Frontiers of Information Technology & Electronic Engineering
www.jzus.zju.edu.cn; engineering.cae.cn; www.springerlink.com
ISSN 2095-9184 (print); ISSN 2095-9230 (online)
E-mail: jzus@zju.edu.cn

Controllable image generation based on causal representation learning*

Shanshan HUANG†1, Yuanhao WANG[1], Zhili GONG[1], Jun LIAO[1], Shu WANG[2], Li LIU†‡1

[1]*School of Big Data and Software Engineering, Chongqing University, Chongqing 401331, China*

[2]*School of Materials and Energy, Southwest University, Chongqing 400715, China*

†E-mail: shanshanhuang@cqu.edu.cn; dcsliuli@cqu.edu.cn

Received May 5, 2023; Revision accepted Oct. 13, 2023; Crosschecked Dec. 19, 2023

Abstract: Artificial intelligence generated content (AIGC) has emerged as an indispensable tool for producing large-scale content in various forms, such as images, thanks to the significant role that AI plays in imitation and production. However, interpretability and controllability remain challenges. Existing AI methods often face challenges in producing images that are both flexible and controllable while considering causal relationships within the images. To address this issue, we have developed a novel method for causal controllable image generation (CCIG) that combines causal representation learning with bi-directional generative adversarial networks (GANs). This approach enables humans to control image attributes while considering the rationality and interpretability of the generated images and also allows for the generation of counterfactual images. The key of our approach, CCIG, lies in the use of a causal structure learning module to learn the causal relationships between image attributes and joint optimization with the encoder, generator, and joint discriminator in the image generation module. By doing so, we can learn causal representations in image's latent space and use causal intervention operations to control image generation. We conduct extensive experiments on a real-world dataset, CelebA. The experimental results illustrate the effectiveness of CCIG.

Key words: Image generation; Controllable image editing; Causal structure learning; Causal representation learning

https://doi.org/10.1631/FITEE.2300303 **CLC number:** TP391.41

1 Introduction

Artificial intelligence generated content (AIGC) has been gradually changing the production of digital content, especially in the image field, and has been widely used for a variety of tasks such as image generation, image editing, image reconstruction, and image restoration (Huang SS et al., 2022). With the development of the AIGC technique, the controllable image generation method has been applied in many fields, including face beautification, virtual try-on, and industrial design. Existing controllable image generation methods (Huang S et al., 2023; Zhang LM et al., 2023) achieve control over the generated images by inputting different conditional information, including label information (e.g., image attributes and semantic segmentation maps), visual information (e.g., sketches, low-resolution images, and image blocks), and text information (i.e., text descriptions), also known as conditional image generation methods. Several approaches (Shen YJ and Zhou, 2021; Shen YJ et al., 2022) are also available to find interpretable, semantically meaningful directions by exploring the latent space of generative adversarial

‡ Corresponding author

* Project supported by the National Major Science and Technology Projects of China (No. 2022YFB3303302), the National Natural Science Foundation of China (Nos. 61977012 and 62207007), and the Central Universities Project in China at Chongqing University (Nos. 2021CDJYGRH011 and 2020CDJSK06PT14)

ORCID: Shanshan HUANG, https://orcid.org/0000-0001-7893-3861; Li LIU, https://orcid.org/0000-0002-4776-5292

networks (GANs), allowing controllable image generation by varying the latent code of desired directions.

Although these methods enable control over the generated images and have yielded many exciting results, there are still several issues: conditional image generation methods require independent sampling of the labels, ignoring the dependencies between labels. The generated image can also be affected by spurious correlations, which can lead to an irrational result; for example, changing the "age" label leads to the change of "glasses," which is obviously not possible. Similarly, another class of approaches, GAN inversion based solutions (Shen YJ et al., 2022; Xia et al., 2023), often assume that the semantic factors can correspond to the latent variables one by one; that is, each dimension of latent encoding measures a different and mutually independent variable factor in the data. However, in practical applications, the semantic meaningful factors of interest are not independent of each other. Instead, there may be an underlying causal structure that makes these semantically meaningful factors interdependent; for example, smile causes mouth opening and eye narrowing.

Fortunately, there have been recent efforts to explore causal learning based methods for controllable image generation, which have shown promising results. Some of these methods (Kocaoglu et al., 2018; Moraffah et al., 2020; Shen XW et al., 2022) learn causal generative models for controlled image generation by given labeled causal graphs; however, obtaining such graphs can be challenging in realistic situations. There are also methods (Yang et al., 2021; Zhu JG et al., 2022) that introduce structural causal models (SCMs) into generative models, learn causal relationships between underlying factors from the data, and use the learned causal relationships to provide interpretability for the generated images. Unfortunately, the quality of the images generated from these methods is often unsatisfactory due to their heavy reliance on variational autoencoder (VAE) models.

To tackle the aforementioned issues, this article presents a causal controllable image generation method by integrating causal structure learning (CSL) with bi-directional GANs. This method, on one hand, avoids the requirement for providing a causal graph beforehand, but instead learns the causal relationships among image attributes from the data. On the other hand, the introduction of bi-directional GANs can improve the image representation learning capability and enhance the fidelity and resolution of the generated images. In summary, the key contributions of this article are as follows:

1. We propose a novel causal controllable image generation (CCIG) framework, which combines a CSL module with an image generation module (IGM), to learn causal graphs of image attributes, and the learned causal graphs are used to constrain the latent representations to understand the image generation mechanism.

2. We propose a bi-directional generative model for image generation and representation learning, which combines an encoder, a generator, and a joint discriminator (JointD) to improve the model representation learning capability, and integrates the attention mechanism and residual structure into both the generator and JointD to enhance the quality of the generated images.

3. We instantiate the proposed framework and conduct extensive experiments over a public dataset, CelebA, verifying the effectiveness and rationality of CCIG.

2 Related works

In this section, we provide a brief review of the latest advances in the fields of causal structure learning, causal representation learning, and controllable image generation.

2.1 Causal structure learning

Causal structure learning is a fundamental task across various scientific disciplines and has been widely used in medical diagnosis, policy-making, social sciences, computer science, and many other fields (Pan et al., 2022; Sun et al., 2022). Many researchers have developed various methods, which can be divided roughly into four categories: score-based, constraint-based, function causal model (FCM) based, and continuous optimization based methods. In particular, with the development of deep learning, continuous optimization based methods have been extensively studied, and the increasing computational power has made it feasible to learn causal structures on large-scale high-dimensional datasets. A growing number of methods that fuse black-box deep learning models and structural learning are transforming the combinatorial graph search

problem back into a continuous optimization problem (Zheng et al., 2018; Yu et al., 2019; Gao et al., 2021; Petkov et al., 2022). NoTEARS (Zheng et al., 2018) is usually considered the pioneering method. This method focuses only on the linear structural equation model (SEM), has a high time complexity for acyclic constraints, and is not applicable to the learning of large-scale causal graphs. To address these issues, DAG-GNN (Yu et al., 2019) extends NoTEARS by incorporating neural network functions and black-box variational inference, and uses the evidence lower bound (ELBO) as a score function to learn the directed acyclic graph (DAG). Ng et al. (2019) proposed a causal framework based on graph autoencoders to support nonlinear SEM and vector-valued variables. Furthermore, Wei et al. (2020) revisited various aspects of DAG-GNN and NoTEARS, and refined the aforementioned methods with acyclic constraints. Similar methods are available in the literature (Lachapelle et al., 2020; Varando, 2020; Gao et al., 2021; Ng et al., 2022; Petkov et al., 2022). In addition, there are methods (Zhu SY et al., 2020; Wang XQ et al., 2021) that combine reinforcement learning with causal structure learning. More related work can be found in Vowels et al. (2023).

