Multi-Scope Representation Learning for Causal Relations Discovery with New Challenging Datasets

Jiageng Zhu[1,2]
jiagengz@isi.edu
Hanchen Xie[1,3]
hanchenx@isi.edu
Jianhua Wu[1,3]
jianhuaw@isi.edu
Mohamed E. Hussein[1]
mehussein@isi.edu
Mahyar Khayatkhoei[1]
mkhayat@isi.edu
Jiazhi Li[1,2]
jiazhil@isi.edu
Wael AbdAlmageed[4]
wabdalm@clemson.edu

[1] Information Sciences Institute
University of Southern California
Marina del Rey, California, USA

[2] Ming Hsieh Department of Electrical and Computer Engineering
University of Southern California
Los Angeles, California, USA

[3] Thomas Lord Department of Computer Science
University of Southern California
Los Angeles, California, USA

[4] Holcombe Department of Electrical and Computer Engineering
Clemson University
Clemson, South Carolina, USA

Abstract

Discovering semantic meaningful latent factors and the causal relations among them is an emergent topic in representation learning with notable impacts on real-world applications. However, many existing Causal Representation Learning (CRL) methods are hindered by strong assumptions, such as full data annotation, the need for counterfactual data, and/or prior knowledge of the causal structure. To address these limitations, we introduce *Causal-Macro*, a weakly supervised architecture that effectively discovers semantic causal factors and learns their causal relations. We theoretically show that *Causal-Macro* is identifiable in the sense that the marginalized posterior distribution of learned factors can be identified up to coordinate-wise reparameterization of ground-truth factors. Additionally, we show that existing CRL datasets are limited to simple causal graphs with a small number of generative factors. Thus, we propose two new datasets with a larger number of generative factors and more sophisticated causal graphs. Our comprehensive evaluations and detailed ablation studies demonstrate the superior performance of *Causal-Macro* over existing methods.

1 Introduction

Learning causality models cause-effect relations among factors [26], enabling interpretation and assessment of interventions [26, 28]. Causal machine learning has applications

© 2024. The copyright of this document resides with its authors.
It may be distributed unchanged freely in print or electronic forms.

in drug discovery [27], public health [9], and computer vision challenges like distribution shift [20], domain adaptation [32], and fairness [40]. However, semantically meaningful factors and causal relations are often unavailable *a priori* [28]. Causal representation learning (CRL) [28] learns semantically meaningful representations and causal relations from high-dimensional raw data. Many CRL methods rely on strong assumptions like intervenable causal mechanisms, counterfactual data [2, 33], temporal information [18, 19, 36], or known causal graphs [29, 33], which are hard to realize in real-world applications.

To mitigate the challenge within the scenario where the underlying causal mechanism cannot be interacted with, and only factual data is available [26], CausalVAE [35] attempts to discover causal structures by a linear causal layer in a fully supervised training manner, which requires extensive annotations. DoVAE [39] uses weak supervision with *do-operations* [26] in latent space to reduce annotation needs. However, DoVAE struggles to learn semantically meaningful representations and causal relations for many factors, uses a randomly initialized Graph Autoencoder (GAE) [24] risking incorrect causal relations, and overlooks nuances of chain-like causal structures [26] by swapping all possible cause factors.

To address the limitations of current methods, we introduce *Causal-Macro*, a novel CRL approach that contains two stages of training. Stage one learns micro-scope [4] features using an autoencoder with *Slots-Attention* [23]. The first stage focuses on basic feature learning without directly engaging in causal relation analysis. Subsequently, in the second stage, the method builds macro-scope representations and causal relations using the micro-scope features. This addresses challenges of underdeveloped visual encoders. *Causal-Macro* uses Hilbert-Schmidt Independence Criterion (HSIC) [10] with GAE for refined causal relations assessment. *Effect-Swap* strategically swaps one effect factor at a time using a mask m to focus on direct causal relations in chains. Root Cause Detection Module (RCDM) identifies root causes and adds training constraints. Further, we provide theoretical proof of the identifiability of *Causal-Macro*, underscoring its reliability and effectiveness.

Meanwhile, we argue existing benchmarks are simple or improperly designed. Pendulum, Flow [35], CelebA(SMILE), and CelebA(BEARD) [35] used in [35, 39] have few factors and simple causal graphs. Further, we observe that the ground-truth causal graphs of CelebA(BEARD) and CelebA(SMILE) are not properly aligned with their statistical distributions. To overcome these deficiencies, we propose *Shadow-Sunlight* and *Shadow-Pointlight*, simulating causal relations of light, floor, objects, and shadows under sunlight or point-light [6], which have more factors and complex causal graphs. Further, We also curate CelebA(BEARD) and CelebA(SMILE) to align statistics with ground-truth causal graphs.

The contributions of this paper are: (1) *Causal-Macro*; a novel and practical CRL method that utilizes an HSIC regression detector, *Effect-Swap* and a RCDM , (2) A proof of identifiability of weakly supervised *Causal-Macro*, (3) *Shadow CRL benchmarks*, which contain more generative factors than current CRL datasets and more sophisticated causal relations, (4) Identifying limitations of existing benchmarks and proposing curating two real datasets, (5) Comprehensive evaluation of existing and proposed datasets demonstrates the superiority of *Causal-Macro*, with ablation studies justifying design choices.

