Fast Diffusion-Based Counterfactuals for Shortcut Removal and Generation

Nina Weng[1*], Paraskevas Pegios[1,2*], Eike Petersen[1], Aasa Feragen[1,2], and Siavash Bigdeli[1]

[1] Technical University of Denmark, Kongens Lyngby, Denmark
{ninwe,ppar,ewipe,afhar,sarbi}@dtu.dk
[2] Pioneer Centre for AI, Copenhagen, Denmark

Abstract. Shortcut learning is when a model – e.g. a cardiac disease classifier – exploits correlations between the target label and a spurious shortcut feature, e.g. a pacemaker, to predict the target label based on the shortcut rather than real discriminative features. This is common in medical imaging, where treatment and clinical annotations correlate with disease labels, making them easy shortcuts to predict disease. We propose a novel detection and quantification of the impact of potential shortcut features via a fast diffusion-based counterfactual image generation that can synthetically remove or add shortcuts. Via a novel self-optimized masking scheme we spatially limit the changes made with no extra inference step, encouraging the removal of spatially constrained shortcut features while ensuring that the shortcut-free counterfactuals preserve their remaining image features to a high degree. Using these, we assess how shortcut features influence model predictions.
This is enabled by our second contribution: An efficient diffusion-based counterfactual explanation method with significant inference speed-up at comparable image quality as state-of-the-art. We confirm this on two large chest X-ray datasets, a skin lesion dataset, and CelebA. Our code is publicly available at `https://fastdime.compute.dtu.dk`.

Keywords: Shortcut Learning · Counterfactuals · Diffusion Models

1 Introduction

Shortcut learning [26] denotes the situation in which a model exploits a spurious correlation between a 'shortcut' feature s and the target outcome y. Known examples include grass being used as a shortcut for sheep, or surgical skin markers being shortcuts for malignant skin lesions [77]. Since the correlation is only spurious, and not causal, shortcut learning results in misleadingly high model performance on validation data that also contains shortcuts, but poor ability to generalize to data without shortcuts. As prediction targets are often more complex than shortcuts, deep models can more easily and accurately identify

* N. Weng and P. Pegios contributed equally to this work.

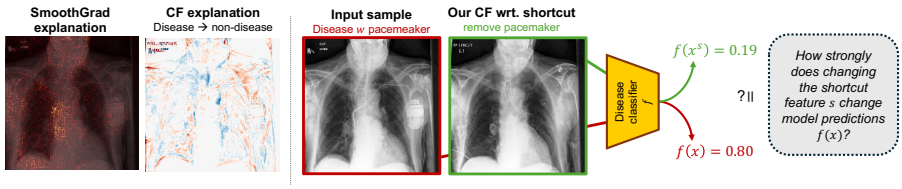

Fig. 1: Shortcut detection: SmoothGrad [71] and CF explanation (*Left*), as two XAI methods, indicate which region in the image could influence the model decision (e.g. from disease to non-disease). Although this includes the shortcut features, it does not clearly indicate it. Therefore an expert is required for further visual inspection. Our counterfactual approach (*Right*) only removes the desired shortcut attribute. With this we can validate if that specific attribute played a role in the model decision. SmoothGrad visualization: highlighting areas crucial to the model decisions. CF explanation: difference map of the original image and CF (*blue*/*red*: information removal/addition).

the shortcuts in the data [28,35,54]. Therefore, shortcut learning is a key obstacle to achieving well-generalized image classification models. This is especially true in medical imaging, where classification tasks often suffer from high intrinsic uncertainty and shortcuts are highly effective in improving classification performance [6,19,35,40,55,77].

Detecting shortcut learning is non-trivial. Stratified analyses of model performance on samples with and without the potential shortcut might yield some indication, but cannot confirm shortcut learning as any performance gaps could result from other distributional differences. Explainable AI (XAI) methods have been proposed to obtain more concrete evidence [56,72,73,78], but require manual inspection of explanations, complicating further use by explanation instability [4,70] and infinite sets of feasible competing explanations [45,58] (see Fig. 1).

We propose a novel approach leveraging diffusion-based counterfactual (CF) generation [5,38] to generate 'shortcut counterfactuals', that *do* or *do not* contain a shortcut feature of interest. Different from standard counterfactual explanations which change the target label, we rather seek to change the shortcut label. By assessing how predictions change with addition or removal of the shortcut feature, we quantify the model's degree of shortcut learning (see Fig. 1).

Our main contributions are the following:

1. A diffusion-based method FastDiME for counterfactual generation, employing approximate gradients in sampling, achieving a 20x speedup over comparable state-of-the-art models while maintaining counterfactual quality.
2. A novel self-optimized masking scheme that confines counterfactual changes to a small region, with which we mitigate the unintentional removal of non-shortcut features in counterfactual images.
3. A novel pipeline for detecting and quantifying shortcut learning via the generation of shortcut counterfactuals from our diffusion-based method.
4. A demonstration of the generality of our method for counterfactual generation as well as its utility in detecting and quantifying shortcut learning in multiple realistic scenarios from different domains.

2 Related work

Counterfactual image explanations. A *counterfactual* describes what the world, some outcome, or a given image would have looked like if some factor c would have had a different value. The most prominent use of counterfactual inference in image analysis is in generating counterfactual explanations which represent images that are similar to a given image but would have been classified differently by a given image classification model. In contrast to adversarial examples, not every minimally changed image that leads to a different classifier output represents a counterfactual example as this should also be realistic, i.e., close to the data manifold, and preserve the semantic properties of the original image. Many methods have been proposed for visual counterfactual explanations, including VAEs [41, 63], GANs [37, 44, 53, 69], flow-based [22] and diffusion-based models [5,38,39,66]. In the medical imaging domain, different methods for generating healthy or diseased counterfactual images have been proposed [17, 24, 50, 68, 73].

In contrast to most of the existing literature, our ultimate goal is *not* only to generate counterfactual explanations of classifier decisions. Instead, we seek to change the state of a potential shortcut feature, such as a person's smile or the presence or absence of a cardiac pacemaker, from a given image, without changing the target class. In this sense, our work is closely related to [57, 59] which use StarGAN [15] for generating demographic counterfactuals in different medical image modalities, and to [18] which combines a hierarchical VAE with structural causal models to perform principled counterfactual inference with respect to demographic attributes in brain MRI and chest X-ray images. These methods target demographic attributes, however, we focus on other, often more localized potential shortcut features such as the presence or absence of cardiac pacemakers or chest drains. Notably, our method can serve dual purposes: a) as a counterfactual explanation method for explaining classifiers' decisions and b) as an image generation tool for our shortcut detection pipeline.

Diffusion-based counterfactuals. Denoising Diffusion Probabilistic Models [32] (DDPMs) have been successfully used for generating counterfactual explanations [5, 38, 39, 66]. DiME [38] is the first method to adapt the original formulation of classifier guidance [21] for counterfactual generation but with a great increase in computational cost, since it requires back-propagation through the whole diffusion process to obtain gradients with respect to a noisy version of the input. DVCE [5] introduces a cone-projection approach, assuming access to an adversarially robust copy of the classifier. Diff-SCM [66] shares the encoder between the target model and the denoiser, making it model-specific. ACE [39] uses a DDPM to turn adversarial attacks into semantically meaningful counterfactuals. Recent studies [75] report the high memory requirements and run time of diffusion-based methods as major challenges for large-scale evaluations. Inspired both by DiME [38] and ACE [39], our method speeds up counterfactual generation and reduces memory usage significantly.

