Segmentor-guided Counterfactual Fine-Tuning for Locally Coherent and Targeted Image Synthesis

Tian Xia[1], Matthew Sinclair[1,2], Andreas Schuh[1,2], Fabio De Sousa Ribeiro[1], Raghav Mehta[1], Rajat Rasal[1], Esther Puyol-Antón[1,2], Samuel Gerber[2], Kersten Petersen[2], Michiel Schaap[1,2], and Ben Glocker[1]

[1] Department of Computing, Imperial College London, UK
t.xia@imperial.ac.uk
[2] Heartflow, Inc., Mountain View, CA, USA

Abstract. Counterfactual image generation is a powerful tool for augmenting training data, de-biasing datasets, and modeling disease. Current approaches rely on external classifiers or regressors to increase the effectiveness of subject-level interventions (e.g., changing the patient's age). For structure-specific interventions (e.g., changing the area of the left lung in a chest radiograph), we show that this is insufficient, and can result in undesirable global effects across the image domain. Previous work used pixel-level label maps as guidance, requiring a user to provide hypothetical segmentations which are tedious and difficult to obtain. We propose *Segmentor-guided Counterfactual Fine-Tuning* (Seg-CFT), which preserves the simplicity of intervening on scalar-valued, structure-specific variables while producing locally coherent and effective counterfactuals. We demonstrate the capability of generating realistic chest radiographs, and we show promising results for modeling coronary artery disease. Code: https://github.com/biomedia-mira/seg-cft.

Keywords: Counterfactual image synthesis · Causal generative models.

1 Introduction

Causal questions, such as *"How would this patient's disease have progressed if treatment A had been administered instead of treatment B?"*, are fundamental to scientific inquiry and clinical decision-making. Addressing such questions requires a causal framework capable of simulating realistic scenarios from observed data—going beyond conventional statistical models that primarily capture correlations. Causal models explicitly represent how changes in one factor influence another, enabling the exploration of both real-world interventions and hypothetical, or counterfactual, scenarios. The ability to generate counterfactual images has become particularly valuable in medical imaging, where it facilitates a range of applications, including data augmentation [11,26], bias mitigation [15], explainability [23], and disease progression modeling [25]. By enabling targeted modifications to patient images, counterfactual generation can support model generalization, improve interpretability, and allow researchers to explore alternative clinical pathways.

Recent efforts [16,14,34,38,29,6] have tried to integrate causality with deep generative models, including GANs [7], VAEs [13], and diffusion models [30,8,31]. However, most methods focus on association or intervention, without a principled approach to counterfactual reasoning—the highest level in Pearl's causal hierarchy. Notable exceptions include Neural Causal Models (NCMs) [35,36,20] and Deep Structural Causal Models (DSCMs) [21,18,5], which integrate causal structures with deep generative models.

Ribeiro et al. [5] proposed to train the generative causal model using a hierarchical variational auto-encoder (HVAE) conditioned on the assumed causal parents. However, relying solely on standard likelihood-based training was found to result in suboptimal axiomatic *effectiveness* [18], meaning that the generative model may fail to enforce counterfactual consistency—ignoring conditioning on intervened-upon parents in the forward model post-abduction [37]. To address this, counterfactual fine-tuning (CFT) was proposed as an additional step, refining the HVAE with pretrained parent classifiers or regressors to improve adherence to causal structure [5]. To this end, previous works [5,37,26,27,10] focused on patient-level characteristics (e.g., sex, age, disease status). The effectiveness of DSCM and CFT has not been validated on structure-specific interventions, such as modifying specific anatomical regions or localized diease patterns.

In this paper, we focus on these structure-specific interventions. We find that the previous CFT with regressors (Reg-CFT) is not sufficient for locally coherent and targeted counterfactual generation. To enable localized control of image generation, one potential approach is using segmentation masks to guide generative models [24,1]. But integrating masks into a causal framework poses challenges: (i) defining their causal role remains unclear, and (ii) requiring predefined counterfactual masks reduces usability as these are difficult to obtain.

Recent work has shown that medical image classifiers often rely on spurious or non-local features, motivating the use of spatial supervision via segmentation to improve specificity and robustness [9,17,28,2]. Building on this insight, we propose Segmentor-guided Counterfactual Fine-Tuning (Seg-CFT), a method for fine-grained anatomical control in counterfactual image generation. We use scalar-valued variables (e.g., the area of the left lung) as guiding signals for counterfactual generation, which preserves the simplicity of the user interaction, avoiding any inputs at the pixel-level. In Seg-CFT, we utilise pre-trained, weight-frozen segmentors to increase the counterfactual effectiveness of DSCMs. The values for the intervened variables are directly determined from the obtained segmentations and compared against the desired user-specified target value in the loss function when fine-tuning the DSCM output. This enables a simple mechanism for intervening directly on structure-specific properties and, as we will show, yields locally coherent and targeted modifications of anatomical structures.