2 Related Works

Representation disentanglement and CRL: Disentangled representation learning focuses on the independence of latent factors [13], whereas CRL seeks causal relationships between them [28]. Variational Autoencoder (VAE) [16], which has commonly been used

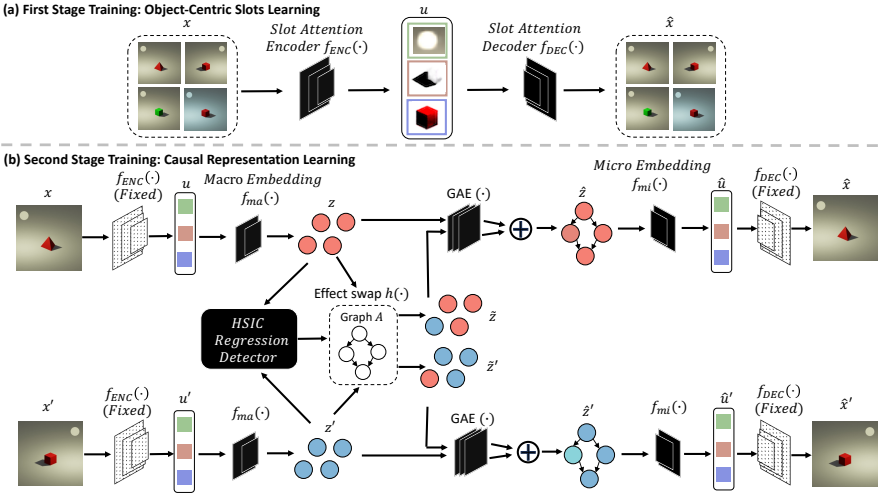

Figure 1: Two training stages of *Causal-Macro*. Micro-scope representation u is first obtained through unsupervised training. Then macro-scope causal representation z is coarsened from u by macro-embedding layer f_{ma} and the causal relations are learned by comparing the new reconstruction $\hat{x}$ and $\hat{x}'$ with the original inputs x and x'.

for both tasks, proposes disentangling the representation by minimizing reconstruction loss with Kullback-Leibler divergence (D_{KL}) as a regularizer on the latent space. Despite previous attempts of seeking disentangled representation via unsupervised VAE-based models [5, 13, 17], supervision is proved to be necessary to identify the relations between latent factor pairs [21]. To fulfill such requirement, many CRL methods focus on learning causal representations by relying on strong assumptions, such as utilizing counterfactual data [2, 33], incorporating temporal information [18, 19, 36] or knowing causal graph [29]. Contrarily, CausalVAE [35] directly uses the generative factor labels and adopts a linear causal discovery layer to discover the causal relations. DoVAE [39] further relaxes the need for strong supervision by leveraging swapping all cause factors to apply *do-operation* [26].

Structure learning: Structure learning, which focuses on learning causal graph structures from *census data*, includes three main approaches: constraint-based, score-based, and continuous optimization-based methods. Constraint-based methods [15, 31] test conditional independence between factors to explore causal graphs. Score-based methods [8, 11, 12] evaluate causal graph candidates using scoring functions. Continuous optimization-based methods, incorporating deep neural networks (DNNs), optimize specific objectives. For instance, NOTEARS [38] learns linear causal relations with an acyclicity constraint, while GOLEM [25] improves discovery performance using a new loss function. DAG-GNN [37] and Graph Autoencoders (GAE) [24] extend these concepts to estimate non-linear relations and enhance overall performance by adapting a graph neural network.

3 Causal-Macro

Causal-Macro sequentially learns micro-scope [4] representations to encode visual features, followed by macro-scope [4] causal representations, benefiting from well-encoded micro-

scope features. Fig. 1 illustrates *Causal-Macro* two-stage training scheme. First, raw input is encoded into micro-scope representations $u = [u_1, u_2, ...u_k]$, where u_i represents object visual features in slot i, containing sufficient information for unsupervised reconstruction. Next, the well-encoded micro-scope representation u is mapped to the macro-scope z (Sec. 3.1). In the macro-scope, causal relations are discovered using a Hilbert-Schmidt Independence Criterion [10] (HSIC) regression detector (Secs. 3.2 and 3.3).

3.1 From Micro to Macro

Micro-variables are fine-grained features containing complete, yet disorganized, information that is enough for reconstruction, obtained via unsupervised learning. As illustrated in Fig. 1(a), during the first stage of training, the *Slot Attention* Encoder f_{ENC} [23] utilizes softmax-based attention mechanism to map inputs to features in each slot, which is randomly initialized and iteratively refined with a recurrent function to align with a specific object within the input features. As empirically shown in Sec. 5.3, compared to regular VAE, utilizing *Slot Attention* module boosts overall CRL performance.

After the f_{ENC} and the decoder f_{DEC} are trained in the first stage, those features in u contain sufficient but unstructured visual information useful for downstream CRL tasks. In the second stage, u are coarsened to a macro-scope representation z via a macro-embedding layer f_{ma}. Since labels are often expensive to annotate, we seek to explore an approach that discovers the causal representation without using labels. Thus, we introduce a weak supervision signal by image-pairing and visual reconstruction through the micro-embedding layer f_{mi} and the decoder f_{DEC}, where f_{mi} serves as the inverse function of f_{ma}.

3.2 Weakly Supervised Causal Generative Model

We leverage input pairing to provide a weak supervision signal [22]. When the ground-truth values of the causal factors are absent, supervision can be introduced by utilizing the Local Markov Property of causality [26], i.e., $p(z) = \prod_i p(z_i|pa_i)$, where z is causal factors and pa_i is the set containing parents of factor i. By using this property and given the fact that the causal mechanisms inherent in a pair of inputs are the same [39], a weak supervision signal can be introduced via probing the possible changes of effect factors of a given image pair when the corresponding causes are not altered. Considering an optimal generator g^* and two causal factors z and z', two corresponding raw inputs x and x' can be generated via $g^*(z)$ and $g^*(z')$. z and z' obey the same causal mechanism $s(\cdot)$, which can also be referred to as solution function [2, 26] that constrains the unidirectional causal effects propagation from causes to their corresponding effect factors via $z_i = s_i(pa_i)$ [26]. Such constraint of s guarantees that varying an effect factor will not affect the output of s if parents of such factor are not altered. The overall generation processes can be expressed by Eqs. (1) to (3).

$$p(z) = \prod_i (z_i|pa_i), \qquad p(z') = \prod_i (z'_i|pa'_i) \tag{1}$$

$$z = s(z) \qquad z' = s(h(z, z', e)) \qquad e \sim p_E \tag{2}$$

$$x = g^*(z) \qquad x' = g^*(s(h(z, z', e))) \tag{3}$$

where E is the set of effect factor indices, e is a randomly sampled effect factor location from E according to probability p_E, and h is the *Effect-Swap* operation that replaces the value of one effect factor z'_e in z' with the corresponding value of its pair z_e. s can be implemented as any neural network that obeys causality constraints [24, 37, 38].