Guided diffusion and image editing. Many approaches have been built to manipulate input images. From these techniques, the most relevant works include Universal Guidance [8], Motion Guidance [27], and GMD [42]. These methods improve the efficiency of the DiME approach by feeding the denoised image into the guiding classifier, and backpropagating the classifier gradients through the denoiser to calculate the gradients wrt. the input image. We, instead, use the denoised image to approximate the gradients for the input image and avoid a heavy back propagation step. The denoised image gradients act as a surrogate for the image gradients as they asymptotically reach the same image.

Shortcut learning detection. XAI-based methods provide explanations for individual decisions and thus may highlight the reliance of the model on potential shortcut features [19,56,72,73,78]. While promising, these methods do not easily allow for quantitative analyses, require inspecting individual explanations, are limited to the detection of spatially localized shortcut features, and are subject to the general challenges of reliability and (non-)uniqueness [4,45,58,70] making their use in helping users recognize the presence of shortcut learning limited [1,2]. Thus, several methods have been proposed for finding potential shortcut features in a dataset [51,54,79], however, they do not evaluate whether a given model has indeed learned to exploit these shortcuts. In [40], model performance is stratified by the presence or absence of a potential shortcut feature, finding large differences in average model performance between these groups. Yet, these performance variations might be influenced by other confounding factors. In [11], models are retrained multiple times using a shortcut feature prediction head in a multi-task fashion, aiming to control and assess the degree to which the shortcut is encoded. Other approaches [52,77], assess how adding surgical skin markings or colored patches to dermoscopic images affects model confidence. However, their shortcut features are simple and relatively easy to add or remove compared to, e.g., a cardiac pacemaker in a chest X-ray. To the best of our knowledge, no prior work has investigated the use of shortcut counterfactuals to quantify the degree to which a model relies on the shortcut feature.

3 Methods

Counterfactual image explanation aims to solve the following problem: Assume given an image classification problem with a particular class c of interest – for instance, whether or not a chest X-ray contains a cardiac pacemaker – and an image x that does not belong to class c. Can we provide an updated image x^c that remains as close as possible to x while both being visually realistic and undergoing sufficient visual change to clearly belong to the class c? In our medical imaging case, this could consist of artificially adding or removing the pacemaker without changing any of the remaining patient anatomy.

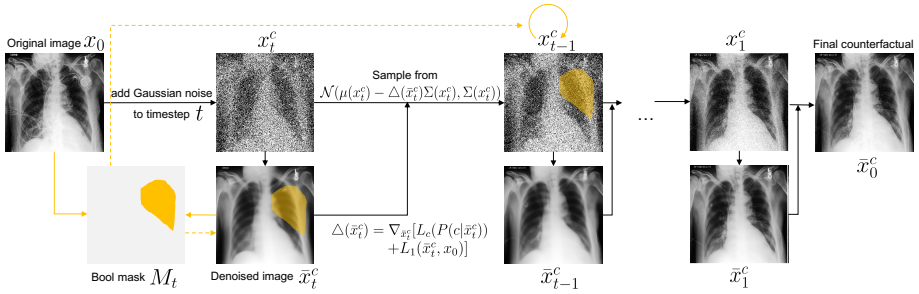

Fig. 2: Proposed FastDiME method. In each step, noised image x_t^c is sampled with the guidance of the counterfactual loss, leveraging information derived from the denoised image $\bar{x}_t^c$. A self-optimized mask is automatically extracted and applied to prevent changes in regions less relevant to the task at each time step.

3.1 Diffusion Models for Counterfactual Explanations (DiME)

Jeanneret et al. [38] utilize DDPMs for generating counterfactual explanations through a guided diffusion process [21]. The image is guided towards the counterfactual class using the classifier's loss objective L_c, while the counterfactual x^c is constrained to remain close to the original x through an L_1 loss and a perceptual loss L_{perc}. The overall gradient for the counterfactual loss term is of the form $\nabla_{CF} = \lambda_c \nabla L_c + \lambda_1 \nabla L_1 + \lambda_p \nabla L_{perc}$, where λ_c, λ_1 and λ_p are hyperparameters, and L_1 is measured between a noisy version and the original image. To obtain meaningful gradients with their classifier, another nested diffusion process is used to synthesize (unconditionally) an image at the cost of running a whole DDPM process for each step. DiME [38] is summarized as follows:

- Corrupt input up to noise level τ with the forward process.
- For every time step $t \in \{\tau, \dots, 0\}$, do:
  1. Denote noisy version of the counterfactual image x_t^c .
  2. Compute the gradients $\nabla_{CF}(\hat{x}_t^c)$ based on a clean image $\hat{x}_t^c$ synthesized using an inner forward process continuing the unconditional generation process from the current step t, i.e, $\hat{x}_t^c = \text{DDPM}(x_t^c, t)$.
  3. Sample x_{t-1}^c from $\mathcal{N}(\mu_g(x_t^c), \Sigma(x_t^c))$, where $\mu_g(x_t^c)$ is the guided mean, $\mu_g(x_t^c) = \mu(x_t^c) - \nabla_{CF} * \Sigma(x_t^c)$ and $\mu(x_t^c)$ is the estimated mean, following the standard guided diffusion scheme [21].
- Return the counterfactual image as $x^c = x_0^c$.

In this process, step 2 is an expensive bottleneck. We therefore propose a modification that significantly reduces complexity and improves stability.

3.2 Fast generation of high-quality counterfactuals

We present FastDiME, which improves on DiME with an efficient gradient estimation and a novel self-optimized masking scheme (see Fig. 2).

Efficient gradient estimation. As explained above, gradients in the original DiME model are retrieved from generated images by re-running the entire DDPM process. This is computationally expensive, with complexity $\mathcal{O}(T^2)$. To speed up inference while maintaining image quality, we propose using the denoised $\bar{x}_t^c$ to calculate the gradients. In the forward diffusion process (q), the noisy image x_t from x_{t-1} is synthesized using a Gaussian distribution at each timestep with scheduled variance β:

$$q(x_t^c|x_{t-1}^c) \coloneqq \mathcal{N}(x_t^c; \sqrt{1-\beta_t}x_{t-1}^c, \beta_t \boldsymbol{I}). \tag{1}$$

As derived by Ho et al. [32], a direct noising process conditioned on the input x_0^c is feasible by marginalizing Eq. (1),

$$q(x_t^c|x_0^c) = \mathcal{N}(x_t^c; \sqrt{\bar{\alpha}_t}x_0^c, (1-\bar{\alpha}_t)\boldsymbol{I}), \tag{2}$$

$$x_t^c = \sqrt{\bar{\alpha}_t}x_0^c + \sqrt{(1-\bar{\alpha}_t)}\epsilon, \tag{3}$$

where $\alpha_t \coloneqq 1-\beta_t$, $\bar{\alpha}_t \coloneqq \prod_{s=0}^{t}\alpha_s$ and $\epsilon \sim \mathcal{N}(0, \boldsymbol{I})$. Given a noise estimate $\bar{\epsilon}$ (from a denoiser), we can use Eq. (3) to retrieve the estimated denoised $\bar{x}_t^c$ at time-step t from the noisy image x_t^c:

$$\bar{x}_t^c = \frac{x_t^c - \sqrt{(1-\bar{\alpha}_t)}\bar{\epsilon}}{\sqrt{\bar{\alpha}_t}}. \tag{4}$$

By definition of the MSE-optimal denoiser [9, 61], we have

$$\bar{x}^c = \bar{x}_0^c = \mathbb{E}[\hat{x}_0^c], \tag{5}$$

which is used for gradient calculation in counterfactual generation. The denoised image $\bar{x}_0^c$ is the expected value of all possible noise-free images, i.e. the intermediate solution of DiME [3]. The denoised image serves as a good surrogate, because: As noise decreases during optimization, the denoised image asymptotically converges to the noise-free sample. Thus, with the mild assumption of continuity of classifier gradients at the input, the gradients of the denoised image converge to those of the noise-free image. The time complexity of adopting denoised images is $\mathcal{O}(T)$, which significantly expedites the entire process (see Sec. 4.3).

