bioengineering

MDPI

Article

Med-cDiff: Conditional Medical Image Generation With Diffusion Models

Alex Ling Yu Hung [1,2,*], Kai Zhao [2], Haoxin Zheng [1,2], Ran Yan [2,3], Steven S. Raman [2], Demetri Terzopoulos [1,4] and Kyunghyun Sung [2]

[1] Computer Science Department, University of California, Los Angeles, CA 90095, USA; haoxinzheng@g.ucla.edu (H.Z.); dt@cs.ucla.edu (D.T.)
[2] Department of Radiology, University of California, Los Angeles, CA 90095, USA; kz@kaizhao.net (K.Z.); ranyan@mednet.ucla.edu (R.Y.); sraman@mednet.ucla.edu (S.S.R.); ksung@mednet.ucla.edu (K.S.)
[3] Bioengineering Department, University of California, Los Angeles, CA 90095, USA
[4] VoxelCloud, Inc., Los Angeles, CA 90024, USA
* Correspondence: alexhung96@ucla.edu

Abstract: Conditional image generation plays a vital role in medical image analysis as it is effective in tasks such as super-resolution, denoising, and inpainting, among others. Diffusion models have been shown to perform at a state-of-the-art level in natural image generation, but they have not been thoroughly studied in medical image generation with specific conditions. Moreover, current medical image generation models have their own problems, limiting their usage in various medical image generation tasks. In this paper, we introduce the use of conditional Denoising Diffusion Probabilistic Models (cDDPMs) for medical image generation, which achieve state-of-the-art performance on several medical image generation tasks.

Keywords: image generation; diffusion models; generative models; super-resolution; denoising; inpainting

Citation: Hung, A.L.Y.; Zhao, K.; Zheng, H.; Yan, R.; Raman, S.S.; Terzopoulos, D.; Sung, K. Med-cDiff: Conditional Medical Image Generation With Diffusion Models. *Bioengineering* **2023**, *10*, 1258. https://doi.org/10.3390/bioengineering10111258

Academic Editors: Alan Wang, Sibusiso Mdletshe, Brady Williamson and Guizhi Xu

Received: 20 September 2023
Revised: 23 October 2023
Accepted: 23 October 2023
Published: 28 October 2023

1. Introduction

Conditional image generation refers to the generation of images using a generative model based on relevant information, which we denote as a condition. When the condition is an image, this is also referred to as image-to-image translation. In the medical domain, this has many important applications such as super-resolution, inpainting, denoising, etc., which can potentially improve healthcare [1]. Super-resolution can help shorten imaging time and improve imaging quality. Denoising helps clinicians and downstream algorithms to make better diagnostic judgments. Medical image inpainting can be beneficial to anomaly detection.

Existing generative models are able to perform some of these jobs decently; e.g., the Hierarchical Probabilistic UNet (HPUNet) [2] for ultrasound image inpainting, and the progressive Generative Adversarial Network (GAN) [3] and SMORE [4] for medical image super-resolution. These methods work to some extent, but they are tailored to specific applications or imaging modalities, making it difficult for researchers to adapt them to different tasks or modalities. MedGAN [5] and UP-GAN [6] target general-purpose medical image generation; however, they are too challenging to train and/or produce underwhelming results.

Models based on Variational Autoencoders (VAE) can be effective in some medical applications [2,7], but the generated images tend to be blurry [8]. Although GAN-based models can generate high-quality medical images [5,9], they suffer from unstable training due to vanishing gradient, convergence, and mode collapse [10]. Normalizing Flow (NF), which has also been used in medical imaging [11,12], can estimate the exact likelihood of the generated sample, making it suitable for certain applications; however, NF requires specifically designed network architectures and the generated image quality fails to impress.

Diffusion models have been dominant in natural image generation due to their ability to generate high-fidelity realistic images [13–16]. They have also been applied to medical image generation [17–20], such as in super-resolution medical imaging [21], but there are only a limited number of studies using conditional diffusion models.

We propose a conditional Denoising Diffusion Probabilistic Model (cDDPM), which we call the medical conditional diffusion model (Med-cDiff), and apply it to a variety of medical image generation tasks, including super-resolution, denoising, and inpainting. In a series of experiments, we show that Med-cDiff achieves state-of-the-art (SOTA) generation performance on these tasks, which demonstrates the great potential of diffusion models in conditional medical image generation.