Theorem 3.1. *Consider the generative process described in Sec. 3.2 and assume that all cause factors are continuously distributed and the total number of generative factors d is known. Let g^* be a diffeomorphic function (i.e., smooth and invertible). Given infinite samples from $p(x,x')$ and the true number of parent factors corresponding to a specific effect, after training a generative model following Eqs. (2) and (3), the posterior $q(\hat{z}|x)$ is* identifiable *such that the marginalized posterior $q(\hat{z}) = \int q(\hat{z}|x)p(x)dx$ is a coordinate-wise reparameterization of the ground-truth causal factors $p(z)$ up to a permutation of the indices of z.* The Proof of identifiability is included in Appendix C.

We want to notice that while both our method and ILCM [7] use pairs of inputs for supervision, the types of input pairs differ significantly. ILCM requires one sample from the original causal mechanism $s(\cdot)$ and another from an intervened and different mechanism $\tilde{s}(\cdot)$, resulting in a counter-factual sample [26]. In contrast, our work focuses on scenarios where intervening the causal mechanism is not possible, so both inputs in our pair are generated by the original mechanism $s(\cdot)$. More Detailed discussions are included in Appendix E.3.

In practice, to introduce a weak supervision signal and align with the generation processes described in Eqs. (2) and (3), we first encode a pair of inputs through encoder f_{ENC} and macro-embedding layer f_{ma} to a pair of macro-scope representations z and z'. Then, we randomly choose one effect factor z_e and swap such effect factor with the corresponding factor in another macro-scope representation z'_e to create two new macro-scope representations $\tilde{z}$ and $\tilde{z}'$. These newly obtained representations are then fed into the solution function s that propagates causal effects from causes to effects. As illustrated in Fig. 1, to account for the nonlinearities of causal relations, we represent s with a graph autoencoder (GAE) [24]. If the GAE learns the correct causal relations between the effect factor and its corresponding causes, the GAE will adjust the effect factor value based on its parents. The whole process of *Effect-Swap* operation can be expressed as shown in Eq. (4):

$$\begin{aligned} \tilde{z} &:= [z_1, ..., z'_e, ..., z_d];\ \hat{z} = m \odot GAE(\tilde{z}) + (1-m) \odot GAE(z) \\ \tilde{z}' &:= [z'_1, ..., z_e, ..., z'_d];\ \hat{z}' = m \odot GAE(\tilde{z}') + (1-m) \odot GAE(z') \end{aligned} \tag{4}$$

where m is a mask with value one on the randomly selected effect factor e's location and zero elsewhere. m ensures GAE focuses on relations between z_e and its parents without affecting z_e's potential children. The new macro-scope latent representations, $\hat{z}$ and $\hat{z}'$, are fed into f_{mi} and f_{DEC} to generate reconstructions $\hat{x}$ and $\hat{x}'$, as shown in Eq. (5)

$$\begin{aligned} \hat{u} &:= f_{mi}(\hat{z});\ \ \hat{x} = Dec(\hat{u}) \\ \hat{u}' &:= f_{mi}(\hat{z}');\ \ \hat{x}' = Dec(\hat{u}') \end{aligned} \tag{5}$$

Given an optimal GAE, the reconstructions, $\hat{x}$ and $\hat{x}'$, should be consistent with the original inputs, x and x', respectively, because only latent effect factors z_e are exchanged by the *Effect-Swap*, and these factors are controlled by their parent factors. Thus, the local causal relation can be discovered by minimizing the distance between the reconstructions, $\hat{x}$ and $\hat{x}'$, and the original inputs x and x', respectively as $d(\hat{x},x) + d(\hat{x}',x')$, where d stands for any proper distance function, such as mean square error (MSE) or binary cross entropy (BCE). Further, throughout the training process, randomly selecting different effect factor location e statistically leads to every local causal relation being traversed so that the complete causal structure can be discovered. In addition, the loss between the inputs and the outputs of GAE [24], and the acyclic causal graph loss, $dag(A)$ [38], need to be included. The overall weak supervision loss is shown in Eq. (6), where α and β are the hyper-parameters.

$$L_{weakly} = d(\hat{x},x) + d(\hat{x}',x') + \alpha(||\hat{z} - z||_2^2 + ||\hat{z}' - z'||_2^2) + \beta dag(A) \tag{6}$$

Please note that, although not needed, if a small amount of ground-truth labels l are available, our model can easily be extended to utilize such a supervision signal:

$$L_{semi} = D_{KL}(q(z|x,l)||p(z|l)) + ||l - GAE(l)||_2^2 + L_{weakly} \tag{7}$$

3.3 Causal Discovery by Regression

Causal directions can be inferred by regression [1, 3, 14, 34]. Thus, we use a regression detector in the macro-scope to discover the correct causal relations in practice. Given two causally associated factors X and Y, the smaller regression loss in a certain causal direction indicates higher confidence in the correctness of such a direction, i.e.,

$$L(Y, \Phi(X)) < L(X, \Psi(Y)) \Rightarrow X \text{ is cause of } Y, \tag{8}$$

where Φ and Ψ are the mapping functions from X to Y and Y to X, respectively. Multiple methods, such as linear, ridge, and HSIC regression, can be applied. However, as discussed in [1], linear or ridge regression based detection requires all variables to be on the same scale because MSE is used to compare the losses between two directions. Thus, we use HSIC regression, which is free of such constraint and achieves the best performance, as discussed in Sec. 5.3. The details of the causal discovery based on HSIC regression are described in Appendix F.3, where the correct causal direction between a pair of factors is indicated by the lower value of comparing the regression losses obtained from each direction. The causal graph detected by HSIC regression will be utilized to apply *Effect-Swap* operation and propagate causal relations within the GAE.

