Causal Coherence in Image Inpainting: Integrating Causal Reasoning with VAEs for Image Restoration

Ruilin Jin
Department of Computer and Data Science
School of Engineering
rxj420@case.edu

Tuo Liang
Department of Computer and Data Science
School of Engineering
txl859@case.edu

Abstract

The field of image inpainting[1] has witnessed substantial growth, fueled by the need for advanced techniques in digital forensics, image restoration, and object removal. Traditional methods, however, often fall short in maintaining semantic coherence. Our research introduces a approach to image inpainting that integrates causal reasoning within a novel generative process, leveraging the strengths of Variational Autoencoders (VAEs) and Structural Causal Models (SCMs). We propose a model that adapts the Causal Layer of CausalVAE, enhanced by the structural elements of NVAE, to address three primary challenges: creating expressive neural networks, scaling up training for larger image sizes and groups, and maintaining training stability. This model focuses on the coherent reconstruction of missing or corrupted image regions through an understanding of causal relationships among features. Our approach not only enhances the semantic coherence and realism of inpainted images but also fosters interpretability in the latent space, paving the way for more reliable and comprehensible image restoration processes.

1 Introduction

1.1 Motivation

Image inpainting, the technique of restoring corrupted or missing parts of images, is indispensable in various domains including digital forensics, image restoration, and object removal. The endeavor is to refill the missing segments in a visually coherent manner [4]. However, conventional methods often stumble in preserving semantic coherence. The recent strides in disentangled representation learning provide a promising pathway to embed causal reasoning in image inpainting, which is anticipated to bolster the semantic coherence and realism of inpainted images, rendering the inpainting process more reliable and the outcomes more interpretable.

1.2 Background & related work

1.2.1 Variational Autoencoders (VAEs). VAEs [3] have emerged as a powerful framework for probabilistic generative modeling. They are especially known for their capability to learn a lower-dimensional representation of input data, which is instrumental for various tasks in computer vision including image inpainting. By imposing a probabilistic graphical model over the variables, VAEs can generate new data that's similar to the training data, making them a popular choice for generative tasks. The structure of VAEs, which includes an encoder and a decoder, has been the backbone for advancing disentangled representation learning.

1.2.2 CausalVAE. A recent advancement in the realm of disentangled representation learning [8] is the introduction of CausalVAE [9], which integrates causal reasoning within the Variational Autoencoder framework. The CausalVAE model augments the traditional VAE with a causal layer, aiming to uncover and leverage causal relationships among the features in the data. This architecture allows for a more structured and interpretable latent space, which is of significant interest for tasks that could benefit from an understanding of causal relationships among the features, like image inpainting. The principles laid down by the CausalVAE framework underline the potential of causal disentanglement in enhancing the generative processes, which is a cornerstone of our proposed advancement in image inpainting.

1.2.3 NVAE. Neural Variational Autoencoders (NVAE) [7] represent an evolution in the field of generative models, diverging from traditional Variational Autoencoders (VAE) and Causal VAEs in several significant ways. NVAE adopts a hierarchical architecture with depthwise separable convolutions, enabling more efficient handling of high-dimensional data compared to the fully connected or standard convolutional layers in regular VAEs. This advanced architecture is complemented by techniques like batch normalization and residual connections, which enhance training stability and address issues such as posterior collapse, a common problem in VAE training. Unlike Causal VAEs, which are tailored to model and infer causal relationships within data, NVAE focuses primarily on improving the quality and efficiency of generative modeling. It achieves this by being more scalable, allowing for training on larger datasets and generating higher quality images. Thus, NVAE stands out for its ability to efficiently generate high-quality models without specifically addressing causal inference, unlike its Causal VAE counterparts.

1.2.4 SCM. Structural Causal Models (SCM) are a framework used in causal inference to represent and understand

the mechanisms behind observed data. SCMs consist of structural equations and a directed acyclic graph (DAG) that depict how variables influence each other. Each equation in an SCM corresponds to a causal mechanism, showing how a particular variable is generated from its direct causes. These models help in distinguishing between correlation and causation, enabling the prediction of effects from interventions and the understanding of counterfactual scenarios.

