (Zheng et al. 2018) is

$$\mathcal{R}_D(\mathbf{A}) = (\text{tr}(\exp(\mathbf{A}\circ\mathbf{A})) - d)^2 \tag{15}$$

Objective. After getting the task-specific parameters ψ_t^*, we use gradient descent-based method *w.r.t.* the variational parameter ϕ, the generative parameter θ and the parameters of function h and minimize the following objective,

$$\mathcal{L} = -\text{ELBO} + \lambda_1\mathcal{R}_D(\mathbf{A}) + \lambda_2\|\mathbf{A}\|_1 \tag{16}$$

where λ_1, λ_2 are hyper parameters which control the "DAG-ness", and $\|\cdot\|_1$ is ℓ_1 norm.

Algorithm 1 shows the steps of the unsupervised meta-training stage. The outputs of unsupervised meta-training stage consists of variational parameter ϕ, the generative parameter θ and the parameters of function h. Similar to the regular meta-training stage, these outputs are also model initialization as it is a bi-level optimization (Liu et al. 2022; Vicol et al. 2022). The inner optimization is to maximize ELBO over task-specific ψ in Equation 8. In th outer loop, our method is to minimize the loss with regard to task-agnostic parameters ϕ, θ and h in Equation 16.

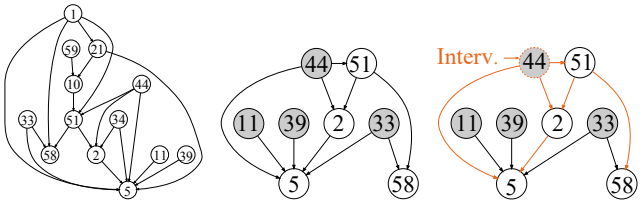
4.4 Supervised Meta-test

With CMoG priors, each causal modality can be seen as a pseudo-class concept. To adapt the causal modality to few-shot classification, we use both support set and query set and draw causal latent codes from the variational posterior q_ϕ. During the meta-test given a task $\mathcal{T} = \{(\mathcal{S}, \mathcal{Q})|\mathcal{S} = \{\mathbf{x}_s, \mathbf{y}_s\}_{s=1}^S, \mathcal{Q} = \{\mathbf{x}_q\}_{q=1}^Q\}$, the goal is to compute the conditional probability $p(\mathbf{y}_q|\mathbf{x}_q, \mathcal{T})$ *w.r.t.* variational posterior q_ϕ, the causal multi-modal prior parameter ψ^* and the backdoor adjustment in Equation 4:

$$p(\mathbf{y}_q|\mathbf{x}_q, \mathcal{T}) = \mathbb{E}_{q_\phi(\mathbf{z}_q|\mathbf{x}_q,\mathcal{T})p(\mathbf{z}_q|h(\mathbf{z}_q))}[p_{\psi^*}(\mathbf{y}_q|\mathbf{z}_q)] \tag{17}$$

Eq. 17 can also be computed by Bayes rule and Monte Carlo sampling. Then the predicted label is

$$\hat{\mathbf{y}}_q = \underset{k}{\text{argmax}}\, p(\mathbf{y}_q = k|\mathbf{z}_q, \mathcal{T}) \tag{18}$$



(a) (b) (c)

Figure 5: (a) DAG on Omniglot by the learned $\mathbf{A}$. Each node represents each dimension of $\mathbf{z}$. Other nodes are not shown because they are independent and have no cause-to-effect relationship (b) Part of DAG to show the causes and effects. The gray nodes represent the causes. (c) Intervention on one cause, *e.g.*, $\mathbf{z}_{44}$, will change the effects *e.g.*, $\mathbf{z}_5$ while will not change other causes, *e.g.*, $\mathbf{z}_{11}$. Best viewed in color.

