Controllable Counterfactual Generation for Interpretable Medical Image Classification

Shiyu Liu[1], Fan Wang[2,3,4,*], Zehua Ren[2], Chunfeng Lian[1,3,*], and Jianhua Ma[2,3,*]

[1] School of Mathematics and Statistics, Xi'an Jiaotong University, China
chunfeng.lian@xjtu.edu.cn
[2] The Key Laboratory of Biomedical Information Engineering of Ministry of Education, School of Life Science and Technology, Xi'an Jiaotong University, Xi'an 710049, China
{fan.wang,jhma}@xjtu.edu.cn
[3] Pazhou Lab (Huangpu), Guangzhou 510000, China
[4] The First Affiliated Hospital of Xi'an Jiaotong University, Xi'an 710049, China

Abstract. Counterfactual generation is used to solve the problem of lack of interpretability and insufficient data in deep diagnostic models. By synthesize counterfactual images based on an image-to-image generation model trained with unpaired data, we can interpret the output of a classification model according to a hypothetical class and enhance the training dataset. Recent counterfactual generation approaches based on autoencoders or generative adversarial models are difficult to train or produce realistic images due to the trade-off between image similarity and class difference. In this paper, we propose a new counterfactual generation method based on diffusion models. Our method combines the class-condition control from classifier-free guidance and the reference-image control with attention injection to transform the input images with unknown labels into a hypothesis class. Our methods can flexibly adjust the generation trade-off in the inference stage instead of the training stage, providing controllable visual explanations consistent with medical knowledge for clinicians. We demonstrate the effectiveness of our method on the ADNI structural MRI dataset for Alzheimer's disease diagnosis and conditional 3D image2image generation tasks. Our codes can be found at https://github.com/ladderlab-xjtu/ControlCG.

Keywords: Counterfactual Reasoning, Diffusion Model, Data Augmentation, Attention Injection.

1 Introduction

Deep learning techniques have shown promising performance in medical diagnostics but still face challenges like poor interpretability [3] and limited generalizability due to small training datasets [2]. Counterfactual generation [6, 11] translates the input image into the image with a hypothetical class label. The generated images can be used as data augmentation for model training and the displayed differences can be used as saliency maps for detailed interpretation of the model outputs. In this paper, we present a novel counterfactual generation method based on diffusion models (DMs) [7] with classifier-

free guidance [8] and attention injection [16]. As the schematic diagram shown in Figure 1, our method trains a general model that can flexibly transforms images between N different categories, rather than an expert model that is fixed to transform images between only two specified categories [13, 15].

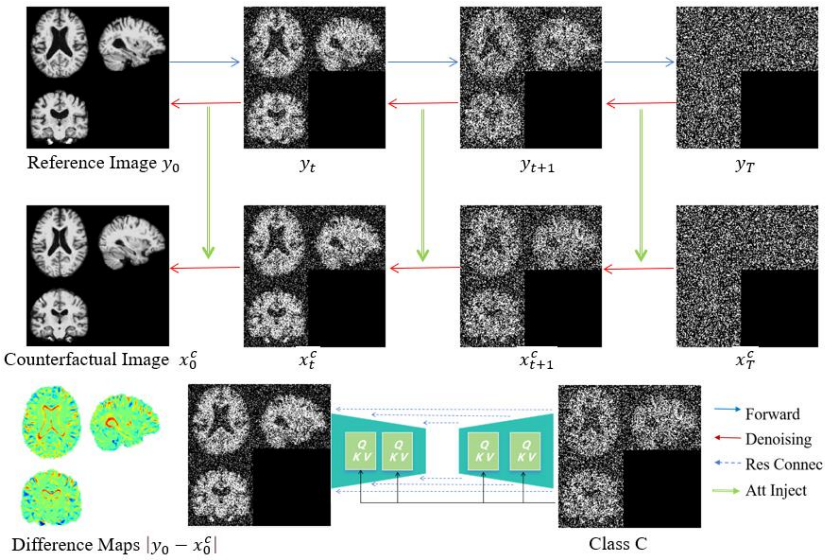

Fig. 1. The overview of our controllable counterfactual generation. Given input image Y_0 and target class C, the diffusion model can generate counterfactual image X^c of Y.