In summary, our key contributions are the following:

- We propose a novel guidance mechanism for counterfactual fine-tuning, which enables fine-grained structural control of localised interventions.

- We provide a comparative analysis against state-of-the-art regressor-based counterfactual fine-tuning of DSCMs, highlighting the improved effectiveness of our approach.
- We demonstrate the capability of generating realistic counterfactual chest radiographs. We also show promising early results on the application of modelling coronary artery disease progression.

2 Method

2.1 Review of DSCM

Structural Causal Models (SCMs) [22] are defined by a triplet $\langle U, A, F \rangle$, where $U = \{u_i\}_{i=1}^{K}$ represents a set of exogenous variables, $A = \{a_i\}_{i=1}^{K}$ a set of endogenous variables, and $F = \{f_i\}_{i=1}^{K}$ a set of functions satisfying $a_k \coloneqq f_k(\mathbf{pa}_k, u_k)$, where $\mathbf{pa}_k \subseteq A \setminus a_k$ are the direct causes (or parents) of a_k. SCMs enable interventions through the do-operator, e.g., by modifying one or more parent variables. Counterfactual inference involves three steps: (i) Abduction: inferring exogenous noise from observed data; (ii) Action: applying an intervention, e.g. $do(a_k \coloneqq c)$; and (iii) Prediction: computing counterfactual outcomes using the modified model and the inferred posterior over the exogenous variables.

Deep Structural Causal Models (DSCMs) were introduced in [21] and later refined in [5] for high-resolution counterfactual image generation. Given an image $\mathbf{x}$, let $\{a_1, \ldots, a_{K-1}\} \supseteq \mathbf{pa}_{\mathbf{x}}$ be its *ancestors*. Each low-dimensional attribute follows an invertible conditional normalizing flow, $a_k = f_k(u_k; \mathbf{pa}_k)$, making abduction explicit and tractable. For high-dimensional variables like images, the generative mechanism is implemented via a Hierarchical Variational Autoencoder (HVAE). To generate a counterfactual image, we first infer the exogenous noise for the image, $\mathbf{z} \sim q_\phi(\mathbf{z} \mid \mathbf{x}, \mathbf{pa}_{\mathbf{x}})$, where q_ϕ is the HVAE encoder. Similarly, the exogenous noise for attributes is inferred as $u_k = f_k^{-1}(a_k; \mathbf{pa}_k)$. We then perform an intervention by setting $a_i \coloneqq c$, allowing for modifications to multiple attributes simultaneously. Using the abducted noise u_k, we compute counterfactual parent values $\widetilde{\mathbf{pa}}_{\mathbf{x}}$ and generate the counterfactual image, $\widetilde{\mathbf{x}} = g_\theta(\mathbf{z}, \widetilde{\mathbf{pa}}_{\mathbf{x}})$.

2.2 Regressor-based Counterfactual Fine-tuning (Reg-CFT)

Previous works [5,37,10] observed that likelihood-based HVAE training may cause *ignored counterfactual conditioning*, where $\widetilde{\mathbf{x}}$ does not respect the intervened counterfactual parents $\widetilde{\mathbf{pa}}_{\mathbf{x}}$. This can be mitigated with *counterfactual fine-tuning (CFT)* [5,37]. The key idea of CFT is to leverage pre-trained classifiers or regressors $q_\xi(\mathbf{pa}_{\mathbf{x}} \mid \mathbf{x})$, and to optimise the HVAE parameters $\{\theta, \phi\}$ by maximising $\log q_\xi(\widetilde{\mathbf{pa}}_{\mathbf{x}} \mid \widetilde{\mathbf{x}})$ while keeping ξ frozen. This fine-tuning step encourages the DSCM to generate faithful counterfactual images that obey the intended interventions by enforcing $\widetilde{\mathbf{pa}}_{\mathbf{x}}$ to be predictable from $\widetilde{\mathbf{x}}$. In this work, we refer to the CFT used in previous studies as Reg-CFT, as they employ pre-trained regressors (or classifiers). For simplicity, we use the term *regressor* to refer to both regressor and classifier throughout the paper.

While Reg-CFT has been demonstrated to be effective for *subject-level interventions*, e.g., changing a subject's sex, it has not been tested on *structure-specific interventions*, e.g., reducing the size of an organ in a medical image. We assess its effectiveness for structure-specific interventions by extracting *areas* of structures as scalar parent variables for $\mathbf{x}$ using 2D medical images. As shown in Section 3, we find that Reg-CFT is not sufficient for these interventions, and produces undesirable global changes. A potential reason is that with Reg-CFT, there is insufficient guidance for DSCMs to capture the exact meaning of (scalar-valued) variables such as organ size, as the regressor could learn potential spurious correlations. For example, it is possible that DSCMs incorrectly learn that the variable *left lung area* corresponds to the *mean pixel intensities of left lung and heart* or some other spuriously correlated characteristics in the images. As such, it may be necessary to incorporate structural and spatially coherent information into CFT to better align the semantic meaning of scalar-valued, structure-specific variables.