To obtain the optimal prior parameters ψ^* in current meta-test task $\mathcal{T}$ and make the causal modality as label, we develop a semi-supervised causal-EM algorithm. In particular, we sample the causal code $\mathbf{z} \sim q_\phi(\mathbf{z}|\mathbf{x}, \mathcal{T})$ first and then get the causal multi-modalities with steps as follows,

$$\begin{aligned}
&\textbf{E: } \omega_{qk} = \frac{\mathcal{N}(\mathbf{z}_{[q]}|\boldsymbol{\mu}_{[k]}, \boldsymbol{\sigma}^2_{[k]})\mathcal{N}(\boldsymbol{\mu}_{[k]}|h(\boldsymbol{\mu}_{[k]}), \gamma^2\boldsymbol{I})}{\sum_k \mathcal{N}(\mathbf{z}_{[q]}|\boldsymbol{\mu}_{[k]}, \boldsymbol{\sigma}^2_{[k]})\mathcal{N}(\boldsymbol{\mu}_{[k]}|h(\boldsymbol{\mu}_{[k]}), \gamma^2\boldsymbol{I})} \\
&\textbf{M: } \quad \tilde{\boldsymbol{\mu}}_{[k]} = \sum_s \tilde{\omega}_{sk}\mathbf{z}_{[s]} + \sum_q \tilde{\omega}_{qk}\mathbf{z}_{[q]} \\
&\quad \boldsymbol{\mu}_{[k]} = \tilde{\boldsymbol{\mu}}_{[k]}(\boldsymbol{I} + \epsilon(\gamma^{-1}\boldsymbol{\sigma}_{[k]})\epsilon^T(\gamma^{-1}\boldsymbol{\sigma}_{[k]}))^{-1} \\
&\quad \boldsymbol{\sigma}^2_{[k]} = \sum_s \tilde{\omega}_{sk}(\mathbf{z}_{[s]} - \boldsymbol{\mu}_{[k]})^2 + \sum_q \tilde{\omega}_{qk}(\mathbf{z}_{[q]} - \boldsymbol{\mu}_{[k]})^2
\end{aligned} \tag{19}$$

where $\tilde{\omega}_{sk} = \frac{\mathbb{1}_{\mathbf{y}_s=k}}{\sum_s \mathbb{1}_{\mathbf{y}_s=k} + \sum_q \omega_{qk}}, \tilde{\omega}_{qk} = \frac{\omega_{qk}}{\sum_s \mathbb{1}_{\mathbf{y}_s=k} + \sum_q \omega_{qk}}$ and $\mathbb{1}$ is the indicator function. We keep the mixture probability fixed to $\frac{1}{K}$ due to the uniformly distributed labels and use diagonal covariance $\boldsymbol{\sigma}^2_{[k]}$ instead of $\boldsymbol{I}$ to obtain more accurate results. The $\boldsymbol{\mu}_{[k]}$ is initialized as: $\boldsymbol{\mu}_{[k]} = \frac{\sum_s \mathbb{1}_{\mathbf{y}_s=k}\mathbf{z}_{[s]}(\boldsymbol{I}+\epsilon(\gamma^{-1}\boldsymbol{I})\epsilon^T(\gamma^{-1}\boldsymbol{I}))^{-1}}{\sum_s \mathbb{1}_{\mathbf{y}_s=k}}$ Finally, we can get the solution of MCP and ψ^* by a few steps iteratively similar to the meta-training.

5 Experiment

In this section we show the empirical performance of our method on few-shot classification tasks.

5.1 Experiment Settings

Dataset. One biased toy dataset and three natural datasets are used to test our algorithm. **1) Toy dataset.** It is a 2-way biased dataset with a synthetic "bird" and "plane" image. (Details in Supp. 5.1.) **2) Omniglot.** Omniglot consists of 1,623 different characters and 20 images per character. Each image is 28 × 28 gray-scale. We take 1200, 100, 323 classes for training, validation and test, respectively. **3) *mini*ImageNet.** It is a subset of ImageNet (Russakovsky et al. 2015) and consists of 100 classes, 600 images per class