Our approach involves three key phases. Initially, we separately develop a baseline classifier and a conditional diffusion model, where the classifier assesses generation quality, including image similarity and class distinction, while the diffusion model creates images from noise based on class conditions. Next, we produce counterfactual images of hypothetical classes using the diffusion model, employing attention injection during inference to shift from the original image to a target class, maintaining voxel-wise similarity but altering class identity. Finally, these generated images serve as novel data augmentation sources, enabling us to refine the baseline classifier. We evaluate our method on a challenging image-to-image task, i.e., Alzheimer's Disease (AD) versus Cognitively Normal (CN) differentiation within the Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset [10]. This challenging task involves 1) global structural changes like in ventricle volume, 2) local structural changes in small regions like hippocampus, and 3) textural changes between gray and white matter. We benchmark our results against other 3D image-to-image techniques.

2 Related Work

Counterfactual generation is used to solve the problem of lack of interpretability and insufficient data in deep learning-based early disease diagnoses by generating a

modified image as close as to the input image but different in class. There naturally exists a trade-off between image similarity and class difference since the more similar in images, the less different in class. More importantly, reliably capturing such small image differences associating with label changes (e.g., early disease) is itself of critical value in early diagnosis and intervention.

Some works based on autoencoders and latent shift try to partially solve the trade-off [4]. An image is encoded to latent variables, shifted to hypothetical class with the gradient of a classifier, and further decoded to images, offering the steepest path to given hypothetical class to improve image similarity. However, this may result in unrealistic images since the smoothness of latent space is not guaranteed and the gradient of classifier may lead to out-of-distribution direction. Our method can control the trade-off without losing the realism quality of our base noise-tp-image model.

Other works based on Generative Adversarial Networks (GANs) adopt regularization tricks like cycle-consistent [17] and sparsity [11] to balance the trade-off among image realism, image similarity, and class difference. However, GANs are hard to train [1] and regularization terms require hyperparameter tuning in the training stage. In our method, the trade-off can be controlled flexibly during inference.

Recently, DDPMs [7] have gained attention due to their higher image quality than GANs in generation [5]. They are also widely used for image-to-image translation, especially in medical area. Some works [15] adopt classifier [7] or classifier-free[8] guidance for class conditioning and DDIM inversion [14], transforming the brain tumor to brain tissue and calculate difference maps for anomaly detection. In anomaly detection, they care more about difference maps than generated images and they fail in generate realistic images, especially in ventricle, as has been reported in existing studies [15]. Therefore, these existing diffusion-based methods may fail in our task of learning interpretable deep diagnostic models, as here we focus on both saliency maps for detailed explanations and realistically generated images for data augmentation.

3 Method

Counterfactual reasoning [6, 11] aims to give visual explanations through the subject's image transitions from one class to another. The difference between the original image and the transformed image can be used as the saliency map to explain the prediction decision of the classifier. The saliency maps are expected to be related to the prior knowledge of the medical expert. In another word, the difference should only appear in the disease-related regions and not affect the disease-irrelevant individual characteristics. To do so, we adopt a base diffusion model for high quality generation [7], classifier-free guidance [8] for controllable generation, and attention injection [16] for controllable translation. We control the generation results in the inference stage by tuning the guidance scale instead of hyperparameter tuning for the weights of the regularization terms in the loss function during the training stage.

Diffusion Models. Diffusion process [7] is a kind of Markov chains that gradually adds random noise $\varepsilon_t \sim N(0, I)$ to ground-truth image $x_0 \sim p(x_0)$ and get x_T in T timesteps.

At each step t from 0 to N, we get the noisy image from previous timestep with the preset diffusion scheduler α_t in the forward process:

$$x_t = \sqrt{\alpha_t} x_{t-1} + \sqrt{1-\alpha_t}\varepsilon_t, t = 1,2,\dots,T, \quad (1)$$

and get the noisy image from image x_0 by making $\varepsilon \sim N(0, I)$ and $\bar{\alpha}_t = \prod_{t=1}^{T} \alpha_t$:

$$x_t = \sqrt{\bar{\alpha}_t} x_0 + \sqrt{1-\bar{\alpha}_t}\varepsilon, t = 1,2,\dots,T. \quad (2)$$

A diffusion model $\varepsilon_\theta(x_t, t)$ is trained to estimate $\varepsilon_t = \varepsilon_\theta(x_t, t)$ from x_t, which can be used to generate x_0 from randomly sampled noise $x_T \sim N(0, I)$ in backward process:

$$x_{t-1} = \sqrt{\frac{\alpha_t - 1}{\alpha_t}} x_t + \left(\sqrt{\frac{1}{\alpha_{t-1}} - 1} - \sqrt{\frac{1}{\alpha_t} - 1} \right) \varepsilon_\theta(x_t, t), t = 1,2,\dots T. \quad (3)$$