2.2 Causal representation learning

The assumption of traditional causal structure learning is that the causal units are random variables connected by the causal graph. In the real world, observations are often not composed of these units, such as objectives in images (Lopez-Paz et al., 2017). Emerging causal representation learning (Schölkopf et al., 2021) endeavors to learn these variables from data and to apply them to downstream tasks. These methods have combined SCMs with representation learning. For example, Leeb et al. (2020) proposed a structured encoder embedded with SCMs to automatically learn the hierarchical structure of disentangled factors. Lu et al. (2021) learned causal representations by observing similar causal models in different environments. Lippe et al. (2022) learned causal representations from time-series data of labeled interventions, assuming that causal effects are not instantaneous but can be resolved over time. There are also methods (Brehmer et al., 2022; Lv et al., 2022; Wang WJ et al., 2022; Ahuja et al., 2023) that use causal representation learning to address domain generalization problems, including image classification and recommendation systems.

2.3 Controllable image generation

Controllable image generation is a method that uses deep generative models along with input conditions, to control the attributes of the generated images. The focus of this study is causal learning based controllable image generation methods. Unlike other approaches, causal controllable image generation methods leverage SCMs and deep generative models to enable intervention and counterfactual queries, offering vital insights to address the current challenges of lacking robustness, interpretability, and fairness in generative models. Motivated by this, numerous work has been conducted, achieving promising results, including VAEs-based (Suter et al., 2019; Reinhold et al., 2021; Yang et al., 2021; Lai, 2022; Zhu JG et al., 2022), GANs-based (Kocaoglu et al., 2018; Moraffah et al., 2020; Zhang XH et al., 2021; Shen XW et al., 2022; Zhang WB et al., 2022), and diffusion-based methods (Sanchez and Tsaftaris, 2022; Sanchez et al., 2022).

CausalGAN is the first method that combines SCMs with GANs, which provides a theoretical guarantee for sampling from the interventional distribution. Similarly, Sauer and Geiger (2021) embedded SCMs into the generator to control the generation of counterfactual images as augmented data for invariant classifier training. In contrast to these methods that require a complete causal graph as a prior, DEAR (Shen XW et al., 2022) requires only causal variables or their causal order. Furthermore, Yang et al. (2021) achieved causal controllable image generation by automatically learning the causal structure using the adjacency matrix. Instead of explicitly encoding SCMs, Reinhold et al. (2021) and Lai (2022) used VAEs to learn the structure assignment for generating counterfactual images. In addition, recent studies (Augustin et al., 2022; Sanchez and Tsaftaris, 2022) have shown that combining generative diffusion models with SCMs can achieve counterfactual image generation and provide counterfactual explanations. In our work, we propose an end-to-end controllable image generation method that combines causal structure learning with the bidirectional generative model. Unlike DEAR, our method, CCIG, does not require a causal order or a causal graph of the variables given as a prior.

3 Method

This section presents the problem definition and the proposed framework, CCIG.

3.1 Problem definition

3.1.1 Generative model

The goal of the generative model is to learn a generator $G_\phi : z \to x$ that can produce user-controllable images. This model is parameterized by ϕ, where $z \in \mathbb{R}^k$ is a latent code sampled from a prior distribution p_z, and the encoder $E_\vartheta : x \to z$, parameterized by ϑ, is used to learn high-level semantic representations of images. Formally, the optimization objective of the generative model is to maximize the likelihood of the observed image by minimizing the Kullback–Leibler (KL) divergence between the encoder joint distribution $p_{\mathrm{e}}(x,z) = p_x(x)p_{\mathrm{e}}(z|x)$ and the generator joint distribution $p_{\mathrm{g}}(x,z) = p_z(z)p_{\mathrm{g}}(x|z)$, as follows:

$$\min_{G,E} \mathcal{L}(p_{\mathrm{e}}(x,z), p_{\mathrm{g}}(x,z)), \tag{1}$$

which is equivalent to maximizing ELBO using variational inference or adversarial training. So, Eq. (1) can be rewritten as follows:

$$\begin{aligned}\mathcal{L}(p_{\mathrm{e}}(x,z), p_{\mathrm{g}}(x,z)) = -\mathbb{E}_{z\sim p_{\mathrm{e}}(z|x)}[\log p_{\mathrm{g}}(x \mid z)] \\ -\mathbb{E}_{x\sim p_x}[D_{\mathrm{KL}}(p_{\mathrm{e}}(z \mid x), p_z(z))] + \mathbb{E}_{x\sim p_x}[\log p_x(x)],\end{aligned} \tag{2}$$

where $\mathbb{E}_{x\sim p_x}[\log p_x(x)]$ is free of parameters. Furthermore, the supervision signals (e.g., true factor labels) are used to train the generative model to acquire the meaningful disentangled image representations. $\zeta \sim p_\zeta \in \mathbb{R}^m$ denotes the underlying ground-truth factor in image x and l denotes the observation (e.g., the image attribute label) corresponding to ζ, which can be discrete or continuous. Without loss of generality, $\zeta_i = \mathbb{E}(l_i \mid x), i = 1, 2, ..., m$. For example, l_1 could be a binary label denoting the gender of a person, while $\zeta_1 = \mathbb{E}(l_1|x) = P(l_1 = 1|x)$ is the possibility that a person is male given an image x. Therefore, the following supervised loss $\mathcal{L}_{\mathrm{sup}}(E)$ will be used to optimize the generative model along with the generative loss:

$$\mathcal{L}_{G,E} = \mathcal{L}(p_{\mathrm{e}}(x,z), p_{\mathrm{g}}(x,z)) + \lambda \mathcal{L}_{\mathrm{sup}}(E), \tag{3}$$

where λ represents the weight of the supervised loss and is a constant. Unlike conditional GANs where supervised labels involve GAN losses, the unsupervised generative modeling losses and supervised regularizers in Eq. (3) are decoupled in terms of taking expectations. Specifically,

$$\mathcal{L}_{\mathrm{sup}}(E) = \mathbb{E}_{x,l}[l_{\mathrm{s}}(E; x, l)], \tag{4}$$

where l_{s} can be a cross-entropy (CE) loss or L2 loss.

3.1.2 Discovery of causal relationships in images

In an ideal scenario, the causal relationships between different attributes should be considered in the image generation process. However, due to the presence of data selection bias, the generative model often learns some spurious correlations. To pursue a more interpretable and reasonable image generation process, we consider the discovery of causal relationships in images as another goal of controllable image generation.

Causal structure learning is an emerging technique for discovering causal relationships, which is believed to be a key to next-generation artificial intelligence with powerful cross-domain performance, transfer learning, and interpretability. It can be formalized as follows: let $l = [l_1, l_2, \cdots, l_m]^{\mathrm{T}} \in \mathbb{R}^{m\times d}$ as a d-dimensional vector with m variables, and let $\mathcal{G} = (V, F)$ be a DAG, representing the image generation mechanism of l. Here, V and F are the node set and edge set, respectively. $h = [h_1, h_2, \cdots, h_m]^{\mathrm{T}} \in \mathbb{R}^{m\times d}$ denotes random noise, and $A = [A_1, A_2, \cdots, A_m] \in \mathbb{R}^{m\times m}$ is an adjacency matrix encoding the causal mechanisms inside l, where $A_{ij} \neq 0$ if l_i is the parent node of l_j and $A_{ij} = 0$ otherwise.