		Omniglot (way, shot)				*mini*ImageNet (way, shot)			
Method	Clustering	(5,1)	(5,5)	(20,1)	(20,5)	(5,1)	(5,5)	(5,20)	(5,50)
Training from Scratch	N/A	52.50	74.78	24.91	47.62	27.59	38.48	51.53	59.63
CACTUs-MAML	BiGAN	58.18	78.66	35.56	58.62	36.24	51.28	61.33	66.91
CACTUs-ProtoNets	BiGAN	54.74	71.69	33.40	50.62	36.62	50.16	59.56	63.27
CACTUs-MAML	ACAI/DC	68.84	87.78	48.09	73.36	39.90	53.97	63.84	69.64
CACTUs-ProtoNets	ACAI/DC	68.12	83.58	47.75	66.27	39.18	53.36	61.54	63.55
UMTRA	N/A	83.80	95.43	74.25	**92.12**	39.93	50.73	61.11	67.15
LASIUM-MAML-RO/N	N/A	83.26	95.29	-	-	40.19	54.56	65.17	69.13
LASIUMs-ProtoNets-RO/N	N/A	80.12	91.10	-	-	40.05	52.53	59.45	61.43
Meta-GMVAE	N/A	94.92	97.09	82.21	90.61	42.82	55.73	63.14	68.26
IFSL†	N/A	94.22	97.01	82.21	90.65	42.90	56.01	63.24	68.90
CMVAE (*ours*)	N/A	**95.11**	**97.14**	**82.58**	90.79	**44.27**	**58.95**	**66.25**	**70.54**
MAML (*Supervised*)	N/A	94.46	98.83	84.60	96.29	46.81	62.13	71.03	75.54
ProtoNets (*Supervised*)	N/A	98.35	99.58	95.31	98.81	46.56	62.29	70.05	72.04

Table 1: Accuracy results (way, shot) in Omniglot and *mini*ImageNet.

with size 84×84. we take 64 classes for training, 16 for validation and 20 for test, respectively. **4) CelebA.** CelebA consists of 202,599 face images with 10,177 number of identities. It has been used in the 5-way few-shot recognition task.

Evaluation metrics. During meta-test, we use the classes in the test set to generate 1000 tasks and compute the mean accuracy and 95% confidence interval on these tasks.

Implementation Details. We adopt the high-level feature reconstruction objective for toy dataset, *mini*-ImageNet and CelebA dataset. The backbone, variational posterior network $q_\phi(\mathbf{z}|\mathbf{x}, \mathcal{T}_t)$ and the high-level feature extractor (SimCLR (Chen et al. 2020)) are same as (Lee et al. 2021) (*i.e.*, 4-layer CNN for Omniglot and 5-layer CNN for others). For the generative network $p_\theta(\mathbf{x}|\mathbf{z}, \mathbf{e})$, we concatenate $\mathbf{z}$ and $\mathbf{e}$ in the last dimension, and it outputs the parameter of Bernoulli distribution for Omniglot and the mean of Gaussian distribution for *mini*ImageNet and CelebA. The function h is defined as described in section 3.4. There are no other parameters in $q_\phi(\mathbf{e}|\mathbf{z}, \mathbf{x})$. The number of iterations for causal-EM steps of all experiment is 10. The hyper-parameters γ, λ_1 and λ_2 are chosen based on the validation accuracy. We train all models for 60,000 iterations using Adam (Kingma and Ba 2015).

5.2 Baselines

We compare the following unsupervised meta-learning baselines with our approach. **CACTUs** (Hsu, Levine, and Finn 2019) extract features by ACAI (Berthelot* et al. 2019), BiGAN (Jeff Donahue 2017), and Deep- Cluster (Caron et al. 2018) and then train MAML or ProtoNets. **UMTRA** (Khodadadeh, Bölöni, and Shah 2019) generates training tasks by random sampling and augmentation for unsupervised meta-training. **Meta-GMVAE** (Lee et al. 2021) learns a set-level latent representation by EM algorithm. **LASIUMs** (Khodadadeh et al. 2021) creates synthetic training data by adding Noise, Random Out-of-class samples, and then train MAML or ProtoNets. **IFSL** (Yue et al. 2020) is a supervised method. We reimplement it by using backdoor adjustments in Meta-GMVAE. Furthermore, we compare the classic supervised methods **MAML** (Finn, Abbeel, and Levine 2017), **ProtoNets** (Snell, Swersky, and Zemel 2017) to indicate the gap between the supervised and unsupervised methods.