Classifier-Free Guidance. Classifier-free guidance [8] is designed to get the trade-off between realism and diversity in conditional image generation with diffusion models. The joint-trained diffusion model can estimate ε_t with or without class condition c by randomly making $c = \emptyset$ with the probability p_{uncond}:

$$x_{t-1} = \sqrt{\frac{\alpha_t - 1}{\alpha_t}} x_t + \left(\sqrt{\frac{1}{\alpha_{t-1}} - 1} - \sqrt{\frac{1}{\alpha_t} - 1} \right) \varepsilon_\theta(x_t, c, t, w), t = 1,2,\dots T. \quad (4)$$

Then the classifier-free guidance scale w is applied to reach the trade-off by mixing the conditional and unconditional ε_t and properly tuning w in the inference time:

$$\varepsilon_\theta(x_t, c, t, w) = (1+w)\varepsilon_\theta(x_t, c, t) - w\varepsilon_\theta(x_t, \emptyset, t). \quad (5)$$

The greater w indicates better sample quality and worse diversity, vice versa.

Attention Injection. Attention injection [16] is applied for class-conditional image-to-image translation task with diffusion model. A conditional U-Net diffusion model with paired self-attention and cross-attention blocks can generate x_0^c from $x_T \sim N(0, I)$ and the class condition c [12]. The diffusion model uses self-attention blocks to get image-related and class-independent information, and use cross-attention blocks to fuse class-related information. With such model, attention injection allows the translation from a reference image y_0 with unknown class to a similar image x_0^c. In the timestep of t, we have y_t from y_0 with Eq. (2) and the l_{th} cross-attention layer uses the latent vector $h = h^l(y_t, t)$ from previous self-attention layer as the query and the class c as the key and value to get $h_{t,c} = h^{l+1}(y_t, t, c)$ as outputs, such as:

$$h^l(y_t, t) = SelfAttention\big(Q, K, V = h(y_t, t)\big), \quad (6)$$

$$h^{l+1}(y_t, t, c) = CrossAttention(Q = h^l(y_t, t), K, V = c). \quad (7)$$

Here $h(y_t, t)$ contains condition-independent information of noisy image y_t as the middle results for self-attention blocks for image translation from reference image y_0:

$$h^l(x_t, t, y_t) = SelfAttention(Q = h(x_t, t), K, V = (h(y_t, t) \oplus h(x_t, t))), \quad (8)$$

where $\oplus$ indicates channel concatenation. We simplify this process as:

$$\varepsilon_\theta(x_t, c, t | y_t) = \varepsilon_\theta(x_t, c, t) \otimes \varepsilon_\theta(y_t, c, t), \quad (9)$$

where $\otimes$ denotes attention injection. The diffusion model can generate image with class condition c and reference image condition y_0 by Eqs. (2), (4), (5), and (9).

4 Experiments

Dataset and Preprocessing. We evaluate our method on the 3D brain Magnetic Resonance (MR) images from the ADNI dataset. After removing subjects appearing in both ADNI-1 and ADNI-2 from ADNI-2, there are 200 AD and 231 CN subjects in ADNI-1, and 159 AD and 205 CN in ADNI-2. We randomly choose the first-visit images from the 80% CN and AD subjects from ADNI1 for training and 20% for validation. We choose CN and AD subjects form ADNI2 for testing. The images are skull-stripped, registered to MNI space, center-cropped to $160\times192\times160$ and min-max normalized to values with 99.5 percentile as max. The classification model is trained with full-size images and the diffusion model is trained with down-sampled images of $80\times96\times80$.

Baseline Model Training. We train a baseline classifier based on a 3D-DenseNet [9] and a baseline conditional diffusion model for noise-to-image generation. The classifier is trained with SGD optimizer with a learning rate of 3×10^{-2} and a batch size of 4. The diffusion model is trained with Adam optimizer with a learning rate of 10^{-4} and a batch size of 1. It consists of 3 down blocks and 3 up blocks with relevant the number of channels of 64,64, and 128. The 3rd down block and 1st up block consists of a self-attention layer and a cross-attention layer for class conditional generation. We use v-prediction and pyramid noise for better performance. The whole training takes about 20 hours on an NVIDIA RTX 4090 GPU. We use PyTorch 2.0.1 and MONAI 1.2 as software framework.