The causal effect signified by a DAG can be quantified by a linear SEM as follows:

$$l = A^{\mathrm{T}} l + h. \tag{5}$$

When the nodes of the graph are sorted in topological order, and matrix A is strictly upper triangular. Hence, ancestral sampling from the DAG is equivalent to generating a random noise h followed by a triangular solver:

$$l = (I - A^{\mathrm{T}})^{-1} h. \tag{6}$$

In our case, we generalize the previous linear SEM to a nonlinear form using a graph autoencoder, as described in Section 3.2. Briefly, to learn the optimal

causal structure of image attributes, the score function with acyclic constraints should be minimized:

$$\min_{A\in\mathbb{R}^{m\times m}} S(A) \quad \text{s.t.} \quad c(A)=0, \tag{7}$$

where $c(A)=0$ is the acyclic constraint that guarantees the acyclicity of the learned causal graph $\mathcal{G}$, and $S(A)$ is the score function, which can usually be calculated by the MSE loss function (Wang YF et al., 2021), regularized negative maximum likelihood function (Lachapelle et al., 2020), or ELBO (Yu et al., 2019). In our case, ELBO is used as the score function, which will be introduced in Section 3.2.

3.2 CCIG framework

In this subsection, we present the proposed framework, CCIG, which consists of three main components: the CSL module for learning causal relationships between image attributes, the IGM for learning image representations and generating images, and the loss function for optimizing the model. Fig. 1 illustrates the workflow of our framework.

3.2.1 Causal structure learning module

The CSL module aims to learn the causal relationships between image attributes to support image causal representation learning of IGM. Inspired by Yu et al. (2019), we use the graph autoencoder to explicitly model the causal structure of image attributes, as shown in Fig. 2. The encoder infers the latent vector h by performing a reparameterized sampling trick on a Gaussian distribution with mean $M_h \in \mathbb{R}^{m\times d}$ and standard deviation $V_h \in \mathbb{R}^{m\times d}$, which can be formulated as follows:

$$\text{Encoder}_l : h = f_2((I-A^{\mathrm{T}})f_1(l)), \tag{8}$$

where f_1 and f_2 are parameterized functions that perform transformations (linear or nonlinear transformations) on l and h, respectively. In our case, f_1 is the multilayer perceptron (MLP) and f_2 is the identity mapping (IM). Then, the CSL module recovers the observations with a decoder, as follows:

$$\text{Decoder}_l : l = f_4((I-A^{\mathrm{T}})^{-1}f_3(h)), \tag{9}$$

where f_3 and f_4 are also parameterized functions that conceptually inverse f_1 and f_2, respectively. In our case, f_3 and f_4 are implemented by IM and MLP, respectively. $M_l \in \mathbb{R}^{m\times d}$ and $V_l \in \mathbb{R}^{m\times d}$ are obtained from Decoder_l, and represent the mean and standard deviation of the likelihood $p(l|h)$ that obeys the Gaussian distribution, respectively. Therefore, Eq. (9) can be rewritten as follows:

$$f_4^{-1}(l) = A^{\mathrm{T}}f_4^{-1}(l) + f_3(h). \tag{10}$$

To obtain a more accurate latent representation h, the following reconstruction loss is used as the

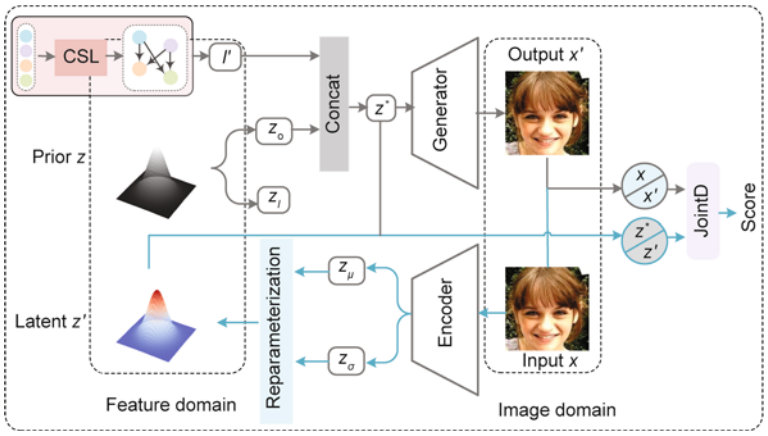

Fig. 1 Overall workflow of our framework. In the training phase, our model first uses the CSL module to learn causality from the image labels, thereby constraining the underlying representation $z'=E(x)$ obtained from the encoder to be consistent with the learned causal relationships. Then, the causal representation l', which encodes causal relationships between factors of interest in image generation, is used to replace a portion z_l of the prior distribution z, which is then concatenated with the other factors z_o needed for image generation to obtain a new latent representation z^*. This representation is then fed into the generator G to obtain the generated image $x'=G(z^*)$. Note that JointD is trained alternatively with the generator G and encoder E. In the inference phase, we can achieve causal controllability of an image attribute by modifying the value of a latent representation along a specific dimension, i.e., causal intervention

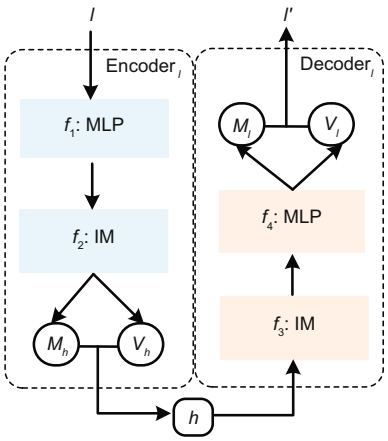

Fig. 2 Workflow of the CSL module, which consists of two components: Encoder$_l$ and Decoder$_l$. Encoder$_l$ acts as an inference model designed to encode the input attribute l as the latent posterior h. Decoder$_l$ acts as the causal generative model for reconstructing the input attributes

optimization objective of the CSL module:

$$\mathcal{L}_{\mathrm{r}} = \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{d} l_{ij}\log(M_l)_{ij}. \tag{11}$$

Furthermore, a regularizer term is added to reconstruction loss to avoid overfitting, as follows:

$$r_l = \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{d}(M_h)_{ij}^2. \tag{12}$$

To ensure the acyclicity of the learned causal graphs, we add the following acyclic constraints:

$$c(A) : \mathrm{tr}[(I+\alpha A\circ A)^m] - m = 0, \tag{13}$$

where "$\circ$" denotes the Hadamard product.

Overall, the objective of CSL consists of the reconstruction error, the regularizer, and the acyclicity constraint. Thus, problem (7) can be rewritten as follows:

$$\min_{A,\varphi}(\mathcal{L}_{\mathrm{r}} + r_l) \quad \text{s.t.} \quad c(A) = 0. \tag{14}$$

To optimize the above objective, we follow Yu et al. (2019) and adopt the augmented Lagrangian method, which introduces a quadratic penalty term to problem (14) and can be rewritten as

$$\mathcal{L}_\nu(A,\varphi,\kappa) = \mathcal{L}_{\mathrm{r}} + r_l + \kappa c(A) + \frac{\nu}{2}|c(A)|^2, \tag{15}$$

where φ is the parameter of the CSL module, and κ and ν are the Lagrange multiplier and penalty parameter, respectively. We solve this optimization problem by iteratively updating κ and ν, following the strategy in Yu et al. (2019).