2.3 Segmentor-guided Counterfactual Fine-Tuning (Seg-CFT)

For the rest of the paper, we focus on structure-specific interventions, i.e. changing areas of anatomical and disease-related structures in 2D images. To improve the effectiveness of this type of intervention, one potential approach is to rely on pixel-level segmentations to guide the generative models [24,1]. This requires users to manually construct hypothetical, clinically plausible label maps, which is tedious and challenging. Additionally, incorporating label maps into a causal graph is inherently difficult, as the causal relationships between label maps and other variables may not be obvious. It is thus desirable to preserve the simplicity of intervening on scalar-valued causal variables while making the DSCM aware of the spatial context of these structure-specific variables.

To this end, we introduce Segmentor-guided Counterfactual Fine-Tuning (*Seg-CFT*), a novel approach that leverages a pre-trained segmentation model

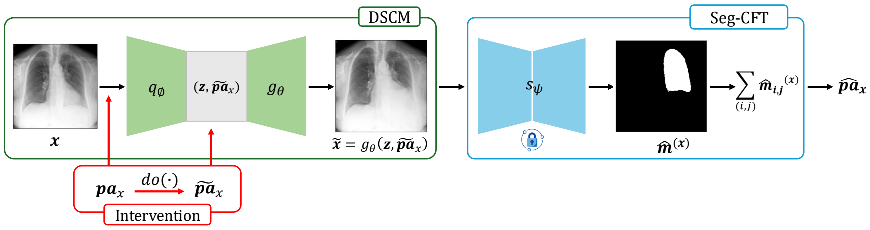

DSCM
$\boldsymbol{x}$
q_ϕ
$(\boldsymbol{z}, \widetilde{\boldsymbol{pa}}_x)$
g_θ
$\widetilde{\boldsymbol{x}} = g_\theta(\boldsymbol{z}, \widetilde{\boldsymbol{pa}}_x)$
$\boldsymbol{pa}_x \xrightarrow{do(\cdot)} \widetilde{\boldsymbol{pa}}_x$
Intervention
Seg-CFT
s_ψ
$\widehat{\boldsymbol{m}}^{(x)}$
$\sum_{(i,j)} \widehat{\boldsymbol{m}}_{i,j}{}^{(x)}$
$\widehat{\boldsymbol{pa}}_x$

Fig. 1: A schematic of the proposed Seg-CFT method, where we utilise pre-trained segmentors to guide counterfactual fine-tuning of DSCMs.

during counterfactual fine-tuning. For Seg-CFT, we retain all variables of interest as scalar-valued, similar to Reg-CFT, including both subject-level and structure-specific variables. This allows for tractable abduction, intervention and counterfactual reasoning for causal variables.

When intervening on one or more variables with counterfactual parents $\widetilde{\mathbf{pa}}_{\mathbf{x}}$, we first predict the segmentation label maps $\widehat{\mathbf{m}}^{(\mathbf{x})}$ for different structures from counterfactual images $\widetilde{\mathbf{x}}$ using segmentors s_ψ: $\widehat{\mathbf{m}}^{(\mathbf{x})} \sim s_\psi(\widehat{\mathbf{m}}^{(\mathbf{x})} \mid \widetilde{\mathbf{x}})$. Next, we compute the *areas* of the structures by summing the pixel-level label probabilities of the predicted segmentations: $\widehat{\mathbf{pa}}_{\mathbf{x}} = \sum_{(i,j)} \widehat{\mathbf{m}}^{(\mathbf{x})}_{i,j}$.

We then optimize the HVAE parameters $\{\theta, \phi\}$ by minimizing the loss function $l(\widehat{\mathbf{pa}}_{\mathbf{x}}, \widetilde{\mathbf{pa}}_{\mathbf{x}})$, where l is designed for scalar-valued variables. A schematic of Seg-CFT is presented in Fig. 1. The key difference between Seg-CFT and Reg-CFT is that Seg-CFT leverages pre-trained segmentors to *indirectly* obtain $\widehat{\mathbf{pa}}_{\mathbf{x}}$, whereas Reg-CFT directly predicts $\widehat{\mathbf{pa}}_{\mathbf{x}}$ using regressors. With Seg-CFT, DSCMs must generate locally coherent and meaningful changes that affect the segmentation masks produced by the segmentor. This effectively forces the DSCM to learn the spatial context of the scalar-valued structure-specific variables. Notably, segmentors are used *only* during training in Seg-CFT. Segmentors are not required during inference, and DSCMs can perform abduction, intervention, and counterfactual generation in the same manner as in previous works [5].