Fig. 3 illustrates the convergence of the proposed FastDiME approach compared to DiME. In addition to the considerable advantage of improved time efficiency, using the denoised image does not degrade the generating quality, as the denoised image represents the expected value of all possible paths of generated images. Thus, the gradients computed on denoised images diminish noise levels, resulting in faster convergence with a more direct convergence route.

Self-optimized masking. Our application of counterfactual generation is to remove shortcut features from images. Shortcut features tend to be highly localized; thus, it is often desirable to constrain changes to the image made by the counterfactual generation process to a small region of the image. To achieve this

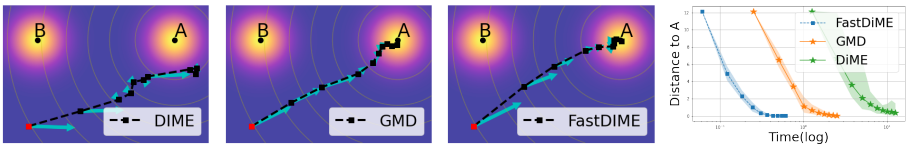

Fig. 3: Toy 2D example for visualization of the counterfactual generation convergence comparing DiME, GMD, and FastDiME. Given the two class distributions A and B, each method tries to bring the initial point x_t (red point) to class A. Blue vectors indicate the gradients of classification loss ∇L_c at each step. When calculating this gradient at each step, DiME uses a new unconditional sample from DDPM, which could lead to noisy gradients. In contrast, FastDiME uses the expected image $\bar{x}_0$ at each step to calculate the gradients, which results in a more stable convergence. The plot on the right shows the convergence of each method in terms of the distance to A. This is averaged over 100 runs and it indicates that FastDiME is accurate and significantly faster than both the other methods.

goal, Jeanneret et al. [39] perform an extra inpainting step after the usual counterfactual generation step using RePaint [49] to ensure highly localized changes to the image. However, due to the in-painting step lacking classifier guidance, their approach does not guarantee that the generated images remain valid counterfactuals – in principle, the class label could change again during the inpainting step. Inspired by their approach, we integrate self-optimized masking directly into the initial counterfactual generation process to preserve more features from the original image, while incorporating other terms in the optimization. To this end, we first extract a mask M_t by binarizing the difference between the denoised image at time-step t and original input x_0, and then masking the sampled image x_t^c and denoised images $\bar{x}_t^c$ accordingly, i.e.,

$$M_t = \delta(\bar{x}_t^c, x_0) \tag{6}$$

$$x_t^{c\prime} = x_t^c \cdot M_t + x_t \cdot (1 - M_t) \tag{7}$$

$$\bar{x}_t^{c\prime} = \bar{x}_t^c \cdot M_t + x_0 \cdot (1 - M_t) \tag{8}$$

where $x_t \sim q(x_t|x_0, t)$ refers to the sample from the forward process on the original image x_0 that adds a Gaussian noise to it, and δ represents the function used for extracting mask. Note that the mask is not given, but automatically generated by our pipeline, and it can vary with each time-step t. Different from [39], our fully gradient-guided masking approach ensures the validity of the counterfactuals. We implement this inside our main process after a warm-up period of τ_w time-steps, with $1 \leq \tau_w \leq \tau$. We further experiment with 2-step approaches keeping M fixed after a completed run of guided diffusion with efficient gradient estimations or a completed run of our full method including our self-optimized mask scheme. We refer to these as FastDiME-2 and FastDiME-2+, respectively.

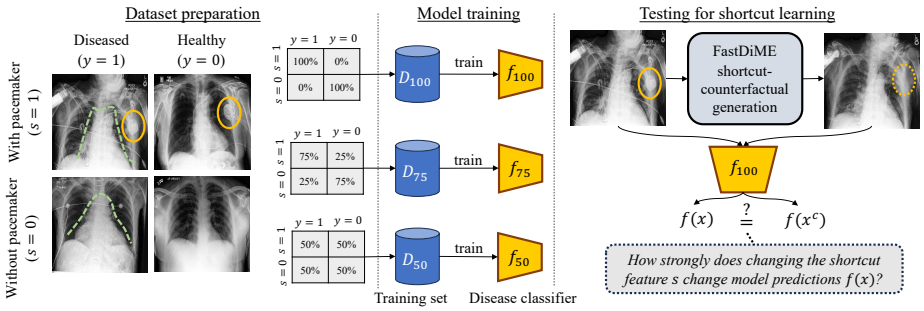

Fig. 4: Validating shortcut learning detection pipeline. In order to test the shortcut detecting ability by shortcut-counterfactuals, we construct three synthetic training datasets with varying degrees of correlation between the shortcut feature s and the target label y and train classifiers based on them (*Left* and *Middle*). By measuring the difference in confidence leveling between the original image and shortcut-counterfactual, the degree of shortcut learning is examined (*Right*). If model predictions differ strongly between natural images and shortcut counterfactual images (while leaving the target label unchanged), shortcut learning has occurred.

3.3 Shortcut learning detection

Assume a classifier $f(x) = \hat{y}$ that we suspect of shortcut learning, different from the one used for counterfactual image generation. For example, f trained to diagnose lung diseases from chest X-ray images but instead might use medical equipment such as cardiac pacemakers or chest drains as shortcuts. We propose the pipeline for quantifying the degree to which f has indeed learned to rely on the shortcut feature as shown in Fig. 4 (*Right*), where we generate the shortcut counterfactuals of all images, i.e. images that remove the pacemaker, and evaluate the alteration of predictions.

In order to validate the proposed counterfactual-based shortcut detection pipeline, we propose the following framework (see Fig. 4):

1. Curate training sets with increasing levels of encoded correlation between the shortcut feature $s = 1$ and the target label $y = 1$, and train models on the increasingly contaminated training sets. In our experiments, we curate three training sets $\mathcal{D}_k, k = \{100, 75, 50\}$, in which $k\%$ of $y = 1$ samples are also positive in s. It is worth noting that $\mathcal{D}_{100}$ is extremely biased by shortcut while $\mathcal{D}_{50}$ is completely balanced.
2. Curate natural test sets, along with a test set where the suspected shortcut feature is counterfactually flipped using FastDiME. In particular, curate
   (a) an i.i.d. test set test_k with the same correlation rate as the training set,
   (b) a balanced test set test_u where each class (shortcut and target label) is represented equally with 50% positive/negative samples,
   (c) a counterfactual balanced test set called test_u^c, which is obtained by generating shortcut counterfactuals x^c for each x in the test_u.

3. Measure the difference in confidence level between original and shortcut-flipped counterfactual images.

If the effect of the shortcut flipping operation on a classification model f is stronger for the models trained on more biased datasets, this is a clear indication that the models are, indeed, prone to learning the suspected shortcut.