2. Related Work

Before diffusion models became popular in medical image analysis or in mainstream computer vision, GANs [22] were the most popular image generation methods. Developed to perform conditional natural image generation, Pix2PixGAN [23] was adapted to medical imaging and several researchers have shown its usefulness in such tasks [24–27]. Zhu et al. [28] proposed CycleGAN to perform conditional image-to-image translation between two domains using unpaired images, and the model has also been extensively used in medical imaging. Du et al. [29] made use of CycleGAN in CT image artifact reduction. Yang et al. [30] used a structure-constrained CycleGAN to perform unpaired MRI-to-CT brain image generation. Liu et al. [31] utilized multi-cycle GAN to synthesize CT images from MRI for head-neck radiotherapy. Harms et al. [32] applied CycleGAN to image correction for cone-beam computed tomography (CBCT). Karras et al. [33] proposed StyleGAN, which has an automatically learned, unsupervised separation of high-level attributes and stochastic variation in the generated images, enabling easier control of the image synthesis process. Fetty et al. [34] manipulated the latent space for high-resolution medical image synthesis via StyleGAN. Su et al. [35] performed data augmentation for brain CT motion artifacts detection using StyleGAN. Hong et al. [9] introduced 3D StyleGAN for volumetric medical image generation. Other GAN-based methods have also been proposed for medical imaging. Progressive GAN [3] was used to perform medical image super-resolution. Upadhyay et al. [6] extended the model by utilizing uncertainty estimation to focus more on the uncertain regions during image generation. Armanious et al. [5] proposed MedGAN, specific to medical image domain adaptation, which captured the high and low frequency components of the desired target modality.

Apart from GANs, other generative models, including VAEs and NFs, are also popular in image generation. The VAE was introduced by Kingma and Welling [36], and it has been the basis for a variety of methods for image generation. Vahdat and Kautz [37] developed Nouveau VAE (NVAE), a hierarchical VAE that is able to generate highly realistic images. Hung et al. [2] adapted some of the features from NVAE into their hierarchical conditional VAE for ultrasound image inpainting. Cui et al. [38] adopted NVAE in positron emission tomography (PET) scan image denoising and uncertainty estimation. As for the NF models, Grover et al. [39] proposed AlignFlow based on a similar concept with NF models instead of GANs. Bui et al. [40] extended AlignFlow into medical imaging for Unpaired multi-contrast MRI conditonal image generation. Wang et al. [41] and Beizaee et al. [42] applied NF to medical image harmonization.

In recent years, diffusion models have become the most dominant algorithm in image generation due to their ability to generate realistic images. On natural images, diffusion models have achieved SOTA results in unconditional image generation by outperforming their GAN counterparts [13,14]. Diffusion models have achieved outstanding performance in tasks such as super-resolution [16,43], image editing [44,45], and unpaired conditional image generation [46], and they have attained SOTA performance in conditional image generation [15]. In medical imaging, unsupervised anomaly detection is an important application of unconditional diffusion models [17,47–49]. Image segmentation is a popular application of conditional diffusion models, where the image to be segmented is used as the

condition [19,50–53]. Diffusion models have also been widely applied to accelerating MRI reconstruction [20,54,55]. Özbey et al. [18] used GANs to shorten the denoising process in diffusion models for medical imaging.

3. Methods

3.1. Background

The goal of conditional image generation is to generate the target image x_0 given a correlated conditional image y. Diffusion models consist of two parts: a forward noising process q, and a reverse denoising process p_θ parameterized by θ. Figure 1 illustrates conditional diffusion models. At a high level, given y, they sample from a data distribution during p_θ, reversing q, which adds noise iteratively to the original image x_0. More specifically, the sampling process starts with a random noise sample x_T, and iteratively generates less-noisy samples, $x_{T-1}, x_{T-2}, \ldots$, based on the conditional image y for T steps until reaching the final output sample x_0. For a specific sample x_t during the process, the larger t is, the more noisy the sample will be. Given the conditional image y, the reverse process p_θ learns to denoise the sample x_t by one step to x_{t-1}.

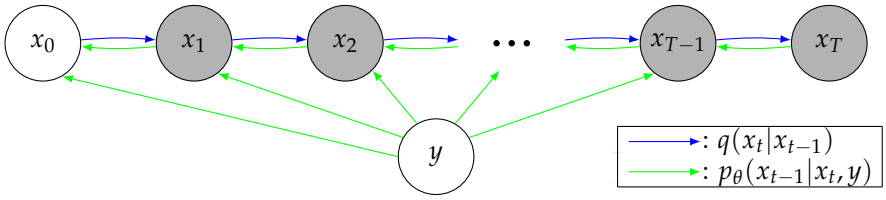

Figure 1. A graphical model representation of conditional diffusion models. The blue and green arrows indicate the forward and reverse processes, respectively.