To further enforce the correctness of the discovered causal graph, we propose a Root Cause Discovery Module (RCDM) to provide additional constraints. For a causal graph, there is at least one root cause factor that is not controlled by any other factors. We utilize $D_{KL}(q(z_i|x)||q(z_i'|x'))$ to find such root cause factors, where i stands for the ith element. Given a pair of latent representation elements describing the same semantic meaning, their D_{KL} is expected to be zero [22]. Given the complexity of modern neural networks, a pair of corresponding root cause factors is more likely to reach such an ideal case because root cause factors are solely controlled by themselves, compared with other factors. Thus, by calculating the D_{KL} among all feature element pairs, we take the factor of the pair with the smallest D_{KL} as the root cause factor. To implement this constraint during training, for a detected pair of root cause factors, we first calculate the average value between them. Then, we use this average value to replace both latent root cause factors. After decoding, their reconstructions should be the same as the original inputs with respect to the semantic factors.

4 Shadows Benchmarks and Curated Real-world Dataset

As shown in Sec. 3.3, existing datasets, both synthetic and real, are too simple for CRL due to their limited factors of variation, making them unsuitable for constructing complex causal graphs. To address this, we propose two novel datasets: *Shadow-Sunlight* and *Shadow-Pointlight*, which contain more factors of variation and enable the creation of more complex causal graphs. Additionally, we observe that the real datasets, CelebA(BEARD) and CelebA(SMILE) [35], are not statistically consistent with their originally proposed causal graphs. Therefore, we propose curating these datasets for correct CRL evaluation.

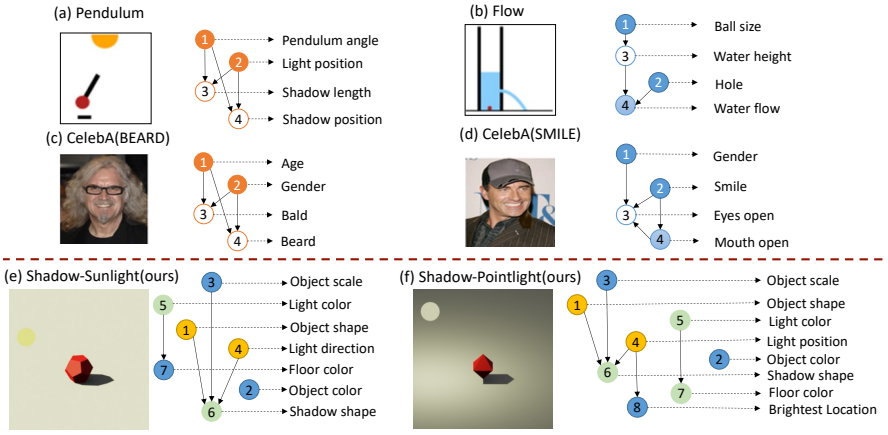

Figure 2: **Datasets:** Samples and ground-truth causal graphs of exisiting datasets and proposed new datasets. First two rows show existing datasets. In the third row, we provide samples and causal graphs of our newly proposed *Shadow-Sunlight* and *Shadow-Pointlight*.

Shadow Benchmarks: To address the limitation of current datasets, we propose two novel datasets, *Shadow-Sunlight* and *Shadow-Pointlight*, which contain seven and eight factors of variation, respectively, so that more complex causal graphs can be constructed. Shadow datasets are generated using Blender [6] with the Cycle rendering engine. The proposed datasets simulate the causal relations between light, object, floor, and shadow. In *Shadow-Sunlight*, we set the light source type to be *sun light*, which emits parallel light rays. In *Shadow-Pointlight*, we set the light source type to be *point light*, in which all light rays are emitted from a single point. Due to the different attributes of the two types of light sources, the causal mechanisms inherent in the two environments are different, which leads to two distinct datasets. The object attributes in the Shadow datasets are the cross-product of seven different object shapes, seven different object colors, and seven different object scales. Further, there are six different light colors. In *Shadow-Sunlight*, there are 20 different light directions, and in *Shadow-Pointlight*, the light directions are controlled by 20 different light positions. Since the shadow shape, the floor color, and the brightest floor position are effect factors, their values are controlled by object shape, object scale, light position/direction, and light color. Object color serves as the nuisance factor, which has no causal relation with other factors. More details are in Appendix A.

Improvements of Real-world Benchmarks: CelebA(BEARD) and CelebA(SMILE) [35] are two real datasets used for CRL, with originally proposed causal graphs shown in Sec. 3.3(c) and (d). We detect inconsistencies between the statistical relations of factor pairs in these datasets and the originally proposed causal graphs, indicating that learning optimal causal graphs from the original data is ill-posed. As discussed in [26], conditional independence tests can assess the correctness of a causal graph. Using the χ^2 test, we evaluate the correctness of CelebA(BEARD)'s originally proposed causal graph [35], which assumes Age and Gender are mutually independent. However, the p-value of the χ^2 test is 10^{-5}, much smaller than the significance level (0.01), indicating that Age and Gender are not mutually independent. To align the dataset with the originally proposed causal graph, we randomly remove samples to make Gender and Age statistically independent. A similar curation process is applied to CelebA(SMILE). More details are provided in Appendix B.