4 Experiments and results

4.1 Datasets and implementation details

To ease comparison with existing work on counterfactual explanations [5, 37–39, 41, 64], we evaluate on **CelebA** [47], containing 128×128 images of faces. First, we compare our counterfactual quality against state-of-the-art methods on the standard 'smile' and 'age' benchmarks. To assess our shortcut detection pipeline, we set 'smile' as a shortcut in predicting 'age'. Furthermore, we validate our method on three real-world, challenging medical datasets. **CheXpert** [34] and **NIH** [76] are chest X-ray datasets with two different suspected shortcuts annotated: cardiac pacemakers [35] and chest drains [40, 55], respectively. Note that 'pacemaker' and 'chest drain' are potential shortcuts to diagnosis as they are common treatments for many cardio and lung diseases, respectively. **ISIC 2018** [16, 74] is a skin lesion dataset for diagnosing malignant or benign lesions with 'ruler markers' as shortcuts [62]. For all medical datasets, images are resized to 224×224, and we evaluate our method on shortcut counterfactuals, while diagnostic labels are used in shortcut detection experiments.

Implementation details. For CelebA, we use the same trained DenseNet121 classifier [33] and DDPM [32] as in DiME [38] and ACE [39] for fair comparisons. For all variants of our method, namely, FastDiME, FastDiME-2, and FastDiME-2+, we follow the same hyperparameters as in DiME [38]. For the medical datasets, we train DDPMs with 1000 steps, use UNet's [65] encoder architecture as the shortcut classifier for counterfactual generation, set τ to 160 out of 400 re-spaced time steps, and compute L_1 between the denoised and original input without including L_{perc}. For all datasets, we set $\tau_w = \frac{\tau}{2}$, and normalize, threshold and dilate the masks similar to ACE [39]. For the shortcut detection experiments, the task classifier suspected of shortcut learning is a ResNet-18 [30].

4.2 Counterfactual explanations

Evaluation criteria. We follow the evaluation protocol of ACE [39] on CelebA [47]. To measure *realism* we use FID [31] (Fréchet Inception Distance) between the original images and their valid counterfactuals as well as sFID proposed in [39] to remove potential biases of FID. We estimate *sparsity* using standard metrics for face attributes such as Mean Number of Attributes Changed (MNAC) which utilizes an oracle pre-trained on VGGFace2 [13] finetuned on CelebA [47], as well as Correlation Difference (CD) proposed in [38] to account for MNAC's

Table 1: Results for CelebA. Results for DiVE [63], DiVE^{100}, STEEX [37], DiME [38] and ACE [39] are from [39]. We compute BKL, MAD and FR metrics for diffusion-based methods, and highlight the **best** and *second-best* performances.

	Smile										Age									
Method	FID	sFID	FVA	FS	MNAC	CD	COUT	BKL	MAD	FR	FID	sFID	FVA	FS	MNAC	CD	COUT	BKL	MAD	FR
DiVE	29.4	-	97.3	-	-	-	-	-	-	-	33.8	-	98.1	-	4.58	-	-	-	-	-
DiVE^{100}	36.8	-	73.4	-	4.63	2.34	-	-	-	-	39.9	-	52.2	-	4.27	-	-	-	-	-
STEEX	10.2	-	96.9	-	4.11	-	-	-	-	-	11.8	-	97.5	-	3.44	-	-	-	-	-
ACE ℓ_1	**1.27**	**3.97**	**99.9**	**0.874**	2.94	*1.73*	**0.783**	0.199	0.72	0.976	**1.45**	**4.12**	*99.6*	*0.782*	3.20	*2.94*	**0.718**	0.266	0.60	0.962
ACE ℓ_2	*1.90*	*4.56*	**99.9**	*0.867*	**2.77**	**1.56**	*0.624*	0.326	0.59	0.843	*2.08*	*4.62*	*99.6*	**0.797**	2.94	**2.82**	*0.564*	0.390	0.48	0.775
DiME	3.17	4.89	98.3	0.729	3.72	2.30	0.526	0.094	*0.82*	0.972	4.15	5.89	95.3	0.671	3.13	3.27	0.444	0.162	**0.71**	0.990
FastDiME	4.18	6.13	*99.8*	0.758	3.12	1.91	0.445	*0.097*	*0.82*	*0.990*	4.82	6.76	99.2	0.738	2.65	3.80	0.356	0.181	*0.69*	0.986
FastDiME-2	3.33	5.49	**99.9**	0.773	3.06	1.89	0.439	**0.083**	**0.83**	**0.994**	4.04	6.01	*99.6*	0.750	*2.63*	3.80	0.369	**0.157**	**0.71**	**0.993**
FastDiME-2+	3.24	5.23	**99.9**	0.785	*2.91*	2.02	0.411	0.098	*0.82*	0.989	3.60	5.59	**99.7**	0.766	**2.44**	3.76	0.323	0.179	*0.69*	*0.987*

Table 2: Results for medical datasets, highlighting the **best** and *second-best* ones.

	CheXpert			NIH			ISIC		
Method	L_1	MAD	FID	L_1	MAD	FID	L_1	MAD	FID
DiME	0.1107	**0.7977**	73.2588	0.0748	**0.2536**	94.3023	0.0779	**0.5240**	121.5135
FastDiME	*0.0897*	0.7554	**61.4010**	*0.0546*	0.2244	*64.1100*	*0.0631*	0.3732	**86.0475**
FastDiME-2	0.0946	*0.7596*	64.0955	0.0584	*0.2386*	67.8026	0.0661	*0.4428*	87.4575
FastDiME-2+	**0.0894**	0.7581	*62.2840*	**0.0536**	0.2263	**63.2997**	**0.0621**	0.3905	*86.9823*

limitations. Moreover, we use Face Verification Accuracy (FVA) [13] and Face Similarity (FS) [39] to measure whether a counterfactual changed the face identity. We also compute the *transition probabilities* between the original image and its counterfactual with the COUnterfactual Transition (COUT) metric [43]. Moreover, we include Bounded remapping of KL divergence (BKL), used in [38] to calculate the *similarity* between prediction and the desired one-hot counterfactual label, with lower values indicating higher similarity. To measure the *validity* of counterfactuals we report Flip Ratio (FR), i.e., the frequency of counterfactuals classified as the target label, and Mean Absolute Difference (MAD) of confidence prediction between original and counterfactual images. For medical datasets, we evaluate the quality of counterfactuals in *closeness* and *realism* using L_1 distance and FID respectively, and *validity* using MAD.

Results. Tab. 1 and Tab. 2 list the results of the counterfactual explanation experiments on CelebA and medical datasets, respectively. For most of the datasets, FastDiME, and its variants outperform DiME in most of the metrics and remain competitive against ACE in image quality. ACE achieves a significantly better COUT as it is sufficient to cross the decision boundary, whereas we explicitly maximize the counterfactual class probability. Results in terms of BKL, MAD, and FR demonstrate that our method indeed produces samples that maximize the counterfactual class probability. Further illustrating this, Fig. 6 shows that our methods tend to produce stronger counterfactual 'smiling' impressions than ACE on CelebA. Fig. 5 shows generated counterfactuals on the medical datasets, successfully removing the shortcut feature from the images without significantly altering the rest of the image.More examples can be found in the Appendix.