3.2.2 Image generation module

As shown in Fig. 1, IGM takes the representation generated by CSL that encodes the causal relationships of image attributes and outputs the generated images.

Specifically, IGM consists of three components, namely, the encoder E for learning high-level causal semantic representations, the generator G for generating user-controllable images, and JointD for distinguishing real-data distribution-latent representation pairs from generated data distribution-prior distributions. These three components play with each other until they reach a global optimum, formally:

$$\min_{E,G}\max_{D}\mathbb{E}_{x\sim p_x, z\sim p_{\mathrm{e}}(z|x)}[\log D(x,E(x))] + \mathbb{E}_{z\sim p_z, x\sim p_{\mathrm{g}}(x|z)}[\log(1-D(G(z^*),z^*))],$$

where z^* represents a hybrid representation that integrates representations sampled from the prior distribution with latent representations obtained by the CSL module that encodes the causal relationships between image attributes. In other words, z^* consists of an m-dimensional causal representation that encodes m interested generative factors' causal relationships and a $(k-m)$-dimensional representation that encodes other factors required for image generation (sampled from the prior distribution of z or $E(x)$). Specifically, the generator–encoder loss (3) can be rewritten as follows:

$$\begin{aligned}\mathcal{L}_{G,E} =& \mathbb{E}_{x\sim p_x, z\sim p_{\mathrm{e}}(z|x)}[l_{G,E}(x,E(x))] \\ &-\mathbb{E}_{z\sim p_z, x\sim p_{\mathrm{g}}(x|z)}[l_{G,E}(G(z^*),z^*)] + \lambda\mathcal{L}_{\mathrm{sup}}(E),\end{aligned} \tag{16}$$

and the discriminator loss is given as follows:

$$\begin{aligned}\mathcal{L}_D =& \mathbb{E}_{x\sim p_x, z\sim p_{\mathrm{e}}(z|x)}[l_D(x,E(x),1)] \\ &+ \mathbb{E}_{z\sim p_z, x\sim p_{\mathrm{g}}(x|z)}[l_D(G(z^*),z^*,-1)],\end{aligned} \tag{17}$$

where $l_{G,E}(x,z) = s_x(x) + s_z(z) + s_{xz}(x,z)$ and $l_D(x,z,\tau) = h(\tau s_x) + h(\tau s_z) + h(\tau s_{xz})$, $\tau\in\{-1,1\}$. Here, $h(t) = \max(0, 1-t)$ is a "hinge" used to regularize the discriminator, and s_x, s_z, and s_{xz} are the learned projections of the discriminators D_x, D_z, and D_{xz}, respectively:

$$\begin{cases} s_x(x) = \theta_x^{\mathrm{T}} D_x^{\Theta}(x), \\ s_z(z) = \theta_z^{\mathrm{T}} D_z^{\Theta}(z), \\ s_{xz}(x,z) = \theta_{xz}^{\mathrm{T}} D_{xz}^{\Theta}(D_x^{\Theta}(x) + D_z^{\Theta}(z)), \end{cases} \tag{18}$$

where D_x, D_z, and D_{zx} are the three key components of JointD, as shown in Fig. 3, each characterized by

the corresponding parameters θ_x, θ_z, and θ_{xz}. The role of D_x is to discriminate between the generated image and the real image, while D_z is responsible for differentiating the latent representation z^* from the output of the encoder $E(x)$. Additionally, D_{zx} is a function that takes into account the outputs f_x and f_z of both D_x and D_z. We design our discriminator in this way as generative models require adversarial learning and optimization in both image and feature domains to accurately learn the underlying data distribution.

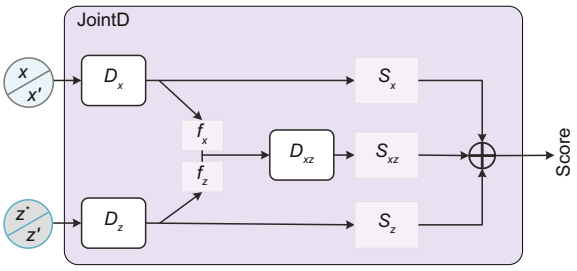

Fig. 3 Workflow of JointD

Overall, the generator and encoder are trained by $\mathcal{L}_{G,E}$ to force them to create matching joint data-latent distributions that JointD cannot predict correctly. JointD is trained with $\mathcal{L}_D$ to correctly distinguish between two joint data-latent distributions from the encoder and the generator, that is, $(x, E(x))$ and $(G(z^*), z^*)$, respectively.

3.3 Model structure

In this subsection, we introduce the model structure of IGM, including the generator, encoder, and discriminator. Note that details of CSL are introduced in Section 3.2.1.

3.3.1 Encoder

We use ResNet50 (He et al., 2016) as our encoder to encode the image into the latent space (i.e., feature domain) to obtain the latent representation of the input image.

3.3.2 Generator

The generator architecture of our proposed framework consists of a GenInBlock, followed by two GenBlocks, an AttentionBlock, and a GenBlock, with batch normalization, rectified linear unit (ReLU) activation, SNConv2d, and Tanh as activation functions. The detailed structures of GenInBlock, GenBlock, and AttentionBlock can be found in Figs. 4b–4d. GenBlock adopts a residual structure to assist the generative model in adapting to the complex data distribution, ultimately improving the quality of the generated images. AttentionBlock introduces an attention mechanism, which assigns different weights to the features at different locations in the image, enhancing the detailed information of the generated images. Moreover, the NoiseInjection layer in both GenInBlock and GenBlock improves the expressiveness of the generative model by injecting noise.

3.3.3 Discriminator

The proposed JointD is composed of three distinct components, namely, D_x, D_z, and D_{xz}, as depicted in Fig. 3. Specifically, D_x is implemented as a convolutional neural network (ConvNet), as shown in Fig. 5a, while D_z and D_{xz} have identical MLP structures.

4 Experiments

To evaluate the proposed CCIG framework, we conduct extensive experiments to answer the following research questions:

RQ1: How does CCIG perform compared with state-of-the-art methods?

RQ2: How does the proposed CSL module impact the interpretability and controllability of the generated images? Does the causal matrix learned in the CSL module conform to human common sense?

RQ3: Do different supervised loss functions affect the quality of the generated images?

RQ4: Do key modules of the discriminator and generator affect the quality of the generated images?

4.1 Experimental setup

4.1.1 Parameter setting

In our implementation, we pre-process the images by taking center crops of 128×128 for CelebA and resizing all images to the 64×64 resolution. We use PyTorch for our implementation and the Adam optimizer with hyperparameters $\beta_1 = 0$, $\beta_2 = 0.999$, and learning rates 1e-3, 5e-5, 5e-5, and 1e-4 for CSL, generator, encoder, and JointD, respectively. The proposed model is trained using a mini-batch size of 64. We employ CE loss as the supervised loss and

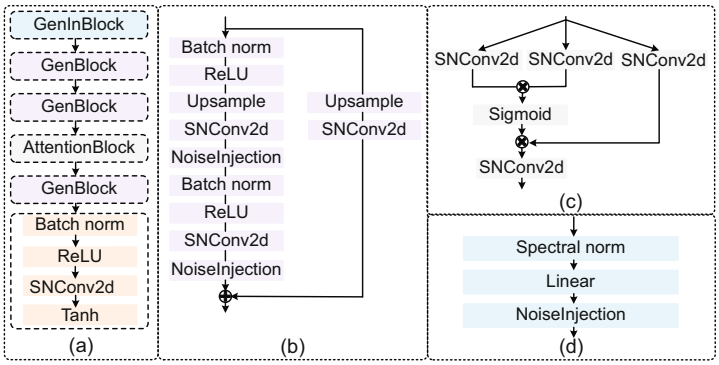

Fig. 4 Details of the generator: (a) generator; (b) GenBlock; (c) AttentionBlock; (d) GenInBlock