1.2.5 Causal Disentanglement. Disentangled representation learning aims to separate out the underlying causal factors of the data into distinct representations. This separation makes the learned representations more interpretable and easier to manipulate. Causal disentanglement takes this a step further by not only seeking to disentangle the representations but also to understand the causal relationships between them. This understanding of causal relationships is crucial as it allows for better generalization across different tasks and can potentially improve the performance of downstream tasks like image inpainting.

2 Method

Our model displayed in Fig. 1 focuses on tackling three challenges: (i) designing expressive neural networks specifically for VAEs, (ii) scaling up the training to a large number of hierarchical groups and image sizes while maintaining training stability, and (iii) implementing structure of NVAE in CausalVAE to improve its generation quality while remain CausalVAE's causal relationship. Our methodology pivots around adapting the Causal Layer of CausalVAE to the image inpainting domain. The causal relationships discovered will be utilized to generate semantically coherent inpaintings.

2.1 Design of New Generative Process

Develop a new generative process to handle the inpainting tasks ensuring it leverages the causal relationships to fill missing or corrupted regions coherently.

- *Residual Cell:* Modified the Encoder part for normal VAE to adopt the residual cell structure of NVAE, where the residual cells expands the number of channels E times before applying the depthwise separable convolution, and then maps it back to C channels.
- *Generative Model:* A new generative model $p_\theta(z, x|m)$ is designed, where z is the latent variable, x is the image data, and m denotes the missing or corrupted regions. The model aims to generate inpaintings that are not only visually plausible but also semantically coherent by leveraging the causal relationships.
- *Probabilistic Formulation:* The inpainting process is formulated as a probabilistic task. The objective is to maximize the a posteriori probability $p_\theta(z|x, m)$ which can be decomposed as $p_\theta(z|x, m) \propto p_\theta(x|z, m)p_\theta(z|m)$ according to Bayes' theorem.
- *Optimization:* The optimization process involves adjusting the model parameters θ to maximize the likelihood of generating realistic inpaintings. An optimization objective can be formulated as $\arg\max_\theta \mathcal{L}(\theta; X, M)$, employing backpropagation and stochastic gradient descent to find the optimal parameters.

2.2 Adaptation of Causal Layer

Integrate causal reasoning within the generative process to ensure the model can discover and leverage causal relationships among features for better inpainting.

- *Causal Graphical Model:* Construct a causal graphical model [6] to represent the relationships among observed variables, latent variables, and missing or corrupted regions in images. Mathematically, the causal graphical model is represented as a directed acyclic graph (DAG) $G = (V, E)$, where V is a set of vertices representing the variables, and E is a set of edges indicating causal relationships.
- *Inference:* The causal relationships are inferred by maximizing the likelihood of the observed data under the causal graphical model [2]. The likelihood can be formulated as $\mathcal{L}(\theta; X, M) = \sum_{i=1}^{N} \log p_\theta(x_i|m_i)$, where θ are the model parameters, X is the observed data, and M represents the missing or corrupted regions.

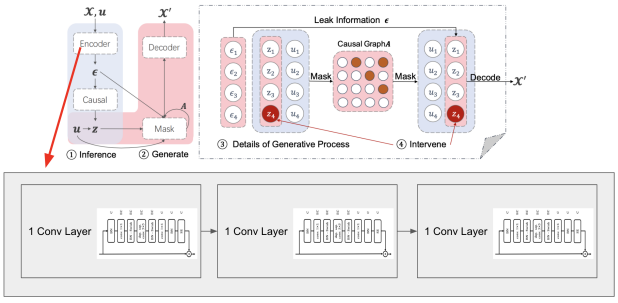

Figure 1. Structure for our model. The encoder takes observation x as inputs to generate independent exogenous variable ϵ, whose prior distribution is assumed to be standard Multivariate Gaussian.

3 Experiments

3.1 Dataset

For the training and evaluation of our proposed model, we have chosen to utilize the CelebA dataset [5], which is a widely recognized dataset offering a rich set of annotations for facial attributes. The CelebA dataset comprises images annotated with 40 attribute labels, providing invaluable information that elucidates the characteristics and features within face images.

We pre-processed the CelebA dataset to tailor it to our project needs, which includes filtering images to retain those with particular attributes. The prepared dataset was employed for both training and evaluating our model, ensuring that it learns to manipulate images in a way that adheres to real-world facial attribute constrains.