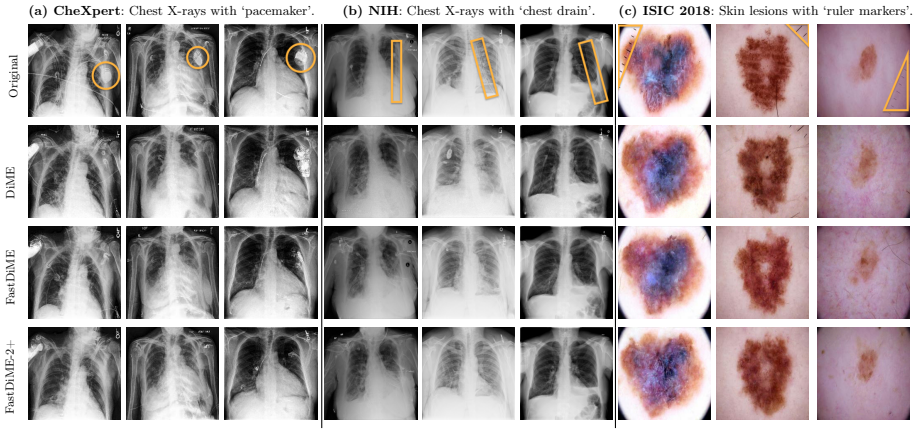

Fig. 5: Shortcut counterfactuals for medical datasets. The shortcuts are highlighted with orange circles or boxes in the original images.

Fig. 6: CelebA counterfactual explanations for the 'smile' attribute.

4.3 Efficiency analysis

We compare diffusion-based methods in terms of inference time and GPU memory usage (MU) in MiB. Using a random subset of 1000 images from CelebA, we generated counterfactuals for the 'smile' attribute with a batch size of 5 on an RTX 5000 Turing 16 GB NVIDIA GPU. We report the average batch time in seconds and the total time in hours together with the theoretical complexity in Tab. 3. Note that we could not fit batch sizes larger than 10 for ACE [39].

4.4 Shortcut learning detection pipeline

We train ResNet-18 [30] classifiers f_k initialised with ImageNet [20] pre-trained weights on each of the datasets $\mathcal{D}_k, k = \{100, 75, 50\}$ and evaluate them on the three curated test sets. Task labels and shortcuts are shown in Tab. 4.

Table 3: Complexity, time, and memory consumption. FastDiME (w/o Mask) refers to our method without self-optimized mask scheme. T denotes the number of diffusion steps and N is the number of adversarial attack update steps in ACE. Total Time (hours) is for a random subset of 1000 CelebA images, Batch Time (seconds) is calculated with a batch of 5, and GPU MU is in MiB.

Method	Complexity	Batch Time	Total Time	GPU MU
DiME	$\mathcal{O}(T^2)$	$218.8_{\pm 72.6}$	12:10:08	**1.6K**
ACE ℓ_1	$\mathcal{O}(NT)$	$41.8_{\pm 0.7}$	02:17:46	13.1K
ACE ℓ_2	$\mathcal{O}(NT)$	$42.9_{\pm 0.1}$	02:18:43	13.1K
GMD	$\mathcal{O}(T)$	$31.4_{\pm 6.6}$	01:45:45	4.2K
FastDiME (w/o Mask)	$\mathcal{O}(T)$	$\mathbf{10.6}_{\pm \mathbf{4.0}}$	**00:35:56**	**1.6K**
FastDiME	$\mathcal{O}(T)$	$\mathit{13.6}_{\pm 5.1}$	*00:45:55*	**1.6K**
FastDiME-2	$\mathcal{O}(T)$	$23.1_{\pm 9.3}$	01:17:40	**1.6K**
FastDiME-2+	$\mathcal{O}(T)$	$27.0_{\pm 10.4}$	01:30:38	**1.6K**

Evaluation criteria. To *quantify* the shortcut learning effect using counterfactuals, we measure the differences in model predictions resulting from counterfactually changing s of the test images. To this end, we measure MAD between original images (test_u) and shortcut counterfactuals (test_u^c), and their Mean confidence Difference (MD) across two subtests $X_{s=0}$ and $X_{s=1}$ according to their true shortcut label s. Furthermore, in order to *validate* the extent of shortcut learning apart from the unbalanced dataset setting, we also evaluate each of the classifiers f_k trained on different correlation levels k with the shortcut feature s using Area Under the Receiver Operating Characteristic (AUROC).

Results. Tab. 4 and Fig. 7 shows the results of our shortcut learning detection experiments. The magnitude of model prediction changes in terms of MAD and MD metrics between original test images and FastDiME shortcut counterfactuals correctly signifies stronger reliance of the models on the shortcut feature for the models trained on the more strongly biased training sets. The significant difference between $test_k$ and $test_u$ in terms of AUROC also indicates that the trained classifiers indeed learn to exploit the shortcut, with the severity of shortcut learning being correlated with the strength $k\%$ of the association in the training set. Fig. 7, demonstrates that the proposed pipeline is indeed appropriate for detecting and quantifying shortcut learning in practice. It is worth noting that the same trend is evident with more powerful backbones including ConvNeXt [46] and ViT [23] (see Appendix). Furthermore, notice that AUROC is only used as an indicator, as it is not a reliable measurement for shortcut learning. The difference between the AUROC of $test_k$ and $test_u$ can only tell whether the shortcut is correlated with the target label, but it can not exclude other potential reasons and cannot reveal a causal relationship. If pacemakers are highly correlated with text markers in X-rays, the AUROC results cannot determine which is the shortcut. It should also be noted that the proposed shortcut detection pipeline is designed to obviate the need for diagnostic labels for new data within $test_u$, thereby facilitating any new databases after the trained counterfactual pipeline.

Dataset	Task Label	Shortcut	Train Set	AUROC			Shortcut Detection Metrics		
				test_k	test_u	test_u^c	MAD	$\text{MD}_{(s=1)}$	$\text{MD}_{(s=0)}$
CheXpert	cardiomegaly	pacemaker	$\mathcal{D}_{100}$	0.98	0.58	0.63	0.36	0.40	-0.26
			$\mathcal{D}_{75}$	0.80	0.71	0.74	0.15	0.14	-0.03
			$\mathcal{D}_{50}$	0.72	0.73	0.74	0.12	-0.01	-0.02
NIH	pneumothorax	chest drain	$\mathcal{D}_{100}$	0.98	0.65	0.65	0.13	0.04	-0.22
			$\mathcal{D}_{75}$	0.87	0.73	0.71	0.07	0.01	-0.10
			$\mathcal{D}_{50}$	0.70	0.71	0.70	0.03	-0.01	-0.02
ISIC	malignant	ruler markers	$\mathcal{D}_{100}$	0.98	0.70	0.79	0.19	0.15	-0.20
			$\mathcal{D}_{75}$	0.86	0.82	0.82	0.10	0.01	-0.09
			$\mathcal{D}_{50}$	0.87	0.85	0.81	0.09	-0.04	-0.03
CelebA	age	smile	$\mathcal{D}_{100}$	0.98	0.65	0.73	0.32	0.36	-0.27
			$\mathcal{D}_{75}$	0.91	0.84	0.83	0.17	0.18	-0.04
			$\mathcal{D}_{50}$	0.87	0.87	0.85	0.14	0.09	0.04

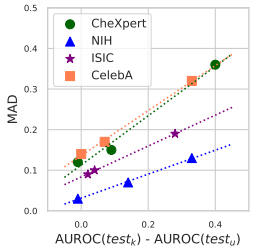

Table 4 & Fig. 7: Results of proposed shortcut detection pipeline. The table (*left*) details the prediction performance on the distinct test sets, $test_k$ (same distribution as trainset) and $test_u$ (unseen balanced set), alongside the associated shortcut detection metrics. The observed performance discrepancy between $test_k$ and $test_u$ quantitatively indicates the extent to which shortcuts affect the main task's predictive accuracy, which correlates with the shortcut detection metrics, as illustrated in the figure (*right*). Although AUROC indicates the presence of shortcuts, it is not a reliable measurement, as it only shows the correlation but no causality.