Models	Pendulum						Flow					
	PosMIC ↑	PosTIC ↑	NegMIC ↓	NegTIC ↓	F_1^{MIC} ↑	F_1^{TIC} ↑	PosMIC ↑	PosTIC ↑	NegMIC ↓	NegTIC ↓	F_1^{MIC} ↑	F_1^{TIC} ↑
Fully Supervised learning methods (all labels are used)												
CausalVAE [36]	53.0±4.5	43.4±3.7	46.6±3.9	37.0±4.2	53.2±3.6	51.4±3.2	45.1 ±4.8	36.7 ±4.2	43.3 ±5.1	33.7 ±3.2	50.2 ±4.4	47.3 ±3.7
ConditionVAE [40]	36.5±3.0	27.8±3.2	34.6±4.2	25.7 ±3.6	46.9 ±4.7	40.5 ±3.5	28.6 ±3.2	21.3 ±3.1	27.2 ±2.8	20.6 ±2.7	41.1 ±5.1	33.6 ±4.0
Unsupervised Learning methods (no label is used)												
CausalVAE(unsup) [36]	20.5 ±2.6	11.8 ±2.7	23.3 ±3.2	14.7 ±1.9	32.4 ±3.4	20.7 ±3.1	22.8 ±2.7	12.5 ±1.4	21.5 ±2.4	12.0 ±1.9	35.3 ±5.6	21.9 ±4.7
β-VAE [5]	21.2 ±2.7	12.7 ±2.9	23.7 ±3.1	12.6 ±1.9	33.2 ±3.3	22.2 ±2.7	23.6 ±3.6	12.5 ±1.9	22.1 ±2.5	11.4 ±1.9	36.2 ±4.9	21.9 ±4.2
LadderVAE [24]	15.2 ±1.9	8.6 ±1.0	**14.2 ±1.7**	**7.9 ±0.9**	25.8 ±3.0	15.7 ±2.8	16.2 ±1.8	10.5 ±1.0	**13.3 ±1.2**	**6.9 ±0.6**	27.3 ± 3.2	18.9 ±2.8
Reduced supervision method (no label is used; supervision source is image pairing)												
Do-VAE [36]	54.1 ±4.5	44.0 ±4.2	40.2 ±3.9	31.6 ±3.2	56.8 ±5.2	53.6 ±4.3	50.7 ±4.7	41.3 ±4.2	36.8 ±3.8	27.2 ±3.0	56.3 ±5.9	52.7 ±4.9
ILCM [8]	52.1 ±4.6	41.2 ±3.9	35.2 ±2.9	27.6 ±2.2	56.2 ±4.2	53.1 ±4.0	56.7 ±4.7	47.3 ±4.1	31.8 ±3.3	25.2 ±2.7	60.3 ±4.3	58.4 ±3.8
Causal-Macro(Ours)	**67.4 ±3.6**	**57.1 ±2.6**	32.0 ±3.0	25.5 ±2.5	**65.3 ±3.1**	**60.9 ±2.5**	**63.9 ±4.0**	**54.7 ±3.2**	25.0 ±4.1	17.9 ±2.9	**65.2 ±3.3**	**64.4 ±3.0**

Table 1: Causal representation metrics tested on Pendulum and Flow.

Models	Shadow-Sunlight						Shadow-Pointlight					
	PosMIC ↑	PosTIC ↑	NegMIC ↓	NegTIC ↓	F_1^{MIC} ↑	F_1^{TIC} ↑	PosMIC ↑	PosTIC ↑	NegMIC ↓	NegTIC ↓	F_1^{MIC} ↑	F_1^{TIC} ↑
Fully Supervised learning methods (all labels are used)												
CausalVAE [36]	43.6±5.2	35.6±3.8	33.1±6.9	22.5±4.5	49.1±5.6	44.9±4.2	39.2 ±3.7	29.1 ±4.5	31.1 ±5.1	25.1 ±3.8	54.2 ±4.6	50.2 ±4.7
ConditionVAE [40]	18.5±3.3	10.6±2.8	25.8±4.1	15.7 ±5.1	29.4 ±4.1	18.6±3.1	11.6 ±2.5	4.8 ±2.2	**14.4 ±2.6**	**6.6 ±2.2**	20.3 ±3.8	12.7 ±3.1
Unsupervised Learning methods (no label is used)												
CausalVAE(unsup) [36]	13.2 ±3.3	7.6 ±2.3	17.5±3.2	**9.7 ±3.9**	22.4 ±4.7	14.4 ±4.8	12.5 ±2.8	5.6 ±1.5	14.9 ±2.4	7.3 ±2.0	19.7 ±3.6	10.9 ±3.5
β-VAE [5]	12.7 ±4.6	6.7 ±4.2	18.8 ±4.3	11.2 ±3.7	21.6 ±3.6	12.2 ±3.3	11.3 ±3.3	4.8 ±1.6	14.6 ±2.5	7.1 ±2.2	19.8 ±5.0	11.7 ±3.8
LadderVAE [24]	13.7 ±3.1	6.0 ±3.5	**17.2 ±4.7**	10.7 ± 2.9	22.8 ±4.5	11.1 ±5.1	7.9 ±4.1	5.2 ±1.9	15.2 ±3.2	8.7 ±2.6	14.3 ± 2.6	9.8 ±2.1
Reduced supervision method (no label is used; supervision source is image pairing)												
DoVAE [36]	32.6 ±3.5	26.7 ±4.1	31.8 ±2.5	24.0 ±2.8	43.9 ±3.3	34.8 ±3.8	34.7 ±4.4	30.8 ±4.3	40.1 ±4.8	29.7 ±3.3	43.3 ±4.4	40.1 ±3.8
ILCM [8]	45.6 ±4.5	39.7 ±3.7	32.8 ±2.9	26.2 ±2.7	56.7 ±3.9	52.7 ±4.0	44.2 ±4.7	36.8 ±3.7	35.2 ±3.8	27.2 ±3.4	55.1 ±3.9	52.4 ±3.8
Causal-Macro(Ours)	**59.7 ±3.0**	**47.1 ±3.6**	25.3 ±3.1	20.5 ±2.2	**62.9 ±3.9**	**59.2 ±3.1**	**60.5 ±3.9**	**50.1 ±3.5**	32.1 ±3.2	25.6 ±2.9	**62.4 ±3.3**	**60.2 ±3.1**

Table 2: Causal representation metrics tested on Shadow-Sunlight and Shadow-Pointlight.