Figure 2. Comparison of our initial result to the ground truth.

3.2 Initial Experiments

Fig. 2 shows the result of out model on real world banchmark dataset CelebA comparing to ground truth, with Fig. 4 showing the experiments on intervening concepts GENDER, SMILE, EYES OPEN and MOUTH OPEN respectively. We observe that when we intervene the cause concept SMILE, the status of MOUTH OPEN also changes. In contrast, intervening effect concept MOUTH OPEN does not cause the cause concept SMILE to change.

Despite our efforts in adjusting hyper-parameters, altering the encoder structure, and varying the length of the latent dimension, we observed that the generated images persistently exhibited blurriness in the background. Consequently, we shifted our focus to explore alternative aspects of the model, particularly by modifying the decoder to address this issue.

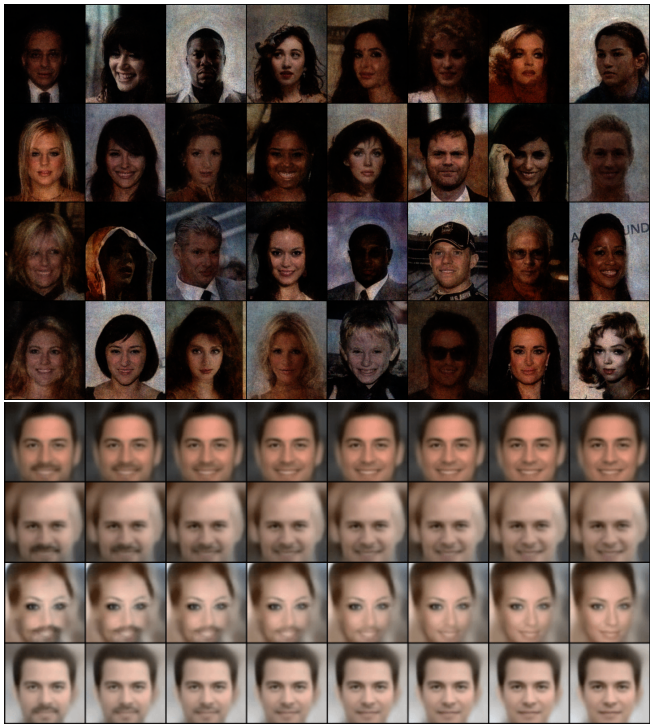

Figure 3. Comparison of our improved result to the result in CausalVAE paper.

3.3 Improved Experiments

Initially, we employed a linear structured decoder derived from the CausalVAE framework. However, to address issues with background clarity in the generated images, we experimented by adapting the decoder to mirror the residual cell structure of NVAE. This modification aimed to leverage NVAE's proficiency in handling high-dimensional data and its enhanced generative capabilities. The results of this adaptation are showcased in Fig. 3, where a marked improvement in background clarity is evident compared to the outputs produced using the original CausalVAE decoder. This improvement can be attributed to the residual cell structure's ability to better capture and reconstruct complex background patterns, leading to sharper and more detailed image generation.

3.4 Learning Process

We show in Fig. 5 the learned adjacency matrix A.As the training epoch increases, we see that the graph learned by our model quickly converges to the true one, which shows that our method is able to correctly learn the causal relationship among the factors.

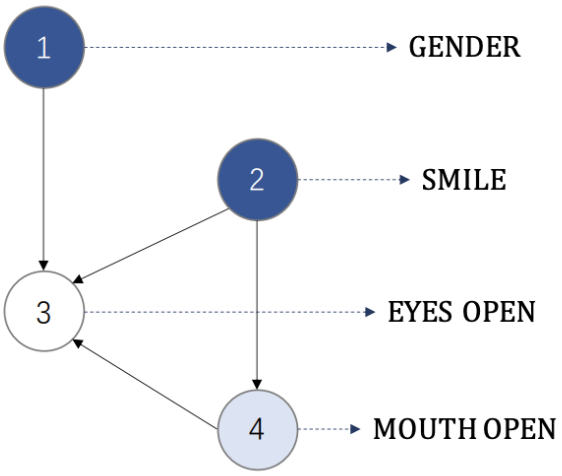

Figure 4. SCM for our experiment

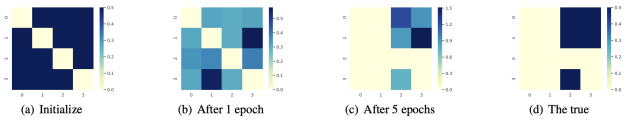

Figure 5. The learning process of causal matrix A. The concepts include: GENDER, SMILE, EYES OPEN, MOUTH OPEN (top-to-bottom and left-to-right order); (c) converged A, (d) ground truth.