Saliency Map Generation. We get the absolute difference maps between input images and respectively generated images as saliency maps. Such saliency maps are threshoded according to individual brain masks to focus on foreground brain regions.

Data Augmentation. We evaluate the quality of generated images with data augmentation as the downstream task. We compare performance of classifier with generated images as extra training data among all methods. For a fair comparison, we down-sampled and up-sampled original training dataset and up-sampled extra training dataset to keep the resolution and blur level the same for all methods. We use log uniform to search for the best learning rate and hyperparameters for each method separately.

Competing Methods and Evaluation Metrics. To evaluate our method, we compare itto the **ILVR** [3] and **DDIM inversion** [13-15]. For qualitative analysis, we show, in Fig. 3, one randomly picked subject of AD, and apply all comparing methods to generate the CN/AD image. For quantitative analysis, we evaluate from three aspects: the image similarity at the per-subject level, class difference and image realism both at the distribution level.

We *first* evaluate image similarity between the paired reference image and generated image with **SSIM** and **MSE**. Higher SSIM and lower MSE indicate better performance in image similarity. *Then*, we evaluate class difference with **FID** between datasets of different labels, where rCN, rAD, gCN and gAD indicate the real CN, real AD, generated CN and generated AD, respectively. A method with higher FID in rCN vs. gAD, and rAD vs. gCN, and lower FID in rCN vs. gCN and rAD vs. gAD indicates better performance in class difference. *Finally*, we evaluate realism between datasets of all

real images and all generated images with FID. We use our pre-trained diagnostic model as feature extractor for calculating all the FID. In Table 2, ILVR-1 indicates setting with n=2, T=250, ILVR-2 for n=2, T=500, and ILVR-3 for n=4, T=250.

As an ablation study and a demonstration of controllable generation, we show the generated mages for different classifier-free guidance on a same AD subject in Fig. 2. The image similarity, class difference and image realism for the total generated dataset are also evaluated in Table 3 in the same way. Furthermore, we adopt data augmentation as a downstream task for evaluating the total generated dataset. We compare the classification results trained with extra data generated by different methods.

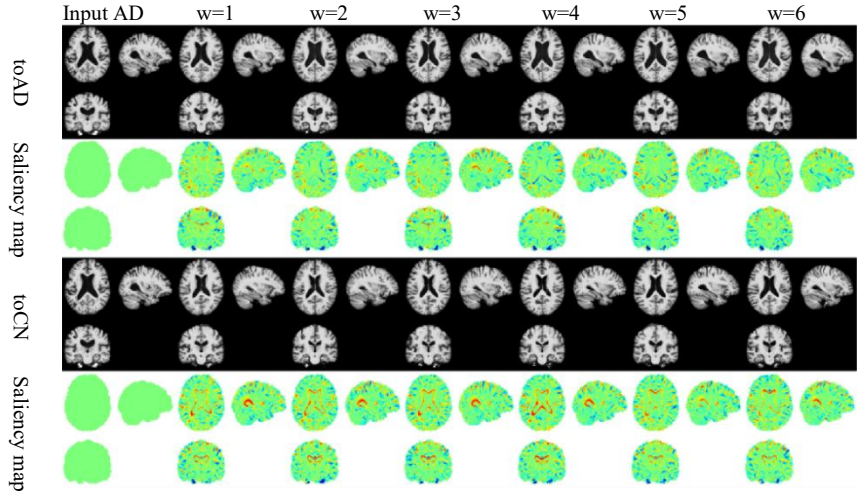

Fig. 2. Controllable generations for our method with different guidance scale w. The greater w indicates better image difference and worse similarity. The left-to-right changes indicate different maps can be manually controlled by guidance scale w. The images are shown in range of 0-1 and saliency maps are shown in range of -1~1.