The forward process q is a Markovian noising process, where Gaussian noise is added to the image x_{t-1} at each time step $t = 1, 2, ..., T$ according to a variance schedule β_t:

$$q(x_t|x_{t-1}) = \mathcal{N}\left(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t I\right), \tag{1}$$

where $\mathcal{N}(\cdot)$ denotes the normal distribution and I is the identity matrix. Note that

$$q(x_1, \ldots, x_T|x_0) = \prod_{t=1}^{T} q(x_t|x_{t-1}), \tag{2}$$

where T is the number of steps. The forward noising process (1) can be used to sample x_t at any timestep t in closed form. In other words, since

$$q(x_t|x_0) = \mathcal{N}\left(x_t; \sqrt{\bar{\alpha}_t}x_0, (1-\bar{\alpha}_t)I\right), \tag{3}$$

then for the original image x_0 and any given timestep t

$$x_t = \sqrt{\bar{\alpha}_t}x_0 + (1-\bar{\alpha}_t)\epsilon, \tag{4}$$

where $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \prod_{i=1}^{t} \alpha_i$, and $\epsilon \sim \mathcal{N}(0, 1)$. When T is large, we can assume that $x_T \sim \mathcal{N}(0, I)$, which is random Gaussian noise containing no information regarding the original image x_0 [13].

In a conditional diffusion model, the objective is to learn the reverse process p_θ so that we can infer x_{t-1} given x_t and the conditional image y. In this way, starting from the Gaussian noise $x_T \sim \mathcal{N}(0, 1)$, and given y, we can iteratively infer the sample at time step $t-1$ from the sample at time step t until we reach the original image x_0. For the reverse process,

$$p_\theta(x_0, \ldots, x_T|y) = p_\theta(x_T) \prod_{t=1}^{T} p_\theta(x_{t-1}|x_t, y). \tag{5}$$

The reverse process can therefore be parameterized as

$$p_\theta(x_{t-1}|x_t, y) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, y, t), \Sigma_\theta(x_t, y, t)), \tag{6}$$

where we set $\Sigma_\theta(x_t, y, t) = \sigma_t^2 I$. As for $\mu_\theta(x_t, y, t)$, Ho et al. [13] showed that it must be parameterized as

$$\mu_\theta(x_t, y, t) = \frac{1}{\sqrt{\alpha_t}}\left(x_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(x_t, y, t)\right), \tag{7}$$

where $\epsilon_\theta(x_t, y, t)$ is a function approximating ϵ.

For a total of T steps, the training objective is to minimize the variational lower bound on the negative log-likelihood:

$$\begin{aligned}\mathbb{E}[-\log p_\theta(x_0|y)] &\leq \mathbb{E}_q\left[-\log \frac{p_\theta(x_0, \ldots, x_T|y)}{q(x_1, \ldots, x_T|x_0)}\right] \\ &= \mathbb{E}_q\left[-\log p_\theta(x_T) - \sum_{t=1}^{T}\log \frac{p_\theta(x_{t-1}|x_t, y)}{q(x_t|x_{t-1})}\right] \\ &= L(\theta).\end{aligned} \tag{8}$$

More efficient training can be achieved by optimizing random terms in the training objective $L(\theta)$ using stochastic gradient descent. Therefore, we can rewrite the training objective as

$$L(\theta) = \mathbb{E}_q\left[\sum_{t=1}^{T} L_t(\theta)\right], \tag{9}$$

where

$$L_t(\theta) = \begin{cases} -\log p_\theta(x_0|x_1) & \text{if } t = 0, \\ D_{KL}(q(x_t|x_{t+1}, x_0)\|p_\theta(x_t|x_{t+1}, y)) & \text{if } 0 < t < T, \\ D_{KL}(q(x_T|x_0)\|p_\theta(x_T)) & \text{if } t = T, \end{cases} \tag{10}$$

and $D_{KL}(.\|.)$ is the Kullback-Leibler (KL) divergence between two distributions. In (10), the term $q(x_t|x_{t+1}, x_0)$ is given by

$$q(x_t|x_{t+1}, x_0) = \mathcal{N}\left(x_t; \tilde{\mu}_{t+1}(x_{t+1}, x_0), \tilde{\beta}_{t+1} I\right), \tag{11}$$

where

$$\tilde{\mu}_{t+1}(x_{t+1}, x_0) = \frac{\sqrt{\bar{\alpha}_t}\beta_{t+1}}{1-\bar{\alpha}_{t+1}}x_0 + \frac{\sqrt{\alpha_{t+1}}(1-\bar{\alpha}_t)}{1-\bar{\alpha}_{t+1}}x_{t+1}, \tag{12}$$

with

$$\tilde{\beta}_{t+1} = \frac{1-\bar{\alpha}_t}{1-\bar{\alpha}_{t+1}}\beta_{t+1}. \tag{13}$$

3.2. Training and Sampling

When $t = T$, $L_T(\theta)$ is a constant with no learnable parameters since β_t is fixed to a constant. Therefore, $L_t(\theta)$ can be ignored during training.