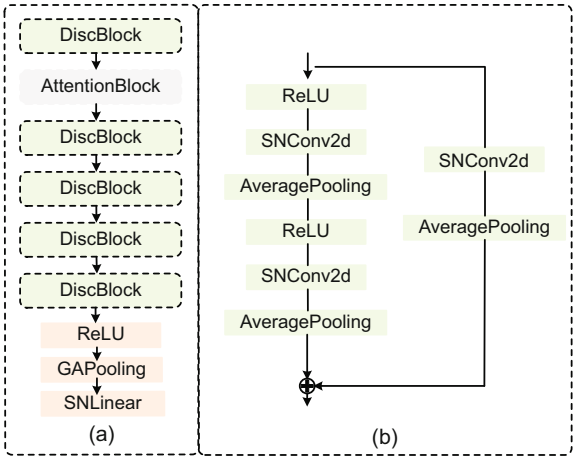

Fig. 5 Detail of discriminator D_x: (a) discriminator; (b) DiscBlock

set the weight λ of this loss to 5. When visualizing the causal matrix, we follow the recommendation of Zheng et al. (2018) and use a thresholding value of 0.3. We train the models for around 150 epochs on the CelebA dataset, including CelebA (smile), CelebA (age), and CelebA (gender) using NVIDIA RTX 2080Ti.

4.1.2 Dataset and evaluation metrics

1. Dataset

We validate the proposed method on the real dataset CelebA (Liu et al., 2015), which contains a total of 202 599 face images with 40 different concepts. Specifically, we select three subsets of causally related attributes, namely, CelebA (smile), CelebA (gender), and CelebA (age). The first dataset contains six attributes: gender (male), smile, narrow eye, mouth open, high cheekbone, and chubby. The second dataset focuses on five attributes: age (young), gender, bald, beard, and mustache. The last dataset focuses on six attributes: age, gender, bag under eye, chubby, heavy makeup, and receding hairline.

2. Evaluation metrics

To evaluate the quality of the generated images by the proposed method, we employ a comprehensive evaluation approach that combines subjective and objective measures, including a subjective evaluation method called the mean opinion score (MOS), as well as two objective image quality evaluation metrics: Fréchet inception distance (FID) (Heusel et al., 2017) and inception score (IS) (Salimans et al., 2016).

4.2 Performance comparison (RQ1)

To demonstrate the effectiveness of CCIG, we compare our method with four baseline methods, namely, CausalGAN (Kocaoglu et al., 2018), CausalVAE (Yang et al., 2021), CMGAN (Zhang WB et al., 2022), and DEAR (Shen XW et al., 2022). The experimental results are shown in Table 1 and Figs. 6–15. We have the following observations:

1. From a subjective perspective, our method, CCIG, not only outperforms the other methods in terms of visual perception quality of the generated images, as shown in the two last columns in Table 1, but also produces more plausible images than the other methods, as illustrated in Figs. 6–15. Indeed, our approach's success may be attributed to two key factors: the IGM used in our method is a bi-directional GAN, which has a strong capability of representation learning and high-quality image generation, and the incorporation of the CSL module is equally significant. This module ensures that the

Table 1 Performance comparison with the baseline methods

Method	IS_MEAN	IS_STD	FID	MOS*	MOS**
CausalVAE	1.3659	0.0694	284.41	4.16	4.52
CMGAN	1.3392	0.0412	218.12	3.52	3.86
CausalGAN	1.8314	**0.0157**	144.16	3.20	4.15
DEAR	2.0140	0.0232	97.16	4.58	4.03
CCIG (L2)	2.0360	0.0214	96.89	4.43	4.60
CCIG (CE)	**2.1461**	0.0347	**96.71**	**4.66**	**4.62**

* Image quality; ** controllability of image editing. FID: Fréchet inception distance; MOS: mean opinion score. IS_MEAN and IS_STD represent the mean and standard deviation of the inception score (IS), respectively. Best results are in bold

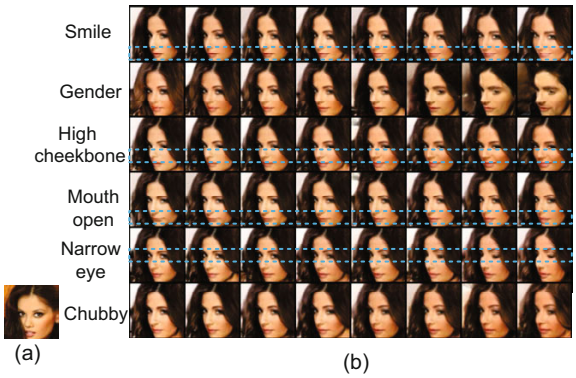

Fig. 6 Results of CCIG in causal controllable generation on the CelebA (smile) dataset: (a) source image; (b) results. For each row, we change only one latent factor and fix all other latent factors. By changing the cause factor (smile or gender), we observe corresponding changes in the effect factors (mouth open and narrow eye). Conversely, by changing the effect factor (mouth open, high cheekbone, narrow eye, or chubby), the reconstruction can become a counterfactual image, while the cause factor remains unchanged

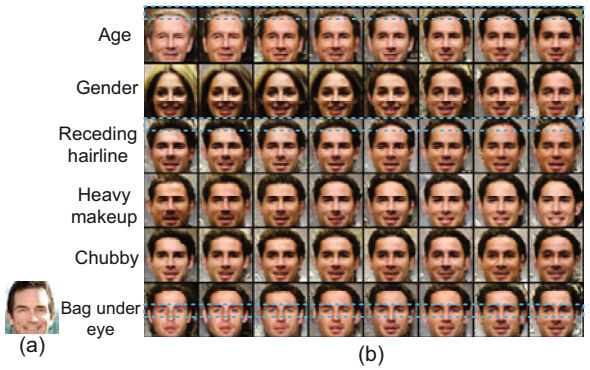

Fig. 7 Results of CCIG in causal controllable generation on the CelebA (age) dataset: (a) source image; (b) results. For each row, we change only one latent factor and fix all other latent factors. By changing the cause factor (age or gender), we observe corresponding changes in the effect factors (receding hairline and bag under eye). Conversely, by changing the effect factor (receding hairline, heavy makeup, chubby, or bag under eye), the reconstruction can become a counterfactual image, while the cause factor remains unchanged

learned image representations effectively encode the causal generative mechanisms among the factors of interest within the generated images.