3 Experiments and results

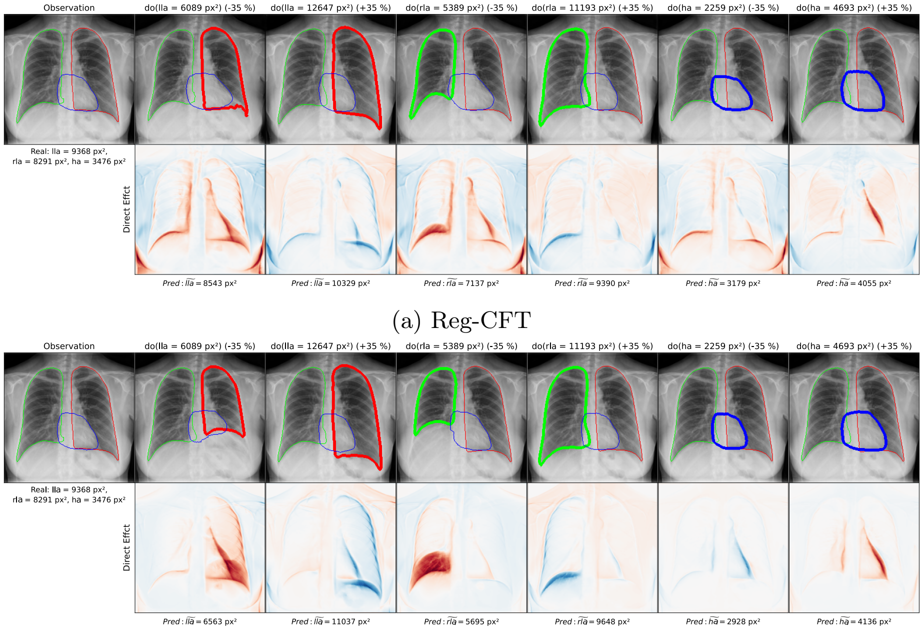
We conduct experiments on two datasets: the publicly available PadChest [3] and an internal coronary computed tomography angiography (CCTA) dataset [32]. Our evaluation compares the proposed Seg-CFT with Reg-CFT for structure-specific interventions, in particular by modifying the areas of targeted structures. For Reg-CFT, we pre-train ResNet-based regressors to predict the area of structures. For Seg-CFT, we pre-train U-Net segmentors using a Dice loss to produce 2D label maps. We evaluate counterfactuals via effectiveness [5,18], which assesses whether generated images obey the counterfactual parents, i.e. $d(\widehat{\mathbf{pa}}_{\mathbf{x}}, \widetilde{\mathbf{pa}}_{\mathbf{x}})$, where d is a metric function. The segmentor used for evaluation is not the same as the one used for fine-tuning; each is trained independently.

3.1 Study 1: Chest radiographs

We begin by evaluating our method on chest radiographs from the PadChest dataset [3], intervening on the size of anatomical structures. We manually selected around 85k subjects, removing mislabeled or low-quality images. The resulting dataset consists of 61,714 images for training, 6,911 for validation and 17,123 for testing. All images were resized to 224×224 pixels. We consider three structure-specific variables: left lung area (LLA), right lung area (RLA), and heart area (HA), with sex and age included as their parents. The original PadChest data does not have segmentation masks. We use masks obtained with the

Table 1: Quantitative evaluation of effectiveness of intervened and unintervened variables. Best results are highlighted as **bold**. For PadChest, we measure MAPE (%). For CCTA, we measure MAE (mm^2). Seg-CFT consistently outperforms other methods.

CFT	PadChest				CCTA			
	Intervention	LLA	RLA	HA	Intervention	NCPA	CPA	LA
No CFT [5]	do(LLA)	16.1	4.3	6.7	do(NCPA)	17.27	**3.45**	5.75
Reg-CFT [5]		13.4	2.2	5.7		14.63	6.83	12.59
Seg-CFT (Ours)		**10.0**	**2.1**	**5.2**		**10.33**	3.98	6.32
No CFT [5]	do(RLA)	3.7	16.9	5.1	do(CPA)	10.41	20.96	7.97
Reg-CFT [5]		**1.3**	13.3	4.6		15.61	13.88	12.10
Seg-CFT (Ours)		1.4	**10.1**	**4.4**		**10.40**	**8.14**	7.49
No CFT [5]	do(HA)	3.7	4.0	14.2	do(LA)	9.79	**3.37**	14.84
Reg-CFT [5]		**1.3**	2.1	11.8		13.51	4.98	12.66
Seg-CFT (Ours)		**1.3**	**2.0**	**8.5**		**9.50**	3.51	**9.68**