When $0 < t < T$, $L_t(\theta)$ can be expressed as

$$L_t(\theta) = \mathbb{E}_q\left[\frac{\|\tilde{\mu}_{t+1}(x_{t+1}, x_0) - \mu_\theta(x_{t+1}, y, t+1)\|^2}{2\sigma_{t+1}^2}\right] + C \tag{14}$$

$$= \mathbb{E}_{x_0,\epsilon}\left[\frac{\left\|\frac{1}{\sqrt{\bar{\alpha}_{t+1}}}\left(x_{t+1}(x_0,\epsilon) - \frac{\beta_{t+1}}{\sqrt{1-\bar{\alpha}_{t+1}}}\epsilon\right) - \mu_\theta(x_{t+1}(x_0,\epsilon), y, t+1)\right\|^2}{2\sigma_{t+1}^2}\right] + C \tag{15}$$

$$= \mathbb{E}_{x_0,\epsilon}\left[\frac{\beta_{t+1}^2\left\|\epsilon - \epsilon_\theta(\sqrt{\bar{\alpha}_{t+1}}x_0 + \sqrt{1-\bar{\alpha}_{t+1}}\epsilon, y, t+1)\right\|^2}{2\sigma_{t+1}^2\alpha_{t+1}(1-\bar{\alpha}_{t+1})}\right] + C, \tag{16}$$

where C is a constant.

When $t = 0$, assuming all the image data have been re-scaled to $[-1, 1]$, the expression of $L_0(\theta)$ can be written as

$$L_0(\theta) = -\log p_\theta(x_0|x_1) = -\sum_{i=1}^{H}\sum_{j=1}^{W}\int_{f(x_0^{i,j}-\delta)}^{f(x_0^{i,j}+\delta)} \mathcal{N}\left(x; \mu^{i,j}, \theta(x_1, 1), \sigma_1^2\right) dx, \tag{17}$$

where H and W are the height and width of the image, respectively, and δ is a small number, and where

$$f(x) = \begin{cases} 1 & \text{if } x > 1, \\ x & \text{if } -1 < x < 1, \\ -1 & \text{if } x < -1. \end{cases} \tag{18}$$

From Equations (16) and (17), we see that the training objective is differentiable with respect to the model parameter θ. During each training step, we sample the image pair (x_0, y) from the dataset x_0, $y \sim p_{data}(x, y)$, the time step t from a uniform distribution $t \sim \mathcal{U}(\{1, 2, ..., T\})$, and ϵ from a normal distribution $\epsilon \sim \mathcal{N}(0, I)$. We then perform gradient descent on

$$\nabla_\theta\left\|\epsilon - \epsilon_\theta(\sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, y, t)\right\|^2, \tag{19}$$

which is an alternative variational lower bound that has been shown to be better for sampling quality [13].

During sampling, x_T is first sampled from a normal distribution $x_T \sim \mathcal{N}(0, I)$. Then we iteratively sample $x_{T-1}, x_{T-2}, \ldots, x_0$ from distribution $x_{t-1} \sim p_\theta(x_{t-1}|x_t, y)$ by

$$x_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(x_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(x_t, y, t)\right) + \sigma_t z, \tag{20}$$

where σ_t is an untrained time dependent constant and $z \sim \mathcal{N}(0, I)$.

4. Experiments

4.1. Datasets

Our method is evaluated on the following datasets:

1. MRI Super Resolution: The dataset consists of 296 patients who underwent pre-operative prostate MRI prior to robotic-assisted laparoscopic prostatectomy. T2-weighted imaging was used for the experiment, acquired by the Turbo Spin Echo (TSE) MRI sequence following the standardized imaging protocol of the European Society of Urogenital Radiology (ESUR) PI-RADS guidelines [56]. Additionally, the dataset includes annotation of the transition zone (TZ) and peripheral zone (PZ) of the prostate. Overall, 238, 29, and 29 patients were used for training, validation, and testing, respectively. To perform super-resolution, we downsampled the images by a factor of $2\sqrt{2}$, 4, $4\sqrt{2}$, 8, $8\sqrt{2}$, and 16.