Table 1. Controllable generation quality when given different hyperparameters for inference. **Bold** numbers indicate the best, underlined 2nd and *italic* 3rd.

w	SSIM↑	MSE↓	rCN2gAD↑	rAD2gAD↓	rCN2gCN↓	rAD2gCN↑	FID↓
1.0	0.823	0.010	**163.605**	28.782	**15.014**	**55.240**	19.739
2.0	0.827	0.010	153.562	25.802	16.509	52.309	18.304
3.0	0.831	0.010	*141.943*	*23.014*	*17.957*	*49.345*	17.052
4.0	*0.835*	0.010	126.626	19.688	18.466	46.430	*15.843*
5.0	0.838	**0.009**	99.339	**18.602**	20.315	42.839	13.504
6.0	**0.841**	**0.009**	58.034	27.735	19.586	43.190	**11.331**

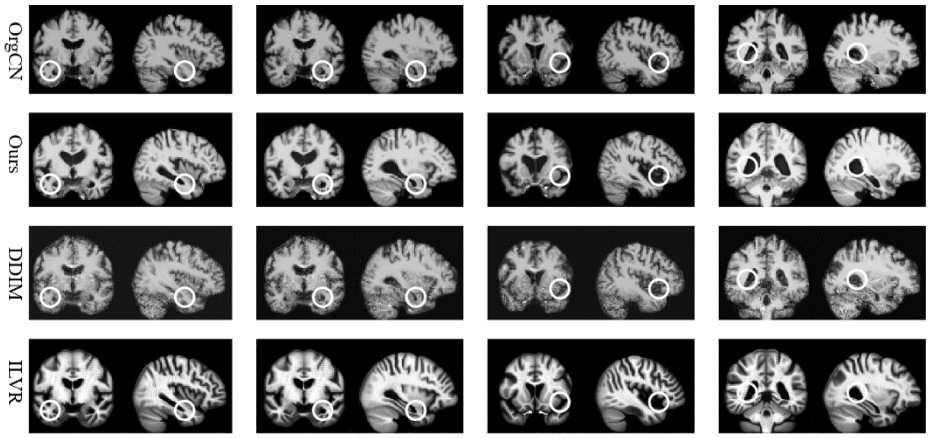

Fig. 3. Counterfactual generation results for a randomly picked CN subject with different method. Our results are consistent with anatomical knowledge in the enlarged temporal poles of the lateral ventricles, the enlargement of the extracerebral space in the anterior Sylvian sulcus and the enlargement of the posterior region of the right lateral ventricle.

Table 2. Comparison among different methods for image similarity, class difference and realism. Methods that fail in any metric are marked with ~~a deletion line in red~~. Our method did not fail in any metric, demonstrating more consistent performance across all measures.

Method	SSIM↑	MSE↓	rCN2gAD↑	rAD2gAD↓	rCN2gCN↓	rAD2gCN↑	FID↓
ILVR-1	~~0.821~~	~~0.014~~	151.521	46.296	31.927	158.631	36.323
ILVR-2	~~0.824~~	~~0.015~~	158.276	65.352	47.995	187.055	~~52.622~~
ILVR-3	0.914	0.004	~~46.162~~	~~46.457~~	50.646	55.748	26.522
DDIM250	0.951	0.003	103.977	4.216	~~42.894~~	~~26.011~~	18.177
DDIM500	0.965	0.002	48.076	20.288	~~61.500~~	~~13.453~~	11.593
Ours	0.835	0.010	126.626	19.688	18.466	46.430	15.843

Table 3. Classification comparison with extra data from different methods. We add FID metric to show roughly positive correlation between image quality and accuracy improvement.

Method	ACC↑	AUC↑	FID↓
Baseline	78.74	~~91.69~~	/
ILVR	77.93	92.90	26.522
DDIM	*82.83*	*93.65*	*18.177*
Ours (w = 1)	76.57	92.86	19.739
Ours (w = 3)	84.47	93.83	17.052
Ours (w = 5)	**85.83**	**93.87**	**13.504**

5 Results and Discussions

Controllable Generations. As can be seen from Fig 2, we observe that with the increase of classifier-free guidance scale w, the class label difference becomes larger and the image similarity becomes smaller, so the trade-off between large class label difference and large image similarity can be controlled by the guidance scale. This manually controlled generation can **1)** enable interpretable diagnosis by medical experts, especially for progressive diseases like AD, **2)** occur at inference time without additional training or fine-tuning, and **3)** produce realistic images that retain disease-irrelevant individual features and modify disease-related features.

Comparison of Generated Results. From Fig. 3, we can observe that ILVR [3] generates unrealistic images for blur in the coronal view. We can also observe that DDIM inversion [15] specializes in altering texture, such as differentiating between white and gray matter, but falls short in adjusting structural features like ventricle size. From the results in Table 2, we can see that our method achieves the best trade-off of class label dissimilarity, image similarity and image realism. Other methods perform badly on at least one metrics. For ILVR [3], image similarity of the generated AD is very poor as FID of rAD2gAD is way too large. For DDIM inversion [15], the generated CN are closer to AD than CN, as FID of rCN2gCN is larger than that of rAD2gCN.