5 Discussion, limitations and conclusion

Adversarially vulnerable classifier. We find, across methods, that the generalization performance of the classifier used for generating counterfactuals is crucial. In addition to classifying samples within the data distribution, the counterfactual generation process introduces out-of-distribution generated samples. This causes most counterfactual failures, as the model ceases to guide once the prediction flips. Examples are demonstrated in Fig. 8. A potential solution would be to use an adversarially robust classifier, to obtain more informative gradients [25] and realistic generated images [10, 54, 67]. This solution mostly applies for counterfactual image synthesis, *not* for counterfactual explanations, as for the latter the goal is to explain any classifier, irrespective of its robustness.

Denoised images and masking parameter τ_w. As expected, our efficient gradient estimation produces denoised images $\bar{x}_t^c$ that tend to be more blurred during the early time steps compared to the expensive generated images $\hat{x}_t^c$ from DiME. Yet, as denoised images diminish noise levels and shortcut features are often highly localized and less sensitive to blurriness, our counterfactuals do not degrade image quality. To further improve them, we apply our self-optimized mask scheme as long as $\bar{x}_t^c$ is not blurred, which is around $\frac{\tau}{2}$. We explored quantitatively and quantitatively the effects of our self-optimized masking in the Appendix. .

Unbiased shortcut generation. One limitation of our work is that it depends on unbiased shortcut classification and image generation – image generation models have also been shown to reproduce (potentially spurious) correlations from training [48]. A counterfactual generator might thus not only change the shortcut

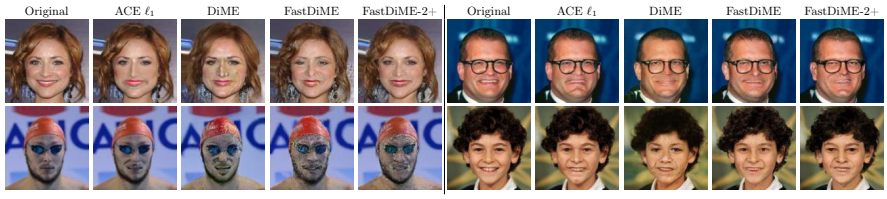

Fig. 8: Limitations for valid counterfactuals. The left-top case shows a hard example (smile → no smile) where all method fails to make semantically meaningful changes. In the left-bottom (no smile → smile), all DiME-based methods fail to produce a suitable sample, while ACE's result appears natural but lacks a noticeable smile alteration. In the right-top (smile → no smile), we observe similar performance for ACE, while our 2-step approach offers a visually better result. Right-bottom case (smile → no smile) shows an example where all methods change more than the targeted smile.

feature but also discriminative ones, e.g., the disease label. We mitigate this by training our shortcut classifier for medical datasets only on diseased samples, and by using self-optimized masks to limit the spatial change of the counterfactual. Our method can sometimes result in counterfactuals that change more than the targeted shortcut (see Fig. 8). Further development of the self-optimized masking and reducing the effect of spurious correlations on the generative shortcut model remain important avenues for future work. Note that it is often easier to train an unbiased shortcut classifier as opposed to for: Shortcut features are typically easier to predict, which is precisely what makes them vulnerable.

Types of shortcut features. In medical imaging, shortcut features vary from demographic features [12, 29] to machine processing [36] and treatments [40]. Our work focuses on localized shortcuts, and FastDiME is designed specifically to cope with them, as emphasized in Sec. 2. Yet, there is a clear potential to be extended to support non-local shortcut features by adding loss terms to capture these global features or altering hyper-parameters to allow more global changes.

Using shortcut counterfactuals to mitigate shortcut learning. While we focus on generating the shortcut counterfactuals and using them to *detect* shortcut learning, they also have the potential to *mitigate* shortcut learning by augmenting the training set with samples that have the shortcut feature added or removed. Similar approaches have been successful in other domains [7, 14, 52, 57, 60].

Conclusion. We present a general and fast method for diffusion-based counterfactual explanations, which is up to 20 times faster than the existing methods while preserving comparable counterfactual quality. We further demonstrate how our method can be used for generating counterfactuals with and without suspected shortcuts in both medical datasets and CelebA. Based on this, we introduce a novel pipeline to automatically detect shortcut learning in practice, eliminating the need for visual inspection which is typical in standard XAI approaches.

Acknowledgements

Work on this project was partially funded by the Independent Research Fund Denmark (DFF, grant number 9131-00097B), the Pioneer Centre for AI (DNRF grant nr P1), the DIREC project EXPLAIN-ME (9142-00001B), and the Novo Nordisk Foundation through the Center for Basic Machine Learning Research in Life Science (MLLS, grant NNF20OC0062606). The funding agencies had no influence on the writing of the manuscript nor on the decision to submit it for publication.

References

1. Adebayo, J., Muelly, M., Abelson, H., Kim, B.: Post hoc explanations may be ineffective for detecting unknown spurious correlation. In: International Conference on Learning Representations (2022)
2. Adebayo, J., Muelly, M., Liccardi, I., Kim, B.: Debugging tests for model explanations. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., Lin, H. (eds.) Advances in Neural Information Processing Systems. vol. 33, pp. 700–712. Curran Associates, Inc. (2020)
3. Alain, G., Bengio, Y.: What regularized auto-encoders learn from the data-generating distribution. The Journal of Machine Learning Research **15**(1), 3563–3593 (2014)
4. Arun, N., Gaw, N., Singh, P., Chang, K., Aggarwal, M., Chen, B., Hoebel, K., Gupta, S., Patel, J., Gidwani, M., Adebayo, J., Li, M.D., Kalpathy-Cramer, J.: Assessing the trustworthiness of saliency maps for localizing abnormalities in medical imaging. Radiology: Artificial Intelligence **3**(6) (Nov 2021)
5. Augustin, M., Boreiko, V., Croce, F., Hein, M.: Diffusion visual counterfactual explanations. In: Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., Oh, A. (eds.) Advances in Neural Information Processing Systems. vol. 35, pp. 364–377. Curran Associates, Inc. (2022)
6. Badgeley, M.A., Zech, J.R., Oakden-Rayner, L., Glicksberg, B.S., Liu, M., Gale, W., McConnell, M.V., Percha, B., Snyder, T.M., Dudley, J.T.: Deep learning predicts hip fracture using confounding patient and healthcare variables. npj Digital Medicine **2**(1) (2019)
7. Balashankar, A., Wang, X., Packer, B., Thain, N., Chi, E., Beutel, A.: Can we improve model robustness through secondary attribute counterfactuals? In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics (2021)
8. Bansal, A., Chu, H.M., Schwarzschild, A., Sengupta, S., Goldblum, M., Geiping, J., Goldstein, T.: Universal guidance for diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 843–852 (2023)
9. Bigdeli, S.A., Lin, G., Dunbar, L.A., Portenier, T., Zwicker, M.: Learning generative models using denoising density estimators. IEEE Transactions on Neural Networks and Learning Systems (2023)
10. Boreiko, V., Augustin, M., Croce, F., Berens, P., Hein, M.: Sparse Visual Counterfactual Explanations in Image Space, pp. 133–148. Springer International Publishing (2022)