Algorithm	$S = 1$	$S = 5$
Training from scratch	34.69	56.50
CACTUs	41.42	62.71
UMTRA	39.30	60.44
LASIUM-RO-GAN-MAML	43.88	66.98
LASIUM-RO-VAE-MAML	41.25	58.22
LASIUM-RO-GAN-ProtoNets	44.39	60.83
LASIUM-RO-VAE-ProtoNets	43.22	61.12
Meta-GMVAE†	58.05	71.95
IFSL†	57.98	72.09
CMVAE (*Ours*)	**61.04**	**74.18**
MAML (*Supervised*)	85.46	94.98
ProtoNets (*Supervised*)	84.17	90.84

Table 2: Accuracy results on CelebA with 5-way, S-shot identity recognition. All the values are from (Khodadadeh et al. 2021), except for ours and † that we reproduce.

5.3 Results

Toy dataset. The 2-way 4-shot classification results in the toy dataset are 78.51 $\pm$ 0.36 for Meta-GMVAE and 93.08 $\pm$ 0.32 for our CMVAE. Since Meta-GMVAE does not take into account the context-bias, its performance is not impressive. While our CMVAE notices the existence of context-bias, the about 15% improvement on the biased toy dataset demonstrates that it offers the ability to alleviate the context-bias. **Natural dataset.** Table 1 reports the results of classification, where ACAI/DC (RO/N) mean ACAI clustering (Random Out-of-class samples) on Omniglot and DeepCluster (Noise) on *mini*ImageNet. Table 2 shows the results of 5-way few-shot identity recognition on CelebA. We can observe that our method outperforms state-of-the-art methods, except for the UMTRA on the 20-shot 5-shot classification in Omniglot. Our CMVAE even outperforms 5-way 1-shot classification supervised MAML in Omniglot. It is noticed

	Omniglot	*mini*ImageNet	CelebA
Default	**95.11 ± 0.47**	43.91 ± 0.74	59.93 ± 0.95
Linear	89.26 ± 0.56	42.68 ± 0.72	51.28 ± 0.91
$\lambda_1 = 10^{-1}$	94.46 ± 0.49	43.06 ± 0.75	59.84 ± 0.95
$\lambda_1 = 10^{5}$	94.42 ± 0.48	42.11 ± 0.74	59.72 ± 0.88
$\lambda_1 = 10^{10}$	91.28 ± 0.61	41.27 ± 0.70	50.92 ± 0.92
$\lambda_2 = 10^{-2}$	94.91 ± 0.50	42.88 ± 0.75	59.29 ± 0.93
$\lambda_2 = 10^{-3}$	94.95 ± 0.48	43.05 ± 0.75	59.27 ± 0.94
$\lambda_2 = 10^{-5}$	93.58 ± 0.49	42.66 ± 0.72	59.59 ± 0.95
$\gamma^2 = s_k^2$	90.34 ± 0.65	42.55 ± 0.75	54.25 ± 0.97
$\gamma^2 = 0.5$	94.76 ± 0.48	43.48 ± 0.74	60.25 ± 0.94
$\gamma^2 = 0.9$	92.40 ± 0.57	43.46 ± 0.74	**61.04 ± 0.94**
$\gamma^2 = 5$	92.52 ± 0.54	**44.27 ± 0.76**	59.04 ± 0.92
$\gamma^2 = 10$	58.29 ± 1.07	44.11 ± 0.75	59.04 ± 0.95

Table 3: Results of 5-way 1-shot classification on Omniglot, *mini*ImageNet and CelebA with different settings. We show the impact of choosing hyper parameters on test accuracies. Default: non-linear, $\lambda_1 = 1$, $\lambda_2 = 10^{-4}$, $\gamma = 1$.

that, for challenging dataset *e.g.*, *mini*ImageNet, our method outperforms Meta-GMVAE by more than about 2.5% average. This shows that 1) Our meta-learning network can capture the causal multi-modal distribution. 2) The causality is a more reliable in the natural images. 3) With causally independent codes and the adjustment for intervention, the confounding effect of meta-knowledge are removed.