Observation
do(lla = 6089 px²) (-35 %)
do(lla = 12647 px²) (+35 %)
do(rla = 5389 px²) (-35 %)
do(rla = 11193 px²) (+35 %)
do(ha = 2259 px²) (-35 %)
do(ha = 4693 px²) (+35 %)
Real: lla = 9368 px²,
rla = 8291 px², ha = 3476 px²
Direct Effct
Pred : $\widetilde{lla}$ = 8543 px²
Pred : $\widetilde{lla}$ = 10329 px²
Pred : $\widetilde{rla}$ = 7137 px²
Pred : $\widetilde{rla}$ = 9390 px²
Pred : $\widetilde{ha}$ = 3179 px²
Pred : $\widetilde{ha}$ = 4055 px²

(a) Reg-CFT

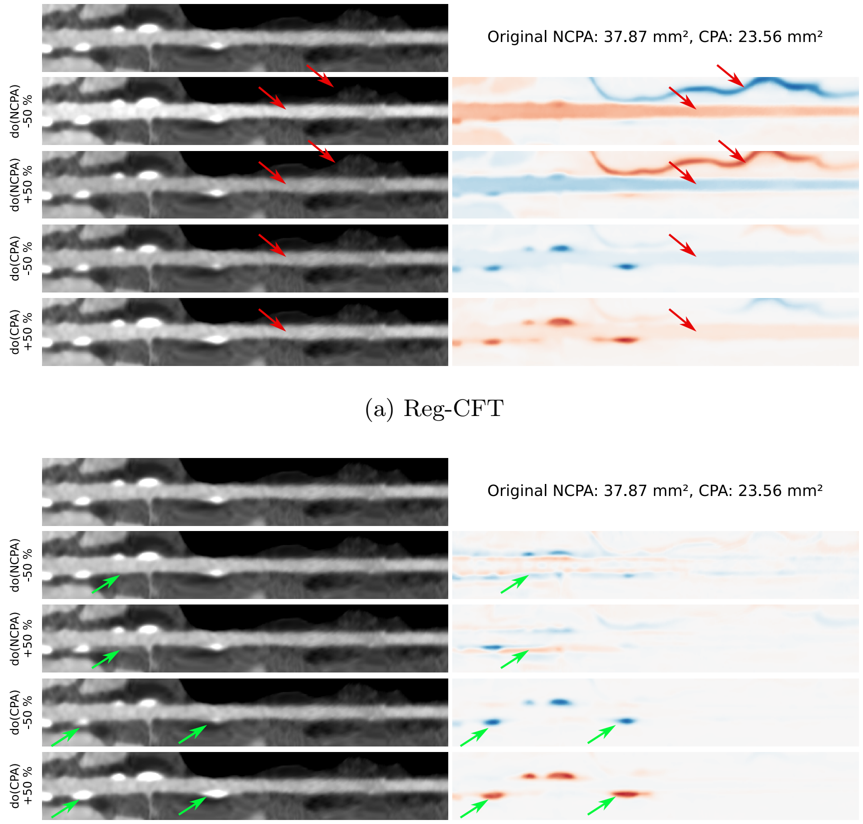
Observation
do(lla = 6089 px²) (-35 %)
do(lla = 12647 px²) (+35 %)
do(rla = 5389 px²) (-35 %)
do(rla = 11193 px²) (+35 %)
do(ha = 2259 px²) (-35 %)
do(ha = 4693 px²) (+35 %)
Real: lla = 9368 px²,
rla = 8291 px², ha = 3476 px²
Direct Effct
Pred : $\widetilde{lla}$ = 6563 px²
Pred : $\widetilde{lla}$ = 11037 px²
Pred : $\widetilde{rla}$ = 5695 px²
Pred : $\widetilde{rla}$ = 9648 px²
Pred : $\widetilde{ha}$ = 2928 px²
Pred : $\widetilde{ha}$ = 4136 px²

(b) Seg-CFT

Fig. 2: Generated counterfactuals (CFs) with (a) Reg-CFT and (b) Seg-CFT. First rows show original image $\mathbf{x}$ and CFs $\widetilde{\mathbf{x}}$ with segmentations for left lung (red), right lung (green) and heart (blue). The intervened structure is highlighted with **thicker** lines. Second rows show direct effect of CFs, i.e. $\widetilde{\mathbf{x}} - \mathbf{x}$. We also report the predicted areas (px^2) by segmentors on the bottom. We observe that Seg-CFT produces more locally coherent and spatially consistent interventions.

pre-trained segmentation model of the *torchxrayvison* [4] package for left lung, right lung, and heart, respectively.