2. From the objective metric perspective, our proposed method shows superior performance on both FID and IS, which suggests that it has advantages in generating realistic and diverse images. Other methods tend to perform poorly on both FID and IS metrics, and the images generated by CausalVAE (Yang et al., 2021) look blurry due to the reliance on MSE or L1 loss functions to minimize the reconstruction loss between the real and the generated images during the training process. For CausalGAN (Kocaoglu et al., 2018), it introduces abrupt changes during the generation process, leading to a more disjointed appearance in the generated images, as shown in Fig. 12. For DEAR (Shen XW et al., 2022), the quality of the generated images is comparable to that of CCIG but still slightly poorer than

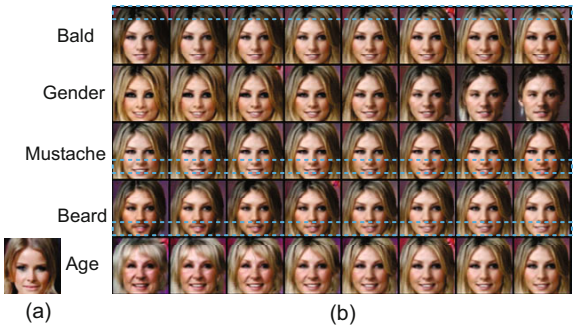

Fig. 8 Results of CCIG in causal controllable generation on the CelebA (gender) dataset: (a) source image; (b) results. For each row, we change only one latent factor and fix all other latent factors. By changing the cause factor (gender or age), we observe corresponding changes in the effect factors (bald, mustache, and beard). By changing the effect factor (bald, mustache, or beard), the reconstruction can become a counterfactual image, while the cause factor remains unchanged

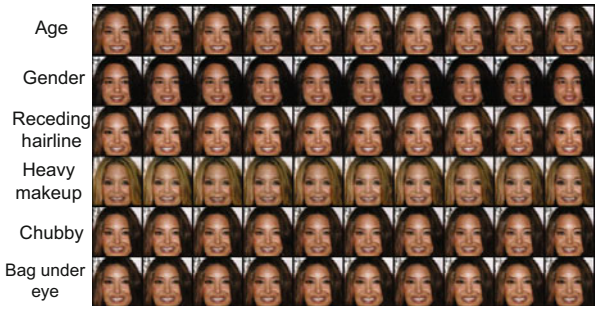

Fig. 9 Results of DEAR (Shen XW et al., 2022) in causal controllable generation on the CelebA (age) dataset

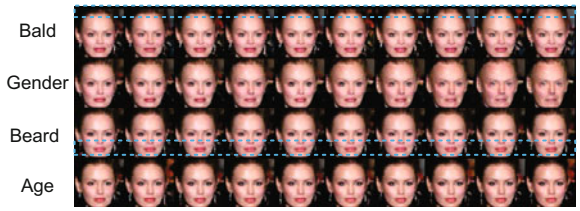

Fig. 10 Results of DEAR (Shen XW et al., 2022) in causal controllable generation on the CelebA (gender) dataset

Fig. 11 Results of DEAR (Shen XW et al., 2022) in causal controllable generation on the CelebA (smile) dataset

Fig. 12 Results of CausalGAN (Kocaoglu et al., 2018) in causal controllable generation on the CelebA (smile) dataset

that of CCIG in terms of the FID metric, which may be due to the improved network structure used by the generator and discriminator. Moreover, there is no significant change in the generated images when intervening on the beard attribute for DEAR since it cannot always generate counterfactual images as

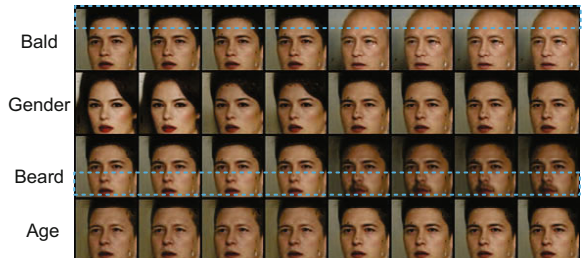

Fig. 13 Results of CausalGAN (Kocaoglu et al., 2018) in causal controllable generation on the CelebA (gender) dataset

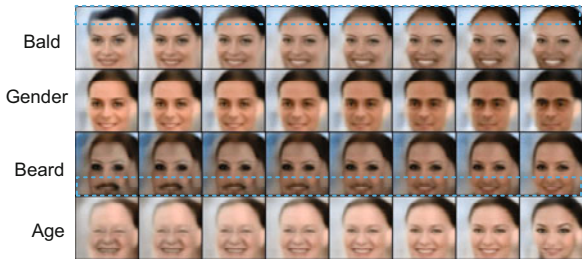

Fig. 14 Results of CausalVAE (Yang et al., 2021) in causal controllable generation on the CelebA (gender) dataset

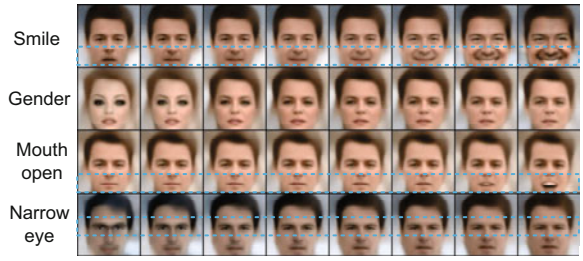

Fig. 15 Results of CausalVAE (Yang et al., 2021) in causal controllable generation on the CelebA (smile) dataset

shown in the third row of Fig. 10. Therefore, its generated image distribution is closer to the true image distribution, which results in better IS_STD values. Overall, our method still outperforms the other methods.

3. In terms of image controllability, our proposed method achieves higher levels of control over generated images. Specifically, our method induces changes in the resulting attribute when intervening on the causal attribute of an image, while leaving the causal attribute unchanged when modifying the resulting attribute. Other methods can also meet these conditions in some cases, such as CausalGAN (Kocaoglu et al., 2018). CausalGAN often generates images with non-smooth changes. Similarly, although the DEAR (Shen XW et al., 2022) method produces high-quality generated images, it may not always flexibly control the changes in image attributes

and produces only subtle and imperceptible changes at times. The controllability of CausalVAE (Yang et al., 2021) is similar to that of our method, but due to its use of VAE as a generative model, the quality of the generated images is significantly poorer than that of other methods.

4.3 Ablation study (RQ2–RQ4)

4.3.1 Study of the CSL module (RQ2)

We conduct experiments with or without the CSL module to verify the existence of causal relationships in image attributes and their impact on controllable image generation. In Fig. 16, we visualize the learned causal matrix of CSL on the three datasets. The corresponding true causal graphs are shown in Fig. 17. It can be observed that although there are certain redundant edges in the learned causal matrices, most of the learned causal relationships are plausible and in line with common sense. Fig. 18 presents the experimental results on the CelebA (smile) dataset with or without the CSL module. The results demonstrate that incorporating the CSL module yields the generated images that more accurately reconstruct the real images.

Furthermore, the CSL module enhances the controllability of the generated images by allowing for more precise manipulation of their variations. As shown in Fig. 18, when intervening on the smile attribute, the generated images without the CSL module exhibit abrupt changes, whereas those produced by the final model display gradual variations. Similar issues are observed for other attributes such as narrow eye and chubby.