11. Brown, A., Tomasev, N., Freyberg, J., Liu, Y., Karthikesalingam, A., Schrouff, J.: Detecting shortcut learning for fair medical AI using shortcut testing. Nature Communications **14**(1) (jul 2023)
12. Brown, A., Tomasev, N., Freyberg, J., Liu, Y., Karthikesalingam, A., Schrouff, J.: Detecting shortcut learning for fair medical ai using shortcut testing. Nature communications **14**(1), 4314 (2023)
13. Cao, Q., Shen, L., Xie, W., Parkhi, O.M., Zisserman, A.: Vggface2: A dataset for recognising faces across pose and age. In: 2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018). pp. 67–74. IEEE (2018)
14. Chang, C.H., Adam, G.A., Goldenberg, A.: Towards robust classification model by counterfactual and invariant data generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 15212–15221 (June 2021)
15. Choi, Y., Choi, M., Kim, M., Ha, J.W., Kim, S., Choo, J.: StarGAN: Unified generative adversarial networks for multi-domain image-to-image translation. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE (jun 2018)
16. Codella, N., Rotemberg, V., Tschandl, P., Celebi, M.E., Dusza, S., Gutman, D., Helba, B., Kalloo, A., Liopyris, K., Marchetti, M., Kittler, H., Halpern, A.: Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (ISIC) (Feb 2019)
17. Cohen, J.P., Brooks, R., En, S., Zucker, E., Pareek, A., Lungren, M.P., Chaudhari, A.: Gifsplanation via latent shift: A simple autoencoder approach to counterfactual generation for chest x-rays. In: Medical Imaging with Deep Learning (2021)
18. De Sousa Ribeiro, F., Xia, T., Monteiro, M., Pawlowski, N., Glocker, B.: High fidelity image counterfactuals with probabilistic causal models. In: Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., Scarlett, J. (eds.) Proceedings of the 40th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 202, pp. 7390–7425. PMLR (23–29 Jul 2023)
19. DeGrave, A.J., Janizek, J.D., Lee, S.I.: AI for radiographic COVID-19 detection selects shortcuts over signal. Nature Machine Intelligence **3**(7), 610–619 (may 2021)
20. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009)
21. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. Advances in neural information processing systems **34**, 8780–8794 (2021)
22. Dombrowski, A.K., Gerken, J.E., Kessel, P.: Diffeomorphic explanations with normalizing flows. In: ICML Workshop on Invertible Neural Networks, Normalizing Flows, and Explicit Likelihood Models (2021)
23. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
24. Dravid, A., Schiffers, F., Gong, B., Katsaggelos, A.K.: medxgan: Visual explanations for medical classifiers through a generative latent space. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2936–2945 (2022)
25. Etmann, C., Lunz, S., Maass, P., Schoenlieb, C.: On the connection between adversarial robustness and saliency map interpretability. In: Chaudhuri, K., Salakhutdinov, R. (eds.) Proceedings of the 36th International Conference on Machine Learn-

ing. Proceedings of Machine Learning Research, vol. 97, pp. 1823–1832. PMLR (09–15 Jun 2019)
26. Geirhos, R., Jacobsen, J.H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., Wichmann, F.A.: Shortcut learning in deep neural networks. Nature Machine Intelligence **2**(11), 665–673 (2020)
27. Geng, D., Owens, A.: Motion guidance: Diffusion-based image editing with differentiable motion estimators. arXiv preprint arXiv:2401.18085 (2024)
28. Gichoya, J.W., Banerjee, I., Bhimireddy, A.R., Burns, J.L., Celi, L.A., Chen, L.C., Correa, R., Dullerud, N., Ghassemi, M., Huang, S.C., Kuo, P.C., Lungren, M.P., Palmer, L.J., Price, B.J., Purkayastha, S., Pyrros, A.T., Oakden-Rayner, L., Okechukwu, C., Seyyed-Kalantari, L., Trivedi, H., Wang, R., Zaiman, Z., Zhang, H.: AI recognition of patient race in medical imaging: a modelling study. The Lancet Digital Health **4**(6), e406–e414 (jun 2022)
29. Gichoya, J.W., Banerjee, I., Bhimireddy, A.R., Burns, J.L., Celi, L.A., Chen, L.C., Correa, R., Dullerud, N., Ghassemi, M., Huang, S.C., et al.: Ai recognition of patient race in medical imaging: a modelling study. The Lancet Digital Health **4**(6), e406–e414 (2022)
30. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
31. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: GANs trained by a two time-scale update rule converge to a local nash equilibrium. Advances in neural information processing systems **30** (2017)
32. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)
33. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4700–4708 (2017)
34. Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., Marklund, H., Haghgoo, B., Ball, R., Shpanskaya, K., Seekins, J., Mong, D.A., Halabi, S.S., Sandberg, J.K., Jones, R., Larson, D.B., Langlotz, C.P., Patel, B.N., Lungren, M.P., Ng, A.Y.: CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison. Proceedings of the AAAI Conference on Artificial Intelligence **33**(01), 590–597 (Jul 2019)
35. Jabbour, S., Fouhey, D., Kazerooni, E., Sjoding, M.W., Wiens, J.: Deep learning applied to chest x-rays: Exploiting and preventing shortcuts. In: Doshi-Velez, F., Fackler, J., Jung, K., Kale, D., Ranganath, R., Wallace, B., Wiens, J. (eds.) Proceedings of the 5th Machine Learning for Healthcare Conference. Proceedings of Machine Learning Research, vol. 126, pp. 750–782. PMLR (07–08 Aug 2020)
36. Jabbour, S., Fouhey, D., Kazerooni, E., Sjoding, M.W., Wiens, J.: Deep learning applied to chest x-rays: exploiting and preventing shortcuts. In: Machine Learning for Healthcare Conference. pp. 750–782. PMLR (2020)
37. Jacob, P., Zablocki, É., Ben-Younes, H., Chen, M., Pérez, P., Cord, M.: STEEX: steering counterfactual explanations with semantics. In: European Conference on Computer Vision. pp. 387–403. Springer (2022)
38. Jeanneret, G., Simon, L., Jurie, F.: Diffusion models for counterfactual explanations. In: Proceedings of the Asian Conference on Computer Vision. pp. 858–876 (2022)
39. Jeanneret, G., Simon, L., Jurie, F.: Adversarial counterfactual visual explanations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16425–16435 (2023)