5 Experimental Evaluation

5.1 Benchmarks

Datasets: To demonstrate the effectiveness of our method, we performed experiments using the newly created Shadow datasets and the Pendulum and Flow benchmarks [35]. We also evaluated our method on real-world datasets, curated versions of CelebA(BEARD) and CelebA(SMILE), as detailed in Sec. 4. In these CelebA experiments, where nuisance factors (36) greatly outnumber causal factors (4), accurate evaluation requires focusing on specific attributes [7]. Thus, we conducted these tests in a semi-supervised setting, using $\{10\%, 20\%, 30\%, 40\%\}$ labeled data [39].

Metrics: To evaluate the CRL performance, we utilize PosMIC/TIC (Positive Maximal/Total Information Coefficient) and NegMIC/TIC (Negative Maximal/Total Information Coefficient) proposed by [39]. To calculate *positive* metrics, latent effect factors are set to zero before the causal discovery layer. Then, MIC and TIC between the latent effect factors after the causal discovery layer and the ground-truth generative effect factors are calculated to be the final scores of PosMIC and PosTIC. Higher *positive* metrics indicate better performance because an effect value is expected to be correctly inferred from its causes. Conversely, NegMIC and NegTIC assess the falseness of causal effects by firstly setting the latent cause factors before the causal discovery layer to zero, then the MIC and TIC between the latent cause factors after the causal discovery layer and the ground-truth generative cause factors are calculated to be the final score. Since causal relations should only propagate from causes to effects [26], the lower scores of *negative* metrics indicate better performance. Further, to combine both *positive* and *negative* metrics, F_1^{MIC} and F_1^{TIC} [39] is obtained by calculating the harmonic mean between the *positive* metrics and one minus the *negative* metrics. All metrics range from zero to one, and we scale them by 100. Details are included in Appendix.

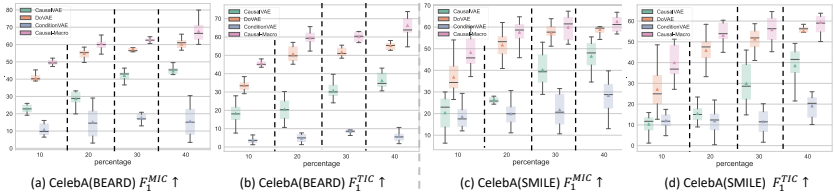

Figure 3: F_1^{MIC} and F_1^{TIC} results on the curated CelebA(BEARD) and CelebA(smile). Full results are included in Appendix E.2.

Training Scheme	*Shadow-Sunlight*	
	F_1^{MIC} ↑	F_1^{TIC} ↑
End-to-End	55.2±5.6	52.4±4.7
Two-stage regular VAE	60.9 ±4.2	55.0 ±4.1
Two-Stages Slots(Ours)	**62.9 ±3.9**	**59.2 ±3.1**

Table 3: Ablation study of comparing two-stage training with different generators and with end-to-end training.

Regression Model	*Shadow-Sunlight*	
	F_1^{MIC} ↑	F_1^{TIC} ↑
Linear	47.9±3.3	46.2±3.6
Ridge	50.2±3.1	47.7±3.9
HSIC(Ours)	**62.9 ±3.9**	**59.2 ±3.1**

Table 4: Ablation study of different regression methods for causal direction detection.

Models	*Shadow-Sunlight*	
	F_1^{MIC} ↑	F_1^{TIC} ↑
w/o RCDM	53.2±4.8	50.4±3.9
w/o average replacement	55.9±6.1	52.8±4.9
full RCDM	**62.9 ±3.9**	**59.2 ±3.1**

Table 5: Ablation study about RCDM. The results are tested on *Shadow-Sunlight* dataset.

5.2 Comparing with SOTA Methods

As shown in Secs. 4 and 5.1 and Fig. 3, the proposed *Causal-Macro* outperforms SOTA methods. CausalVAE [35] limits the causal discovery layer to be linear, thus showing sub-optimal performance. ConditionVAE [30] forces latent factors to be mutually independent so that, although it achieves good results on NegMIC and NegTIC, it cannot discover the correct causal relations. Unsupervised methods, CausalVAE(unsup) [35], β-VAE [13] and LadderVAE [17], show poor performances on PosMIC, PosTIC, F_1^{MIC}, and F_1^{TIC} because they cannot learn the expected latent representation and encode semantic information [21]. Unsupervised methods achieve low value on NegMIC and NegTIC due to barely learning semantic information. Besides, introducing supervision signal is necessary since Locatello *et al.* [21] shows that unsupervised learning method is impossible and is highly unstable for identifying the latent representation relationships. DoVAE [39], which also depends on weak supervision, merely relies on an underdeveloped GAE for discovering causal relations during training so that it may fail when a large number of causal factors exists, as shown in Sec. 5.1. ILCM [2], which requires one sample in an input pair to be counter-factual data generated by applying intervention on the underlying causal mechanism, fails to achieve good performance when the underlying causal mechanism can not be interacted with and only factual data is available. In contrast, *Causal-Macro* only requires factual samples where the underlying causal mechanism need not be intervened. By first learning to effectively encode visual information and then employing a nonlinear causal discovery layer with advanced techniques, including HSIC regression detector and RCDM. This comprehensive approach allows *Causal-Macro* to consistently excel in F_1^{MIC} and F_1^{TIC}, demonstrating its superiority under the scenario where only factual data is available. We also include qualitative visualizations of *Causal-Macro* in Appendix G.