Moreover, the introduction of the CSL module

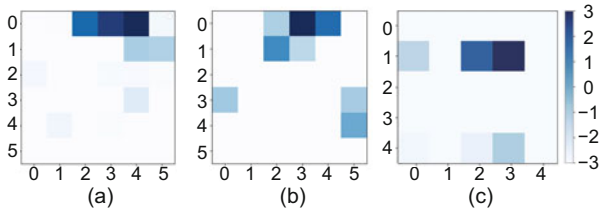

Fig. 16 Visualization of the learned causal matrix: (a) CelebA (smile) (0–5 correspond to smile, gender, high cheekbone, mouth open, narrow eye, and chubby, respectively); (b) CelebA (age) (0–5 correspond to age, gender, receding hairline, heavy makeup, chubby, and bag under eye, respectively); (c) CelebA (gender) (0–4 correspond to bald, gender, mustache, beard, and age, respectively)

leads to more reasonably generated images, as it ensures that changing the cause attribute results in corresponding changes in the result attribute, while keeping the cause attribute unchanged when changing the result attribute. Taking smile as an example, there exists a causal relationship among mouth open ← smile → narrow eye, where smile is the cause and narrow eye and mouth open are the effects. As shown in Fig. 18, changing the smile attribute always seems to have a consistent effect on the narrow eye attribute, no matter whether the CSL module is added. However, modifying the narrow eye attribute reveals that CCIG without the CSL module significantly alters the smile attribute, which is evidently irrational. The incorporation of the CSL module resolves this issue and ensures that a change in one attribute does not result in an unexpected change in another attribute. This further highlights the CSL module's ability to capture the causal relationships between image attributes and enable control of image attributes through causal intervention operations. Here, causal intervention refers to changing only a single dimension of a latent representation while fixing all other dimensions.

4.3.2 Study of the supervised loss L_{sup} (RQ3)

The quality of the images generated using different loss functions as supervised loss is shown in Table 1. The results show that the image quality obtained using either L2 or CE loss as the supervised loss is comparable. Overall, CE loss exhibits better performance, so we use CE loss as the final supervised loss to constrain the latent representation with supervision information (i.e., the labels).

4.3.3 Study of generator and discriminator structures (RQ4)

We verify the effectiveness of the modified generator and discriminator structures by ablating key blocks in the generators and discriminators. There are four variants in the experiments: the generator without the AttentionBlock (CCIG_GNA), the discriminator without the AttentionBlock (CCIG_DNA), the discriminator that considers only the output of the D_{xz} module (CCIG_DNs_{xz}), and the final method (CCIG). The experimental results are displayed in Table 2. We notice that the quality of the generated images is

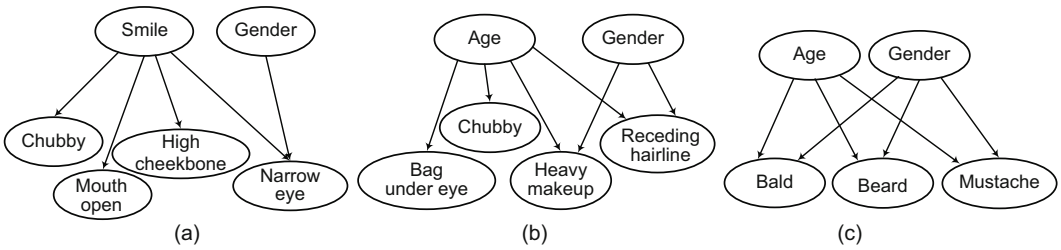

Fig. 17 True causal graph: (a) CelebA (smile); (b) CelebA (age); (c) CelebA (gender)

Fig. 18 Experimental results of the proposed method with or without the CSL module based on the CelebA (smile) dataset: (a) source images; (b) CCIG with the CSL module; (c) CCIG without the CSL module

Table 2 Results of ablation experiments

Method	IS_MEAN	IS_STD	FID	MOS*	MOS**
CCIG_GNA	**2.296**	0.0740	104.63	4.58	4.33
CCIG_DNA	2.1107	0.0649	100.77	4.40	4.37
CCIG_DNs_{xz}	1.9482	0.0482	102.36	4.33	4.43
CCIG (CE)	2.1461	**0.0347**	**96.71**	**4.66**	**4.62**

* Image quality; ** controllability of image editing. FID: Fréchet inception distance; MOS: mean opinion score. IS_MEAN and IS_STD represent the mean and standard deviation of the inception score (IS), respectively. Best results are in bold

significantly improved by adopting the AttentionBlock either in the generator or in the discriminator. This is due to the fact that the introduction of the AttentionBlock helps distribute the weights of the features more efficiently, allowing the model to pay more attention to the key details of the information in the image. In addition, when we consider only the output of D_{xz}, that is, when s_x and s_z are removed and only s_{xz} is considered, the experimental results show that CCIG significantly outperforms

CCIG_DNs_{xz} in terms of FID and IS of the generated images. This result is intuitively reasonable as it matches the standard generator loss. Overall, the quality of the generated images is significantly improved when the AttentionBlock is employed and the incorporation of JointD enhances the quality of the generated images, while improving the model representation capability by imposing joint constraints on the image-latent representation distribution.

5 Conclusions

In this article, we propose a novel framework, CCIG, for controllable image generation that uses causal representation learning. We adopt a CSL module to capture the causal relationships between image attributes and constrain the latent representation of the image. The CSL module is jointly optimized with the encoder, generator, and JointD in IGM. Through causal intervention operations on the latent representation, we can generate counterfactual images and perform real image editing. This work is a bold attempt to explore the interpretability and causal controllability of image generation. In the future, we will make more efforts to explore image causal representation and controllable generation, including but not limited to new representation generation mechanisms and more practical and efficient causal structure learning methods.

Contributors

Shanshan HUANG designed the research. Shanshan HUANG, Yuanhao WANG, and Zhili GONG processed the data. Shanshan HUANG drafted the paper. Yuanhao WANG, Jun LIAO, and Shu WANG helped organize the paper. Shanshan HUANG and Li LIU revised and finalized the paper.

Compliance with ethics guidelines

All the authors declare that they have no conflict of interest.

Data availability

The data that support the findings of this study are available from the corresponding author upon reasonable request.

References

Ahuja K, Mahajan D, Wang YX, et al., 2023. Interventional causal representation learning. Proc 43th Int Conf on Machine Learning, p.372-407.

Augustin M, Boreiko V, Croce F, et al., 2022. Diffusion visual counterfactual explanations. Proc 36th Advances in Neural Information Processing Systems, p.364-377.

Brehmer J, de Haan P, Lippe P, et al., 2022. Weakly supervised causal representation learning. Proc 36th Advances in Neural Information Processing Systems, p.38319-38331.

Gao YH, Shen L, Xia ST, 2021. DAG-GAN: causal structure learning with generative adversarial nets. Proc IEEE Int Conf on Acoustics, Speech and Signal Processing, p.3320-3324. https://doi.org/10.1109/ICASSP39728.2021.9414770

He KM, Zhang XY, Ren SQ, et al., 2016. Deep residual learning for image recognition. Proc IEEE Conf on Computer Vision and Pattern Recognition, p.770-778. https://doi.org/10.1109/CVPR.2016.90

Heusel M, Ramsauer H, Unterthiner T, et al., 2017. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. Proc 31st Int Conf on Neural Information Processing Systems, p.6629-6640.

Huang S, Li Q, Liao J, et al., 2023. An overview of controllable image synthesis: current challenges and future trends. *SSRN*, Article 4187269. https://ssrn.com/abstract=4187269

Huang SS, Jin X, Jiang Q, et al., 2022. Deep learning for image colorization: current and future prospects. *Eng Appl Artif Intell*, 114:105006. https://doi.org/10.1016/j.engappai.2022.105006

Kocaoglu M, Snyder C, Dimakis AG, et al., 2018. CausalGAN: learning causal implicit generative models with adversarial training. Proc Int Conf on Learning Representations.

Lachapelle S, Brouillard P, Deleu T, et al., 2020. Gradient-based neural DAG learning. Proc 8th Int Conf on Learning Representations.