40. Jiménez-Sánchez, A., Juodelyte, D., Chamberlain, B., Cheplygina, V.: Detecting shortcuts in medical images – a case study in chest x-rays. In: International Symposium on Biomedical Imaging (ISBI) (2023)
41. Joshi, S., Koyejo, O., Kim, B., Ghosh, J.: xGEMs: Generating examplars to explain black-box models. arXiv preprint arXiv:1806.08867 (2018)
42. Karunratanakul, K., Preechakul, K., Suwajanakorn, S., Tang, S.: Guided motion diffusion for controllable human motion synthesis. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2151–2162 (2023)
43. Khorram, S., Fuxin, L.: Cycle-consistent counterfactuals by latent transformations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10203–10212 (2022)
44. Lang, O., Gandelsman, Y., Yarom, M., Wald, Y., Elidan, G., Hassidim, A., Freeman, W.T., Isola, P., Globerson, A., Irani, M., et al.: Explaining in style: Training a gan to explain a classifier in stylespace. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 693–702 (2021)
45. Laugel, T., Jeyasothy, A., Lesot, M.J., Marsala, C., Detyniecki, M.: Achieving diversity in counterfactual explanations: a review and discussion. In: 2023 ACM Conference on Fairness, Accountability, and Transparency. FAccT '23, ACM (jun 2023)
46. Liu, Z., Mao, H., Wu, C.Y., Feichtenhofer, C., Darrell, T., Xie, S.: A convnet for the 2020s. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11976–11986 (2022)
47. Liu, Z., Luo, P., Wang, X., Tang, X.: Deep learning face attributes in the wild. In: Proceedings of International Conference on Computer Vision (ICCV) (December 2015)
48. Luccioni, A.S., Akiki, C., Mitchell, M., Jernite, Y.: Stable bias: Analyzing societal representations in diffusion models. In: NeurIPS (2023)
49. Lugmayr, A., Danelljan, M., Romero, A., Yu, F., Timofte, R., Van Gool, L.: Repaint: Inpainting using denoising diffusion probabilistic models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11461–11471 (2022)
50. Mertes, S., Huber, T., Weitz, K., Heimerl, A., André, E.: GANterfactual—counterfactual explanations for medical non-experts using generative adversarial learning. Frontiers in Artificial Intelligence **5** (Apr 2022)
51. Müller, N.M., Roschmann, S., Khan, S., Sperl, P., Böttinger, K.: Shortcut detection with variational autoencoders. In: ICML Workshop on Spurious Correlations,Invariance, and Stability (2023)
52. Nauta, M., Walsh, R., Dubowski, A., Seifert, C.: Uncovering and correcting shortcut learning in machine learning models for skin cancer diagnosis. Diagnostics **12**(1), 40 (Dec 2021)
53. Nemirovsky, D., Thiebaut, N., Xu, Y., Gupta, A.: Countergan: Generating counterfactuals for real-time recourse and interpretability using residual gans. In: Uncertainty in Artificial Intelligence. pp. 1488–1497. PMLR (2022)
54. Neuhaus, Y., Augustin, M., Boreiko, V., Hein, M.: Spurious features everywhere - large-scale detection of harmful spurious features in imagenet. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 20235–20246 (October 2023)
55. Oakden-Rayner, L., Dunnmon, J., Carneiro, G., Re, C.: Hidden stratification causes clinically meaningful failures in machine learning for medical imaging. In: Proceedings of the ACM Conference on Health, Inference, and Learning. ACM (2020)

56. Pahde, F., Dreyer, M., Samek, W., Lapuschkin, S.: Reveal to revise: An explainable AI life cycle for iterative bias correction of deep models. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2023. pp. 596–606. Springer Nature Switzerland (2023)
57. Pakzad, A., Abhishek, K., Hamarneh, G.: CIRCLe: Color invariant representation learning for unbiased classification of skin lesions. In: Lecture Notes in Computer Science. pp. 203–219. Springer Nature Switzerland (2023)
58. Pawelczyk, M., Broelemann, K., Kasneci, G.: On counterfactual explanations under predictive multiplicity. In: Peters, J., Sontag, D. (eds.) Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence (UAI). Proceedings of Machine Learning Research, vol. 124, pp. 809–818. PMLR (03–06 Aug 2020)
59. Pombo, G., Gray, R., Cardoso, M.J., Ourselin, S., Rees, G., Ashburner, J., Nachev, P.: Equitable modelling of brain imaging by counterfactual augmentation with morphologically constrained 3D deep generative models. Medical Image Analysis **84**, 102723 (feb 2023)
60. Qiang, Y., Li, C., Brocanelli, M., Zhu, D.: Counterfactual interpolation augmentation (CIA): A unified approach to enhance fairness and explainability of DNN. In: Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence. IJCAI-2022, International Joint Conferences on Artificial Intelligence Organization (Jul 2022)
61. Raphan, M., Simoncelli, E.P.: Least squares estimation without priors or supervision. Neural computation **23**(2), 374–420 (2011)
62. Rieger, L., Singh, C., Murdoch, W., Yu, B.: Interpretations are useful: penalizing explanations to align neural networks with prior knowledge. In: International conference on machine learning. pp. 8116–8126. PMLR (2020)
63. Rodriguez, P., Caccia, M., Lacoste, A., Zamparo, L., Laradji, I., Charlin, L., Vazquez, D.: Beyond trivial counterfactual explanations with diverse valuable explanations. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV). IEEE (Oct 2021)
64. Rodríguez, P., Caccia, M., Lacoste, A., Zamparo, L., Laradji, I., Charlin, L., Vazquez, D.: Beyond trivial counterfactual explanations with diverse valuable explanations. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1056–1065 (2021)
65. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18. pp. 234–241. Springer (2015)
66. Sanchez, P., Tsaftaris, S.A.: Diffusion causal models for counterfactual estimation. In: First Conference on Causal Learning and Reasoning (2022)
67. Santurkar, S., Ilyas, A., Tsipras, D., Engstrom, L., Tran, B., Madry, A.: Image synthesis with a single (robust) classifier. In: Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., Garnett, R. (eds.) Advances in Neural Information Processing Systems. vol. 32. Curran Associates, Inc. (2019)
68. Singla, S., Eslami, M., Pollack, B., Wallace, S., Batmanghelich, K.: Explaining the black-box smoothly—a counterfactual approach. Medical Image Analysis **84**, 102721 (2023)
69. Singla, S., Pollack, B., Chen, J., Batmanghelich, K.: Explanation by progressive exaggeration. In: International Conference on Learning Representations (2020)
70. Slack, D., Hilgard, A., Lakkaraju, H., Singh, S.: Counterfactual explanations can be manipulated. In: Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., Vaughan,

J.W. (eds.) Advances in Neural Information Processing Systems. vol. 34, pp. 62–75. Curran Associates, Inc. (2021)
71. Smilkov, D., Thorat, N., Kim, B., Viégas, F., Wattenberg, M.: Smoothgrad: removing noise by adding noise. arXiv preprint arXiv:1706.03825 (2017)
72. Sun, S., Koch, L.M., Baumgartner, C.F.: Right for the wrong reason: Can interpretable ML techniques detect spurious correlations? In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2023. pp. 425–434. Springer Nature Switzerland (2023)
73. Thiagarajan, J.J., Thopalli, K., Rajan, D., Turaga, P.: Training calibration-based counterfactual explainers for deep learning models in medical image analysis. Scientific Reports **12**(1) (Jan 2022)
74. Tschandl, P., Rosendahl, C., Kittler, H.: The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. Scientific Data **5**(1) (Aug 2018)
75. Vaeth, P., Fruehwald, A.M., Paassen, B., Gregorova, M.: Diffusion-based visual counterfactual explanations–towards systematic quantitative evaluation. arXiv preprint arXiv:2308.06100 (2023)
76. Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., Summers, R.M.: ChestX-Ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE (Jul 2017)
77. Winkler, J.K., Fink, C., Toberer, F., Enk, A., Deinlein, T., Hofmann-Wellenhof, R., Thomas, L., Lallas, A., Blum, A., Stolz, W., Haenssle, H.A.: Association between surgical skin markings in dermoscopic images and diagnostic performance of a deep learning convolutional neural network for melanoma recognition. JAMA Dermatology **155**(10), 1135 (2019)
78. Zech, J.R., Badgeley, M.A., Liu, M., Costa, A.B., Titano, J.J., Oermann, E.K.: Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. PLOS Medicine **15**(11), e1002683 (nov 2018)
79. Zhang, R., Griner, D., Garrett, J.W., Qi, Z., Chen, G.H.: Training certified detectives to track down the intrinsic shortcuts in COVID-19 chest x-ray data sets. Scientific Reports **13**(1) (Aug 2023)