The quantitative effectiveness results are shown in Table 1, where we measure the mean absolute percentage error (MAPE) (%) of LLA, RLA and HA for counterfactuals upon different interventions. We observe that without CFT, counterfactuals have the highest MAPE, highlighting the importance of CFT to mitigate ignored counterfactual conditioning. Seg-CFT achieves the best results with the lowest MAPE for all intervened variables.



(a) Reg-CFT

(b) Seg-CFT

Fig. 3: Generated CFs with (a) Reg-CFT and (b) Seg-CFT. Left columns show original image $\mathbf{x}$ and CFs $\widetilde{\mathbf{x}}$; right columns show direct effect of CFs, i.e. $\widetilde{\mathbf{x}} - \mathbf{x}$. Seg-CFT produces more locally coherent and spatially consistent interventions. Green arrows indicate expected local changes in plaque with Seg-CFT, while red arrows highlight undesirable changes of non-target structures with Reg-CFT.

Visual examples in Fig. 2 show that both Reg-CFT and Seg-CFT produce plausible counterfactuals, but we observe that Reg-CFT yields undesired effects outside the intervened structures. By contrast, with Seg-CFT, we observe more locally coherent and targeted changes, resulting in a more accurate intervention as reflected in the LLA, RLA and HA values predicted from counterfactuals. This suggests that with Seg-CFT, DSCMs obtain a better understanding of which part of an image should be changed upon structure-specific interventions.

3.2 Study 2: Coronary artery disease

CCTA is an important modality for the assessment of coronary artery disease (CAD), including evaluation of the composition and volume of atherosclerotic plaques. For the CCTA images, straightened curvilinear planar reformation (sCPR) was used to create 2D images from the longitudinal cross-section of the centerline of the left anterior descending (LAD) artery [12]. All images were sampled at a resolution of 0.25×0.25 mm and cropped to 64×384 pixels.

Segmentation masks of the coronary lumen and plaque were also generated. To achieve this, 3D meshes of the coronary lumen and outer wall were sampled and rasterised in the 2D sCPR image plane, generating masks of the lumen, calcified plaque, and non-calcified plaque [19,33]. A total of 18,433 CCTA images were used to generate samples, with 12,903 samples for training, 1,843 for validation, and 3,687 for testing. We consider three structure-specific variables: calcified plaque area (CPA), non-calcified plaque area (NCPA), and lumen area (LA). For simplicity, we assume that these are independent of each other.

The quantitative effectiveness of our approach is reported in Table 1, where we measure the mean absolute error (MAE) of NCPA, CPA, and LA in mm^2. Across all intervened variables, the proposed Seg-CFT achieves the best performance, followed by Reg-CFT. Notably, Reg-CFT results in significantly higher MAE for unintervened variables (indicated in gray text). This is likely due to the regressors learning spurious correlations, which subsequently affect DSCMs during fine-tuning. The visual results in Fig. 3 illustrate that Reg-CFT introduces unintended global effects across non-target structures. Note how interventions on NCPA affect the global intensity of the lumen area. In contrast, Seg-CFT yields much more localised effects, focusing specifically on the intervened structures. This demonstrates the advantage of incorporating segmentor-guidance in CFT.

4 Conclusion

By integrating pre-trained segmentation models during counterfactual fine-tuning, Seg-CFT enables locally coherent and targeted interventions while maintaining the simplicity of scalar-valued causal variables. Our experiments on PadChest for counterfactual chest radiographs, and on a CCTA dataset for simulating coronary artery disease progression, demonstrate that Seg-CFT outperforms regressor-based fine-tuning. Seg-CFT results in more targeted and structure-specific modifications while minimizing unintended global changes in unintervened regions. These findings highlight the importance of incorporating segmentation information to improve anatomical consistency. Future work should explore the causal relationship between structure-specific variables. We currently assume independence for simplicity. Practical applications of counterfactuals in areas such as disease progression modelling, treatment effect estimation, bias mitigation, and data augmentation should be explored. Extending Seg-CFT to 3D medical imaging, including volumetric CT and MRI scans, is another important direction that could unlock advanced counterfactual reasoning in high-dimensional data. Beyond controlling the area of anatomical structures, future research could investigate whether other characteristics, such as shape, location, or texture, can be explicitly modified within the counterfactual generation process, providing even greater flexibility in medical image synthesis.

Acknowledgments

This research was funded by Heartflow. B.G. received support from the Royal Academy of Engineering as part of his Kheiron/RAEng Research Chair. B.G. and F.R. acknowledge the support of the UKRI AI programme, and the EPSRC, for CHAI - EPSRC Causality in Healthcare AI Hub (grant no. EP/Y028856/1). R.M. is funded through the European Union's Horizon Europe research and innovation programme under grant agreement 10108030. R.R. is supported by the Engineering and Physical Sciences Research Council (EPSRC) through a Doctoral Training Partnerships PhD Scholarship.