Lai PK, 2022. DeepSCM: an efficient convolutional neural network surrogate model for the screening of therapeutic antibody viscosity. *Comput Struct Biotechnol J*, 20:2143-2152. https://doi.org/10.1016/j.csbj.2022.04.035

Leeb F, Annadani Y, Bauer S, et al., 2020. Structural autoencoders improve representations for generation and transfer. https://arxiv.org/abs/2006.07796v1

Lippe P, Magliacane S, Löwe S, et al., 2022. CITRIS: causal identifiability from temporal intervened sequences. Proc 39th Int Conf on Machine Learning, p.13557-13603.

Liu ZW, Luo P, Wang XG, et al., 2015. Deep learning face attributes in the wild. Proc IEEE Int Conf on Computer Vision, p.3730-3738. https://doi.org/10.1109/ICCV.2015.425

Lopez-Paz D, Nishihara R, Chintala S, et al., 2017. Discovering causal signals in images. Proc IEEE Conf on Computer Vision and Pattern Recognition, p.6979-6987. https://doi.org/10.1109/CVPR.2017.14

Lu CC, Wu YH, Hernández-Lobato JM, et al., 2021. Nonlinear invariant risk minimization: a causal approach. https://arxiv.org/abs/2102.12353

Lv FR, Liang J, Li S, et al., 2022. Causality inspired representation learning for domain generalization. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.8046-8056. https://doi.org/10.1109/CVPR52688.2022.00788

Moraffah R, Moraffah B, Karami M, et al., 2020. Causal adversarial network for learning conditional and interventional distributions. https://arxiv.org/abs/2008.11376

Ng I, Zhu SY, Chen ZT, et al., 2019. A graph autoencoder approach to causal structure learning. https://arxiv.org/abs/1911.07420

Ng I, Zhu S, Fang Z, et al., 2022. Masked gradient-based causal structure learning. Proc SIAM Int Conf on Data Mining, p.424-432. https://doi.org/10.1137/1.9781611977172.48

Pan YH, Li ZC, Zhang LY, et al., 2022. Causal inference with knowledge distilling and curriculum learning for unbiased VQA. *ACM Trans Multim Comput Commun Appl*, 18(3):67. https://doi.org/10.1145/3487042

Petkov H, Hanley C, Dong F, 2022. DAG-WGAN: causal structure learning with Wasserstein generative adversarial networks. https://arxiv.org/abs/2204.00387

Reinhold JC, Carass A, Prince JL, 2021. A structural causal model for MR images of multiple sclerosis. Proc 24th Int Conf on Medical Image Computing and Computer-Assisted Intervention, p.782-792. https://doi.org/10.1007/978-3-030-87240-3_75

Salimans T, Goodfellow I, Zaremba W, et al., 2016. Improved techniques for training GANs. Proc 30th Int Conf on Neural Information Processing Systems, p.2234-2242.

Sanchez P, Tsaftaris SA, 2022. Diffusion causal models for counterfactual estimation. Proc 1st Conf on Causal Learning and Reasoning, p.647-668.

Sanchez P, Kascenas A, Liu X, et al., 2022. What is healthy? Generative counterfactual diffusion for lesion localization. Proc 2nd MICCAI Workshop on Deep Generative Models, p.34-44. https://doi.org/10.1007/978-3-031-18576-2_4

Sauer A, Geiger A, 2021. Counterfactual generative networks. Proc 9th Int Conf on Learning Representations.

Schölkopf B, Locatello F, Bauer S, et al., 2021. Toward causal representation learning. *Proc IEEE*, 109(5):612-634. https://doi.org/10.1109/JPROC.2021.3058954

Shen XW, Liu FR, Dong HZ, et al., 2022. Weakly supervised disentangled generative causal representation learning. *J Mach Learn Res*, 23(1):241.

Shen YJ, Zhou BL, 2021. Closed-form factorization of latent semantics in GANs. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.1532-1540. https://doi.org/10.1109/CVPR46437.2021.00158

Shen YJ, Yang CY, Tang XO, et al., 2022. InterFaceGAN: interpreting the disentangled face representation learned by GANs. *IEEE Trans Patt Anal Mach Intell*, 44(4):2004-2018. https://doi.org/10.1109/TPAMI.2020.3034267

Sun YP, Chen Q, He XY, et al., 2022. Singular value fine-tuning: few-shot segmentation requires few-parameters fine-tuning. Proc 36th Advances in Neural Information Processing Systems, p.37484-37496.

Suter R, Miladinovic D, Schölkopf B, et al., 2019. Robustly disentangled causal mechanisms: validating deep representations for interventional robustness. Proc 36th Int Conf on Machine Learning, p.6056-6065.

Varando G, 2020. Learning DAGs without imposing acyclicity. https://arxiv.org/abs/2006.03005v1

Vowels MJ, Camgoz NC, Bowden R, 2023. D'ya like DAGs? A survey on structure learning and causal discovery. *ACM Comput Surv*, 55(4):82. https://doi.org/10.1145/3527154

Wang WJ, Lin XY, Feng FL, et al., 2022. Causal representation learning for out-of-distribution recommendation. Proc ACM Web Conf, p.3562-3571. https://doi.org/10.1145/3485447.3512251

Wang XQ, Du YL, Zhu SY, et al., 2021. Ordering-based causal discovery with reinforcement learning. Proc 30th Int Joint Conf on Artificial Intelligence, p.3566-3573.

Wang YF, Zhu YL, Hang TT, et al., 2021. Incorporating proportional sparse penalty for causal structure learning. Proc IEEE 33rd Int Conf on Tools with Artificial Intelligence, p.105-112. https://doi.org/10.1109/ICTAI52525.2021.00023

Wei D, Gao T, Yu Y, 2020. DAGs with no fears: a closer look at continuous optimization for learning Bayesian networks. Proc 34th Int Conf on Neural Information Processing Systems, p.328.

Xia WH, Zhang YL, Yang YJ, et al., 2023. GAN inversion: a survey. *IEEE Trans Patt Anal Mach Intell*, 45(3):3121-3138. https://doi.org/10.1109/TPAMI.2022.3181070

Yang MY, Liu FR, Chen ZT, et al., 2021. CausalVAE: disentangled representation learning via neural structural causal models. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.9593-9602. https://doi.org/10.1109/CVPR46437.2021.00947

Yu Y, Chen J, Gao T, et al., 2019. DAG-GNN: DAG structure learning with graph neural networks. Proc 36th Int Conf on Machine Learning, p.7154-7163.

Zhang LM, Rao A, Agrawala M, 2023. Adding conditional control to text-to-image diffusion models. https://arxiv.org/abs/2302.05543

Zhang WB, Liao J, Zhang Y, et al., 2022. CMGAN: a generative adversarial network embedded with causal matrix. *Appl Intell*, 52(14):16233-16245. https://doi.org/10.1007/S10489-021-03094-8

Zhang XH, Wong Y, Wu XF, et al., 2021. Learning causal representation for training cross-domain pose estimator via generative interventions. Proc IEEE/CVF Int Conf on Computer Vision, p.11270-11280. https://doi.org/10.1109/ICCV48922.2021.01108

Zheng X, Aragam B, Ravikumar P, et al., 2018. DAGs with NO TEARS: continuous optimization for structure learning. Proc 32nd Int Conf on Neural Information Processing Systems, p.9492-9503.

Zhu JG, Xie HC, AbdAlmageed W, 2022. Do-operation guided causal representation learning with reduced supervision strength. https://arxiv.org/abs/2206.01802v1

Zhu SY, Ng I, Chen ZT, 2020. Causal discovery with reinforcement learning. Proc 8th Int Conf on Learning Representations.