Visualization. To better understand how CMVAE learns in the supervised meta-test stage, we visualize the real instances and ones generated by $p_\theta(\mathbf{x}|\mathbf{z}, \mathbf{e})$ in Figure 4a, 4b, where each row represents each modality. We can observe that 1) The distinction between real samples and generated samples reveals how well our generative ability for network $p(x|z, h)$ from output distribution. 2) Our CMVAE can capture the similar visual structure in each modality and make it as a class-concept in the meta-test stage.

5.4 Ablation Study

Counterfactual samples. To further demonstrate the effectiveness of the causality learned by CMVAE, we plot the DAG structure after obtaining $\mathbf{A}$ based on h in Figure 5. The nodes are a collection of dimensions of latent codes *i.e.*, $\mathbf{V} = \{\mathbf{z}_0, \cdots, \mathbf{z}_{63}\}$, and the edges represent cause-to-effect. Note that all the nodes are codes with semantics of interest. We can discover that $\mathbf{z}_1, \cdots, \mathbf{z}_{59}$ are the causes.

Figure 5c shows the intervention propagation. Because intervening causes will change the effects while intervening effects will not change the causes, the image will change more massively when intervening causes. Although we do not know which parts of the image these causes are responsible for generating, they are the most relevant to image generation. To this end, we generate counterfactual samples by intervening the causes and the effects, respectively, with the same amount (*e.g.*, 7 causes or 7 effects) and intervention value (*e.g.*, fixed to 0). Figure 4c, 4d show the visual results. Comparing them, we conclude as follows: 1) Intervention on the causes from the DAG results in larger changes. Since the intervention can propagate from causes to effects, the DAG learned by our CMVAE is reliable. 2) The causes are the most relevant to the images though we do not know what they means in complex real-world scenes.

	EM	Inverse	Causal EM
1-shot	129.59	139.45 (+7.6%)	145.31 (+12.1%)
5-shot	143.36	150.52 (+5.0%)	156.46 (+9.1%)

Table 4: Time (s) cost over 10000 20-way tasks on Omniglot during the meta-test stage. Inverse: Matrix inversion.

DAG type. We compare the performance of CMVAE with regard to the DAG type, *i.e.*, when the DAG function h is linear or non-linear. The results are shown in the Rows 1-2 of Table 3. We can observe that performances get worse when the function h is linear, which is in line with the common sense that the cause-to-effect is not a simple linear but a complex non-linear relation in the natural images.

Influence of λ_1, λ_2. The hyper parameters λ_1, λ_2 control the "DAGness". The larger λ_1 and λ_2, the more strongly causal relations are enforced. Rows 3-8 of Table 3 show that the setting when $\lambda_1 = 1, \lambda_2 = 10^{-4}$ outperforms other settings. This is because in the real-world images, factors with semantics are unknown and uncountable. The weak constraints can avoid overfitting the causal relations.

Effects of γ. The value of hype parameter γ controls the influence of causal regularization on modalities. We tuned this parameter using the validation classes with the following values: $[s_k^2, 0.5, 0.9, 1.0, 5, 10]$ where $s_k^2 = \sum_i (\frac{w_{ik}}{\sum_i w_{ik}})^2$ for the meta-training and $s_k^2 = \sum_s (\frac{\tilde{\omega}_{sk}}{\sum_s \tilde{\omega}_{sk} + \sum_q \tilde{\omega}_{qk}})^2 + \sum_q (\frac{\tilde{\omega}_{qk}}{\sum_s \tilde{\omega}_{sk} + \sum_q \tilde{\omega}_{qk}})^2$ for the meta-test based on the causal EM algorithm, and select the best γ corresponding to the best average 5-way 1-shot accuracy over meta-validation data for inference over the meta-test data. The last 5 rows of Table 3 shows the test class accuracies with respect to different values of γ. Though $\gamma^2 = 1, 5, 0.9$ provide the best results for different datasets, the default value already outperforms SOTA and it is user-friendly in practice.