5.3 Discussion

Two-stage Training Scheme: The two-stage training approach first trains the model to encode visual information into a lower-dimensional representation. Subsequently, this refined

visual representation is used to simplify learning causal relationships. In contrast, the end-to-end approach trains the model to process visual information and identify causal connections simultaneously, potentially leading to poorer outcomes if the visual representation is not fully developed. As shown in Tab. 3, the two-stage method yields superior results, where models that incorporate a Slot Attention module in the first stage outperform regular VAEs.

Regression Model Alternatives: Besides HSIC regression, linear and ridge regression can also be used for causal discovery [1, 3]. However, they are limited by requiring that all variables are on the same scale [1] because they use MSE to compare regression losses between two directions, and variables on different scales can lead to false judgment. Unfortunately, such a requirement cannot be guaranteed in CRL. As shown in Tab. 4, HSIC regression achieves the best performance because of being free of such a constraint.

Discussion on RCDM: Since all causal graphs have at least one root cause that should only be controlled by itself, including RCDM can help the model to identify the root causes and enforce the self-dependency property. As shown in Tab. 5, comparing $\text{Row}_{1,3}$, having RCDM boost F_1^{MIC} and F_1^{TIC}. Comparing $\text{Row}_{2,3}$, instead of replacing the values of the detected root cause factors with their average, simply swapping them with each other ignores the self-dependency property of root cause factors, which can degrade the CRL performance, especially under sophisticated scenarios.

6 Conclusion

In this study, we introduce *Causal-Macro*, a method that first learns fine-scale representations and then uncovers causal relationships at a broader scale using tools like *Effect-Swap*, HSIC regression, and RCDM. We also develop the Shadow datasets to intensify challenges in Causal Representation Learning (CRL) and improve real-world datasets through curation. Our experiments across these datasets showcase our method's superiority, corroborated by detailed ablation studies of *Causal-Macro*.

References

[1] Patrick Blöbaum, Dominik Janzing, Takashi Washio, Shohei Shimizu, and Bernhard Schölkopf. Cause-effect inference by comparing regression errors. In *AISTATS*, pages 900–909, 2018. URL http://proceedings.mlr.press/v84/bloebaum18a.html.

[2] Johann Brehmer, Pim De Haan, Phillip Lippe, and Taco Cohen. Weakly supervised causal representation learning. In *Advances in Neural Information Processing Systems*, 2022.

[3] Kailash Budhathoki and Jilles Vreeken. Origo: causal inference by compression. *Knowledge and Information Systems*, 56(2):285–307, November 2017. doi: 10.1007/s10115-017-1130-5. URL https://doi.org/10.1007/s10115-017-1130-5.

[4] Krzysztof Chalupka, Frederick Eberhardt, and Pietro Perona. Causal feature learning: an overview. *Behaviormetrika*, 44(1):137–164, January 2017.

[5] Ricky T. Q. Chen, Xuechen Li, Roger Grosse, and David Duvenaud. Isolating sources of disentanglement in variational autoencoders, 2019.

[6] Blender Online Community. *Blender - a 3D modelling and rendering package*. Blender Foundation, Stichting Blender Foundation, Amsterdam, 2022. URL https://www.blender.org.

[7] Zunlei Feng, Xinchao Wang, Chenglong Ke, An-Xiang Zeng, Dacheng Tao, and Mingli Song. Dual swap disentangling. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper/2018/file/fdf1bc5669e8ff5ba45d02fded729feb-Paper.pdf.

[8] Dan Geiger and David Heckerman. Learning gaussian networks. In *Proceedings of the Tenth International Conference on Uncertainty in Artificial Intelligence*, UAI'94, page 235–243, San Francisco, CA, USA, 1994. Morgan Kaufmann Publishers Inc. ISBN 1558603328.

[9] Thomas A. Glass, Steven N. Goodman, Miguel A. Hernán, and Jonathan M. Samet. Causal inference in public health. *Annual Review of Public Health*, 34(1):61–75, 2013. doi: 10.1146/annurev-publhealth-031811-124606.

[10] Arthur Gretton, Olivier Bousquet, Alex Smola, and Bernhard Schölkopf. Measuring statistical dependence with hilbert-schmidt norms. In Sanjay Jain, Hans Ulrich Simon, and Etsuji Tomita, editors, *Algorithmic Learning Theory*, pages 63–77, Berlin, Heidelberg, 2005. Springer Berlin Heidelberg. ISBN 978-3-540-31696-1.

[11] Peter D. Grünwald and Paul M.B. Vitányi. Algorithmic information theory. In Pieter Adriaans and Johan van Benthem, editors, *Philosophy of Information*, Handbook of the Philosophy of Science, pages 281–317. North-Holland, Amsterdam, 2008. doi: https://doi.org/10.1016/B978-0-444-51726-5.50013-3. URL https://www.sciencedirect.com/science/article/pii/B9780444517265500133.

[12] David Heckerman, Dan Geiger, and David M. Chickering. Learning bayesian networks: The combination of knowledge and statistical data. *Machine Learning*, 20 (3):197–243, Sep 1995. ISSN 1573-0565. doi: 10.1023/A:1022623210503. URL https://doi.org/10.1023/A:1022623210503.

[13] Irina Higgins, Loïc Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. URL https://openreview.net/forum?id=Sy2fzU9gl.

[14] Patrik Hoyer, Dominik Janzing, Joris M Mooij, Jonas Peters, and Bernhard Schölkopf. Nonlinear causal discovery with additive noise models. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Advances in Neural Information Processing Systems*, volume 21. Curran Associates, Inc., 2008. URL https://proceedings.neurips.cc/paper/2008/file/f7664060cc52bc6f3d620bcedc94a4b6-Paper.pdf.

[15] Markus Kalisch and Peter Bühlman. Estimating high-dimensional directed acyclic graphs with the pc-algorithm. *Journal of Machine Learning Research*, 8(3), 2007.