Disclosure of interests

M.S., A.S., E.P.A., S.G., K.P. and M.S. are employees of Heartflow. B.G. is part-time employee of DeepHealth. No other competing interests.

References

1. Alaya, M.B., Lang, D.M., Wiestler, B., Schnabel, J.A., Bercea, C.I.: Mededit: Counterfactual diffusion-based image editing on brain mri. In: International Workshop on Simulation and Synthesis in Medical Imaging. pp. 167–176. Springer (2024)
2. Aslani, S., Lilaonitkul, W., Gnananthan, V., Raj, D., Rangelov, B., Young, A.L., Hu, Y., Taylor, P., Alexander, D.C., Collaborative, N., et al.: Optimising chest x-rays for image analysis by identifying and removing confounding factors. In: International Conference on Medical Imaging and Computer-Aided Diagnosis. pp. 245–254. Springer (2022)
3. Bustos, A., Pertusa, A., Salinas, J.M., De La Iglesia-Vaya, M.: Padchest: A large chest x-ray image dataset with multi-label annotated reports. Medical image analysis **66**, 101797 (2020)
4. Cohen, J.P., Viviano, J.D., Bertin, P., Morrison, P., Torabian, P., Guarrera, M., Lungren, M.P., Chaudhari, A., Brooks, R., Hashir, M., et al.: Torchxrayvision: A library of chest x-ray datasets and models. In: International Conference on Medical Imaging with Deep Learning. pp. 231–249. PMLR (2022)

5. De Sousa Ribeiro, F., Xia, T., Monteiro, M., Pawlowski, N., Glocker, B.: High fidelity image counterfactuals with probabilistic causal models. ICML (2023)
6. Geffner, T., Antoran, J., Foster, A., Gong, W., Ma, C., Kiciman, E., Sharma, A., Lamb, A., Kukla, M., Pawlowski, N., et al.: Deep end-to-end causal inference. arXiv preprint arXiv:2202.02195 (2022)
7. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial networks. Communications of the ACM **63**(11), 139–144 (2020)
8. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in Neural Information Processing Systems **33**, 6840–6851 (2020)
9. Hooper, S., Chen, M., Saab, K., Bhatia, K., Langlotz, C., Ré, C.: A case for reframing automated medical image classification as segmentation. Advances in Neural Information Processing Systems **36**, 55415–55441 (2023)
10. Ibrahim, Y., Warr, H., Kamnitsas, K.: Semi-supervised learning for deep causal generative models. In: MICCAI. pp. 294–303. Springer (2024)
11. Ilse, M., Tomczak, J.M., Forré, P.: Selecting data augmentation for simulating interventions. In: ICML. pp. 4555–4562. PMLR (2021)
12. Kanitsar, A., Fleischmann, D., Wegenkittl, R., Felkel, P., Groller, E.: CPR - curved planar reformation. In: IEEE Visualization, 2002. VIS 2002. pp. 37–44. IEEE. https://doi.org/10.1109/VISUAL.2002.1183754
13. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 (2013)
14. Kocaoglu, M., Snyder, C., Dimakis, A.G., Vishwanath, S.: Causalgan: Learning causal implicit generative models with adversarial training. arXiv preprint arXiv:1709.02023 (2017)
15. Kumar, A., Fathi, N., Mehta, R., Nichyporuk, B., Falet, J.P.R., Tsaftaris, S., Arbel, T.: Debiasing counterfactuals in the presence of spurious correlations. In: Workshop on Clinical Image-Based Procedures. pp. 276–286. Springer (2023)
16. Louizos, C., Shalit, U., Mooij, J.M., Sontag, D., Zemel, R., Welling, M.: Causal effect inference with deep latent-variable models. Advances in neural information processing systems **30** (2017)
17. Luo, L., Chen, H., Xiao, Y., Zhou, Y., Wang, X., Vardhanabhuti, V., Wu, M., Han, C., Liu, Z., Fang, X.H.B., et al.: Rethinking annotation granularity for overcoming shortcuts in deep learning–based radiograph diagnosis: A multicenter study. Radiology: Artificial Intelligence **4**(5), e210299 (2022)
18. Monteiro, M., Ribeiro, F.D.S., Pawlowski, N., Castro, D.C., Glocker, B.: Measuring axiomatic soundness of counterfactual image models. In: The Eleventh ICLR (2023), https://openreview.net/forum?id=lZOUQQvwI3q
19. Narula, J., Stuckey, T.D., Nakazawa, G., Ahmadi, A., Matsumura, M., Petersen, K., Mirza, S., Ng, N., Mullen, S., Schaap, M., Leipsic, J., Rogers, C., Taylor, C.A., Yacoub, H., Gupta, H., Matsuo, H., Rinehart, S., Maehara, A.: Prospective deep learning-based quantitative assessment of coronary plaque by computed tomography angiography compared with intravascular ultrasound: the REVEALPLAQUE study **25**(9), 1287–1295 (2024). https://doi.org/10.1093/ehjci/jeae115
20. Pan, Y., Bareinboim, E.: Counterfactual image editing. arXiv preprint arXiv:2403.09683 (2024)
21. Pawlowski, N., Coelho de Castro, D., Glocker, B.: Deep structural causal models for tractable counterfactual inference. Advances in Neural Information Processing Systems **33**, 857–869 (2020)
22. Pearl, J.: Causality. Cambridge university press (2009)