Time complexity. Compared with the original EM, the inference of causal-EM comes with more time cost, as matrix operations (*i.e.*, inversion) have cubic time complexity. Table 4 reports that causal-EM costs about 10% more time, which is acceptable compared to the better accuracy.

6 Conclusion

The context-bias arises when the priors cause spurious corrections between inputs and predictions in unsupervised meta-learning. In this work, we offer an adjustment formulation that performs intervention on inputs to achieve bias-removal. We also develop CMVAE that carries out classification in causal latent space. Extensive experiments demonstrate that our approach has a better generalization ability across different tasks and datasets. CMVAE is also flexible for the extension to supervised learning. The limitation is that CMVAE may lack identifiability without any additional observation. We leave these questions for future work.

References

Bengio, Y.; Deleu, T.; Rahaman, N.; Ke, N. R.; Lachapelle, S.; Bilaniuk, O.; Goyal, A.; and Pal, C. J. 2020. A Meta-Transfer Objective for Learning to Disentangle Causal Mechanisms. In *International Conference on Learning Representations*.

Bernstein, D.; Saeed, B.; Squires, C.; and Uhler, C. 2020. Ordering-Based Causal Structure Learning in the Presence of Latent Variables. In *International Conference on Artificial Intelligence and Statistics*, volume 108, 4098–4108.

Berthelot*, D.; Raffel*, C.; Roy, A.; and Goodfellow, I. 2019. Understanding and Improving Interpolation in Autoencoders via an Adversarial Regularizer. In *International Conference on Learning Representations*.

Caron, M.; Bojanowski, P.; Joulin, A.; and Douze, M. 2018. Deep clustering for unsupervised learning of visual features. In *European Conference on Computer Vision (ECCV)*, 132–149.

Charpentier, B.; Kibler, S.; and Günnemann, S. 2022. Differentiable DAG Sampling. In *International Conference on Learning Representations*.

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. 2020. A Simple Framework for Contrastive Learning of Visual Representations. In *International Conference on Machine Learning*, 1597–1607.

Finn, C.; Abbeel, P.; and Levine, S. 2017. Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks. In *International Conference on Machine Learning*, volume 70, 1126–1135.

Glymour, M.; Pearl, J.; and Jewell, N. P. 2016. *Causal Inference in Statistics: A Primer*. John Wiley & Sons.

Higgins, I.; Matthey, L.; Pal, A.; Burgess, C.; Glorot, X.; Botvinick, M.; Mohamed, S.; and Lerchner, A. 2016. beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework. In *International Conference on Learning Representations*.

Hsu, K.; Levine, S.; and Finn, C. 2019. Unsupervised Learning via Meta-Learning. In *International Conference on Learning Representations*.

Jeff Donahue, T. D., Philipp Krähenbühl. 2017. Adversarial Feature Learning. In *International Conference on Learning Representations*.

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.

Khodadadeh, S.; Bölöni, L.; and Shah, M. 2019. Unsupervised Meta-Learning for Few-Shot Image Classification. In *Advances in Neural Information Processing Systems*, 10132–10142.

Khodadadeh, S.; Zehtabian, S.; Vahidian, S.; Wang, W.; Lin, B.; and Boloni, L. 2021. Unsupervised Meta-Learning through Latent-Space Interpolation in Generative Models. In *International Conference on Learning Representations*.

Kim, H.; Shin, S.; Jang, J.; Song, K.; Joo, W.; Kang, W.; and Moon, I.-C. 2021. Counterfactual Fairness with Disentangled Causal Effect Variational Autoencoder. *AAAI Conference on Artificial Intelligence*, 35: 8128–8136.

Kingma, D. P.; and Ba, J. 2015. Adam: A Method for Stochastic Optimization. In *International Conference on Learning Representations*.

Kingma, D. P.; and Welling, M. 2014. Auto-Encoding Variational Bayes. In *International Conference on Learning Representations*.

Kyono, T.; Zhang, Y.; and van der Schaar, M. 2020. CASTLE: Regularization via Auxiliary Causal Graph Discovery. In *Advances in Neural Information Processing Systems*.