Data Augmentation Performance. Table 3 shows the classification performance with extra data generated by different generation methods. We observe an increase in accuracy when implementing our method comparing with other method. For ablation study, we note that improper selection of hyperparameters decreases the classification accuracy. We also notice that the classification accuracy of different methods and their generation quality are roughly positively correlated.

Overall Evaluation. We evaluate our generated images qualitatively and quantitatively for visual explanations and evaluate our method on downstream data augmentation task. We observe qualitatively, from Figure 3, a large difference between generated AD and original CN, containing 1) structural changes in small but important regions like the atrophy of the hippocampus, 2) structural changes in large regions like the enlargement of the ventricles and 3) texture transformation between gray matter and white matter. Such findings match the medical prior knowledge. As a contrast, DDIM inversion fails in generating structural changes and ILVR fails in getting the trade-off between realism and class difference. We also observe quantitatively from Table 2, our model achieves best trade-off among image realism, image similarity and class difference. Data augmentation can support the conclusion as a downstream task in Table 3.

6 Conclusions

In this paper, we use conditional diffusion models for counterfactual reasoning when given hypothetical class label. Our counterfactual generation quality outperforms state-of-the-art works in reaching the flexible trade-off between image similarity, class difference and image realism. We only modify the inference formula instead of tuning the

regularization weights. Our saliency maps can be tuned in inference time to explain classifier prediction in progressive way instead of training time. This provides medical experts with new tools for interpretable disease diagnosis, especially for early diagnosis of progressive diseases like AD. We demonstrate the high quality of our generated images through a downstream data augmentation task.

Acknowledgments. This work was supported in part by NSFC Grants (Nos. 12326616, 62101431, & 62101430).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

Reference

1. Arjovsky, M., Chintala, S., Bottou, L.: Wasserstein generative adversarial networks. In: International conference on machine learning, pp. 214-223. PMLR, (2017)
2. Chlap, P., Min, H., Vandenberg, N., Dowling, J., Holloway, L., Haworth, A.: A review of medical image data augmentation techniques for deep learning applications. Journal of Medical Imaging and Radiation Oncology 65, 545-563 (2021)
3. Choi, J., Kim, S., Jeong, Y., Gwon, Y., Yoon, S.: Ilvr: Conditioning method for denoising diffusion probabilistic models. arXiv preprint arXiv:2108.02938 (2021)
4. 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, pp. 74-104. PMLR, (2021)
5. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. Advances in neural information processing systems 34, 8780-8794 (2021)
6. Goyal, Y., Wu, Z., Ernst, J., Batra, D., Parikh, D., Lee, S.: Counterfactual visual explanations. In: International Conference on Machine Learning, pp. 2376-2384. PMLR, (2019)
7. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems 33, 6840-6851 (2020)
8. Ho, J., Salimans, T.: Classifier-free diffusion guidance. arXiv preprint arXiv:2207.12598 (2022)
9. 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)
10. Mueller, S.G., Weiner, M.W., Thal, L.J., Petersen, R.C., Jack, C., Jagust, W., Trojanowski, J.Q., Toga, A.W., Beckett, L.: The Alzheimer's disease neuroimaging initiative. Neuroimaging Clinics 15, 869-877 (2005)
11. Oh, K., Yoon, J.S., Suk, H.-I.: Learn-explain-reinforce: counterfactual reasoning and its guidance to reinforce an Alzheimer's Disease diagnosis model. IEEE Transactions on Pattern Analysis and Machine Intelligence 45, 4843-4857 (2022)
12. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 10684-10695. (2022)

13. Sanchez, P., Kascenas, A., Liu, X., O'Neil, A.Q., Tsaftaris, S.A.: What is healthy? generative counterfactual diffusion for lesion localization. In: MICCAI Workshop on Deep Generative Models, pp. 34-44. Springer, (2022)
14. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502 (2020)
15. Wolleb, J., Bieder, F., Sandkühler, R., Cattin, P.C.: Diffusion models for medical anomaly detection. In: International Conference on Medical image computing and computer-assisted intervention, pp. 35-45. Springer, (2022)
16. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 3836-3847. (2023)
17. Zhu, J.-Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE international conference on computer vision, pp. 2223-2232. (2017)