23. Pegios, P., Lin, M., Weng, N., Svendsen, M.B.S., Bashir, Z., Bigdeli, S., Christensen, A.N., Tolsgaard, M., Feragen, A.: Diffusion-based iterative counterfactual explanations for fetal ultrasound image quality assessment. arXiv preprint arXiv:2403.08700 (2024)
24. Pérez-García, F., Bond-Taylor, S., Sanchez, P.P., van Breugel, B., Castro, D.C., Sharma, H., Salvatelli, V., Wetscherek, M.T., Richardson, H., Lungren, M.P., et al.: Radedit: stress-testing biomedical vision models via diffusion image editing. In: ECCV. pp. 358–376. Springer (2024)
25. Puglisi, L., Alexander, D.C., Ravì, D.: Enhancing spatiotemporal disease progression models via latent diffusion and prior knowledge. In: MICCAI. pp. 173–183. Springer (2024)
26. Roschewitz, M., Ribeiro, F.D.S., Xia, T., Khara, G., Glocker, B.: Robust image representations with counterfactual contrastive learning. arXiv preprint arXiv:2409.10365 (2024)
27. Roschewitz, M., de Sousa Ribeiro, F., Xia, T., Khara, G., Glocker, B.: Counterfactual contrastive learning: robust representations via causal image synthesis. In: MICCAI Workshop on Data Engineering in Medical Imaging. pp. 22–32. Springer (2024)
28. Saab, K., Hooper, S., Chen, M., Zhang, M., Rubin, D., Ré, C.: Reducing reliance on spurious features in medical image classification with spatial specificity. In: Machine Learning for Healthcare Conference. pp. 760–784. PMLR (2022)
29. Sanchez, P., Liu, X., O'Neil, A.Q., Tsaftaris, S.A.: Diffusion models for causal discovery via topological ordering. arXiv preprint arXiv:2210.06201 (2022)
30. Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., Ganguli, S.: Deep unsupervised learning using nonequilibrium thermodynamics. In: International Conference on Machine Learning. pp. 2256–2265. PMLR (2015)
31. Song, Y., Sohl-Dickstein, J., Kingma, D.P., Kumar, A., Ermon, S., Poole, B.: Score-based generative modeling through stochastic differential equations. In: ICLR (2021), https://openreview.net/forum?id=PxTIG12RRHS
32. Taylor, C.A., Petersen, K., Xiao, N., Sinclair, M., Bai, Y., Lynch, S.R., UpdePac, A., Schaap, M.: Patient-specific modeling of blood flow in the coronary arteries **417** (2023). https://doi.org/10.1016/j.cma.2023.116414
33. Taylor, C.A., Petersen, K., Xiao, N., Sinclair, M., Bai, Y., Lynch, S.R., UpdePac, A., Schaap, M.: Patient-specific modeling of blood flow in the coronary arteries. Computer Methods in Applied Mechanics and Engineering **417**, 116414 (2023)
34. Tran, D., Blei, D.M.: Implicit causal models for genome-wide association studies. arXiv preprint arXiv:1710.10742 (2017)
35. Xia, K., Lee, K.Z., Bengio, Y., Bareinboim, E.: The causal-neural connection: Expressiveness, learnability, and inference. Advances in Neural Information Processing Systems **34**, 10823–10836 (2021)
36. Xia, K.M., Pan, Y., Bareinboim, E.: Neural causal models for counterfactual identification and estimation. In: The Eleventh ICLR (2023), https://openreview.net/forum?id=vouQcZS8KfW
37. Xia, T., Roschewitz, M., De Sousa Ribeiro, F., Jones, C., Glocker, B.: Mitigating attribute amplification in counterfactual image generation. In: MICCAI (2024)
38. Yang, M., Liu, F., Chen, Z., Shen, X., Hao, J., Wang, J.: Causalvae: Disentangled representation learning via neural structural causal models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9593–9602 (2021)