4 Limitation and Future Works

4.1 Image Blurriness

While modifying the decoder to the residual cell structure of NVAE improved background clarity, the images remain somewhat blurry. This partial enhancement highlights the need for further refinements in our approach, suggesting areas like parameter optimization and architectural adjustments for future work to achieve sharper image quality.

4.2 Trade-off Between Quality and Intervention Capabilities

After enhancing the image quality through the decoder modifications, we observed an unexpected trade-off: a decrease in the model's intervention capabilities. This suggests a complex balance between achieving higher image resolution and maintaining the model's ability to effectively manipulate and intervene in the generative process. Future efforts will focus on optimizing this balance, aiming to simultaneously retain high-quality image generation while preserving robust intervention capabilities.

4.3 Slow Convergence

The slow convergence of our model, requiring over 250 epochs to produce discernible images, is largely due to the complexities of merging VAEs with causal reasoning. VAEs inherently need extensive training to learn data distributions, and incorporating causal relationships adds to this complexity. This dual task of capturing probabilistic generative properties and understanding causal links results in a longer training duration for achieving coherent visual outputs.

4.4 Based on SCM Assumption

In the current iteration of our model, the application of SCMs relies on a pre-determined causal framework. This approach, while effective in integrating causal reasoning with the generative process, has its limitations. It relies heavily on predefined causal relationships, which may not fully capture the complexity and variability inherent in real-world data. For future developments, a significant advancement would be to enable the model to autonomously generate and refine its own SCM. This evolution would involve the model not only discovering the causal relationships within data but also continuously adjusting and optimizing these relationships as it processes more information.

5 Conclusion

In conclusion, this research introduces a novel image inpainting method that combines causal reasoning with advanced VAEs and SCMs, as shown in our CelebA dataset experiments. This method effectively generates semantically coherent and visually convincing inpaintings by merging the Causal Layer of CausalVAE with NVAE's structure. This blend not only enhances technical prowess but also ensures effective generation and disentanglement, balancing theoretical and practical aspects of generative models. Despite challenges like image blurriness and slow convergence, the potential in applications such as digital forensics is significant. Future work will focus on refining these techniques, enhancing their application in realistic scenarios.

References

[1] Marcelo Bertalmio, Guillermo Sapiro, Vincent Caselles, and Coloma Ballester. 2000. Image inpainting. In *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*. 417–424.

[2] Thomas L Griffiths and Joshua B Tenenbaum. 2009. Theory-based causal induction. *Psychological review* 116, 4 (2009), 661.

[3] Diederik P Kingma and Max Welling. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114* (2013).

[4] Hongyu Liu, Bin Jiang, Yi Xiao, and Chao Yang. 2019. Coherent semantic attention for image inpainting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 4170–4179.

[5] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. 2018. Large-scale celebfaces attributes (celeba) dataset. *Retrieved August* 15, 2018 (2018), 11.

[6] Judea Pearl. 1998. Graphical models for probabilistic and causal reasoning. *Quantified representation of uncertainty and imprecision* (1998), 367–389.

[7] Arash Vahdat and Jan Kautz. 2020. NVAE: A deep hierarchical variational autoencoder. *Advances in neural information processing systems* 33 (2020), 19667–19679.

[8] Xin Wang, Hong Chen, Si'ao Tang, Zihao Wu, and Wenwu Zhu. 2022. Disentangled representation learning. *arXiv preprint arXiv:2211.11695* (2022).

[9] Mengyue Yang, Furui Liu, Zhitang Chen, Xinwei Shen, Jianye Hao, and Jun Wang. 2020. Causalvae: Structured causal disentanglement in variational autoencoder. *arXiv preprint arXiv:2004.08697* (2020).