Lee, D. B.; Min, D.; Lee, S.; and Hwang, S. J. 2021. Meta-GMVAE: Mixture of Gaussian VAE for Unsupervised Meta-Learning. In *International Conference on Learning Representations*.

Liu, R.; Gao, J.; Zhang, J.; Meng, D.; and Lin, Z. 2022. Investigating Bi-Level Optimization for Learning and Vision From a Unified Perspective: A Survey and Beyond. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44: 10045–10067.

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.

Lopez-Paz, D.; Nishihara, R.; Chintala, S.; Schölkopf, B.; and Bottou, L. 2017. Discovering Causal Signals in Images. In *Conference on Computer Vision and Pattern Recognition*, 58–66.

Magliacane, S.; Van Ommen, T.; Claassen, T.; Bongers, S.; Versteeg, P.; and Mooij, J. M. 2018. Domain Adaptation by Using Causal Inference to Predict Invariant Conditional Distributions. *Advances in Neural Information Processing Systems*, 31.

Pearl, J.; et al. 2000. CAUSALITY: Models, Reasoning and Inference. *Cambridge, UK: CambridgeUniversityPress*, 19.

Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. 2015. Imagenet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, 115: 211–252.

Scanagatta, M.; Corani, G.; De Campos, C. P.; and Zaffalon, M. 2016. Learning Treewidth-Bounded Bayesian Networks with thousands of Variables. *Advances in Neural Information Processing Systems*, 29.

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: 612–634.

Schölkopf, B.; Janzing, D.; Peters, J.; Sgouritsa, E.; Zhang, K.; and Mooij, J. M. 2012. On Causal and Anticausal Learning. In *ICML*.

Snell, J.; Swersky, K.; and Zemel, R. S. 2017. Prototypical Networks for Few-shot Learning. In *Advances in Neural Information Processing Systems*, 4077–4087.

Träuble, F.; Creager, E.; Kilbertus, N.; Locatello, F.; Dittadi, A.; Goyal, A.; Schölkopf, B.; and Bauer, S. 2021. On Disentangled Representations Learned from Correlated Data. In *International Conference on Machine Learning*, volume 139, 10401–10412.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, 5998–6008.

Vicol, P.; Lorraine, J. P.; Pedregosa, F.; Duvenaud, D.; and Grosse, R. B. 2022. On Implicit Bias in Overparameterized Bilevel Optimization. In *International Conference on Machine Learning*, volume 162, 22234–22259.

Viinikka, J.; Hyttinen, A.; Pensar, J.; and Koivisto, M. 2020. Towards Scalable Bayesian Learning of Causal DAGs. *Advances in Neural Information Processing Systems*, 33: 6584–6594.

Wang, T.; Huang, J.; Zhang, H.; and Sun, Q. 2020. Visual Commonsense R-CNN. In *Conference on Computer Vision and Pattern Recognition*, 10757–10767.

Wang, Y.; and Blei, D. M. 2019. The Blessings of Multiple Causes. *Journal of the American Statistical Association*, 114: 1574–1596.

Yang, M.; Liu, F.; Chen, Z.; Shen, X.; Hao, J.; and Wang, J. 2021a. CausalVAE: Disentangled Representation Learning via Neural Structural Causal Models. In *Conference on Computer Vision and Pattern Recognition*, 9593–9602.

Yang, X.; Zhang, H.; Qi, G.; and Cai, J. 2021b. Causal Attention for Vision-Language Tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9847–9857.

Yue, Z.; Zhang, H.; Sun, Q.; and Hua, X.-S. 2020. Interventional Few-Shot Learning. In *Advances in Neural Information Processing Systems*, volume 33, 2734–2746.

Zheng, X.; Aragam, B.; Ravikumar, P.; and Xing, E. P. 2018. DAGs with NO TEARS: Continuous Optimization for Structure Learning. In *Advances in Neural Information Processing Systems*, 9492–9503.

Zheng, X.; Dan, C.; Aragam, B.; Ravikumar, P.; and Xing, E. 2020. Learning Sparse Nonparametric DAGs. In *International Conference on Artificial Intelligence and Statistics*, 3414–3425.