[16] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In Yoshua Bengio and Yann LeCun, editors, *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014. URL http://arxiv.org/abs/1312.6114.

[17] Daniel D. Lee, Masashi Sugiyama, Ulrike von Luxburg, Isabelle Guyon, and Roman Garnett, editors. *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, 2016. URL https://proceedings.neurips.cc/paper/2016.

[18] Phillip Lippe, Sara Magliacane, Sindy Löwe, Yuki M Asano, Taco Cohen, and Efstratios Gavves. CITRIS: Causal identifiability from temporal intervened sequences. In *ICLR2022 Workshop on the Elements of Reasoning: Objects, Structure and Causality*, 2022. URL https://openreview.net/forum?id=H87xrH_Lcg9.

[19] Phillip Lippe, Sara Magliacane, Sindy Löwe, Yuki M Asano, Taco Cohen, and Efstratios Gavves. Causal representation learning for instantaneous and temporal effects in interactive systems. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=itZ6ggvMnzS.

[20] Chang Liu, Xinwei Sun, Jindong Wang, Haoyue Tang, Tao Li, Tao Qin, Wei Chen, and Tie-Yan Liu. Learning causal semantic representation for out-of-distribution prediction. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 6155–6170. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper/2021/file/310614fca8fb8e5491295336298c340f-Paper.pdf.

[21] Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Rätsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations, 2019.

[22] Francesco Locatello, Michael Tschannen, Stefan Bauer, Gunnar Rätsch, Bernhard Schölkopf, and Olivier Bachem. Disentangling factors of variations using few labels. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=SygagpEKwB.

[23] Francesco Locatello, Dirk Weissenborn, Thomas Unterthiner, Aravindh Mahendran, Georg Heigold, Jakob Uszkoreit, Alexey Dosovitskiy, and Thomas Kipf. Object-centric learning with slot attention. In *NeurIPS 2020*, 2020.

[24] Ignavier Ng, Shengyu Zhu, Zhitang Chen, and Zhuangyan Fang. A graph autoencoder approach to causal structure learning. *arXiv preprint arXiv:1911.07420*, 2019.

[25] Ignavier Ng, AmirEmad Ghassami, and Kun Zhang. On the role of sparsity and dag constraints for learning linear dags. In *NeurIPS*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/d04d42cdf14579cd294e5079e0745411-Abstract.html.

[26] Judea Pearl. *Causality: Models, Reasoning and Inference*. Cambridge University Press, 2nd edition, 2009.

[27] Daniel Rivas-Barragan, Sarah Mubeen, Francesc Guim Bernat, Martin Hofmann-Apitius, and Daniel Domingo-Fernández. Drug2ways: Reasoning over causal paths in biological networks for drug discovery. *PLoS computational biology*, 16(12): e1008464, 2020.

[28] Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021. doi: 10.1109/JPROC.2021.3058954.

[29] Xinwei Shen, Furui Liu, Hanze Dong, Qing Lian, Zhitang Chen, and Tong Zhang. Weakly supervised disentangled generative causal representation learning. *Journal of Machine Learning Research*, 23(241):1–55, 2022. URL http://jmlr.org/papers/v23/21-0080.html.

[30] Kihyuk Sohn, Honglak Lee, and Xinchen Yan. Learning structured output representation using deep conditional generative models. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015. URL https://proceedings.neurips.cc/paper/2015/file/8d55a249e6baa5c06772297520da2051-Paper.pdf.

[31] Pater Spirtes, Clark Glymour, Richard Scheines, Stuart Kauffman, Valerio Aimale, and Frank Wimberly. Constructing bayesian network models of gene expression networks from microarray data. 2000.

[32] Xinwei Sun, Botong Wu, Xiangyu Zheng, Chang Liu, Wei Chen, Tao Qin, and Tie-Yan Liu. Recovering latent causal factor for generalization to distributional shifts. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 16846–16859. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper/2021/file/8c6744c9d42ec2cb9e8885b54ff744d0-Paper.pdf.

[33] Julius von Kügelgen, Yash Sharma, Luigi Gresele, Wieland Brendel, Bernhard Schölkopf, Michel Besserve, and Francesco Locatello. Self-supervised learning with data augmentations provably isolates content from style. In *Advances in Neural Information Processing Systems*, 2021.

[34] Lan Wang and Vishnu Naresh Boddeti. Do learned representations respect causal relationships? In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[35] Mengyue Yang, Furui Liu, Zhitang Chen, Xinwei Shen, Jianye Hao, and Jun Wang. Causalvae: Disentangled representation learning via neural structural causal models. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 9593–9602. Computer Vision Foundation / IEEE, 2021. URL https://openaccess.thecvf.com/content/CVPR2021/html/Yang_CausalVAE_Disentangled_Representation_Learning_via_Neural_Structural_Causal_Models_CVPR_2021_paper.html.

[36] Weiran Yao, Yuewen Sun, Alex Ho, Changyin Sun, and Kun Zhang. Learning temporally causal latent processes from general temporal data. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=RDlLMjLJXdq.

[37] Yue Yu, Jie Chen, Tian Gao, and Mo Yu. DAG-GNN: DAG structure learning with graph neural networks. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 7154–7163. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/yu19a.html.

[38] Xun Zheng, Bryon Aragam, Pradeep Ravikumar, and Eric P. Xing. DAGs with NO TEARS: Continuous Optimization for Structure Learning. In *Advances in Neural Information Processing Systems*, 2018.

[39] Jiageng Zhu, Hanchen Xie, and Wael AbdAlmgaeed. Do-operation guided causal representation learning with reduced supervision strength. In *NeurIPS 2022 Workshop on Causality for Real-world Impact*, 2022. URL https://openreview.net/forum?id=KbjUEXmlKWJ.

[40] Shengyu Zhu, Ignavier Ng, and Zhitang Chen. Causal discovery with reinforcement learning. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=S1g2skStPB.