2. X-ray Denoising: The public chest X-ray dataset [57] contains 5863 X-ray images with pneumonia and normal patients. Overall, 624 images were used for testing. Pneumonia patients were further categorized as virus- or bacteria-infected patients. We randomly added Gaussian noise as well as salt and pepper noise to the images and used the original images as the ground truth.
3. MRI Inpainting: The dataset consists of 18,813 T1-weighted prostate MRI images that were acquired by the Spoiled Gradient Echo (SPGR) sequence. We used 6271 of them for testing. The masks were randomly generated during training, and they were fixed among different tests for testing.

4.2. Implementation and Evaluation Details

For Med-cDiff, $\epsilon_\theta(x_t, y, t)$ was parameterized by a U-Net [58] while using group normalization [59]. The total number of steps was set to $T = 2000$. The forward process variances were set to constants that linearly increase from $\beta_1 = 10^{-4}$ and $\beta_T = 0.02$. We also set $\sigma_t^2 = \beta_t$. All the images used were resized to 128×128, and the pixel values are normalized to the range $[-1, 1]$ in a patient-wise manner. The models were all trained for 2×10^5 iterations with a learning rate of 1×10^{-4}.

For quantitative evaluation, we used the following metrics: Learned Perceptual Image Patch Similarity (LISPS) (v1.0) [60] with AlexNet [61] as the backbone, Fréchet Inception Distance (FID) [62], accutance (acc) [63], which measures the sharpness of an image, Dice similarity coefficient (DSC) [64], classification accuracy, and the 2-alternative forced-choice (2AFC) paradigm [65].

Due to the domain gaps [66,67] between different datasets and different tasks, combining datasets and training a combined network would yield a worse performance than separately training the networks. Thus, we trained and tested our methods on different tasks separately.

4.3. MRI Super-Resolution

For MRI super-resolution, we downsampled the MRI images by a factor of $2\sqrt{2}$, 4, and $4\sqrt{2}$, and then we upscaled the images to their original size. We compared the performance of Med-cDiff against bilinear interpolation, pix2pixGAN [23], and SRGAN [68] both visually and quantitatively, evaluated by LPIPS, FID, and accutance, as well as performance comparison on the downstream zonal segmentation task.

Figure 2 shows qualitative results. Clearly, images generated by the other methods are blurry and lack realistic textures, whereas Med-cDiff is able to recover the shape of the prostate as well as relevant textures. For zonal segmentation, we utilized the pretrained CAT-nnUNet [69] and calculated the 3D patient-wise DSC for evaluation. The quantitative results are reported in Table 1, confirming that the images generated by Med-cDiff are the most realistic with the best sharpness and are useful in downstream zonal segmentation. Furthermore, to show the effectiveness of Med-cDiff on zonal segmentation, we further downsampled the original images by a factor of 8, $8\sqrt{2}$, and 16 and performed MRI super-resolution. The results on downstream zonal segmentation are plotted in Figure 3, which reveals that Med-cDiff clearly outperforms bilinear interpolation and pix2pixGAN. CAT-nnUNet performs similarly on images generated by Med-cDiff and SRGAN for PZ segmentation, but it performs better on images generated by Med-cDiff for TZ segmentation. The segmentation performance using bilinear interpolation and pix2pixGAN drops drastically as the upscaling factor increases, while the segmentation performance using images generated by SRGAN and Med-cDiff does not decrease much.

Table 1. Numerical comparison of Med-cDiff against other super-resolution methods.

Factor		LPIPS ($\times 10^{-4}$)↓	FID↓	acc.↑	PZ DSC (%)↑	TZ DSC (%)↑
$2\sqrt{2}$	bilinear	2787.847	**1.19**	6.75	**82.8**	87.7
	pix2pixGAN	**1.53**	1.20	12.72	81.5	87.2
	SRGAN	3.30	**1.19**	5.60	82.7	88.0
	Med-cDiff	2.74	**1.19**	**22.84**	81.7	**88.2**
4	bilinear	4339.392	1.20	4.51	78.2	84.2
	pix2pixGAN	**1.96**	1.22	11.31	78.3	86.1
	SRGAN	5.03	**1.19**	5.11	**80.2**	86.2
	Med-cDiff	4.62	**1.19**	**21.44**	77.8	**86.3**
$4\sqrt{2}$	bilinear	5773.238	1.21	3.28	68.9	75.9
	pix2pixGAN	**2.50**	1.22	12.68	69.2	81.1
	SRGAN	6.09	1.21	4.39	72.6	82.7
	Med-cDiff	5.09	**1.20**	**21.37**	**74.2**	**84.3**

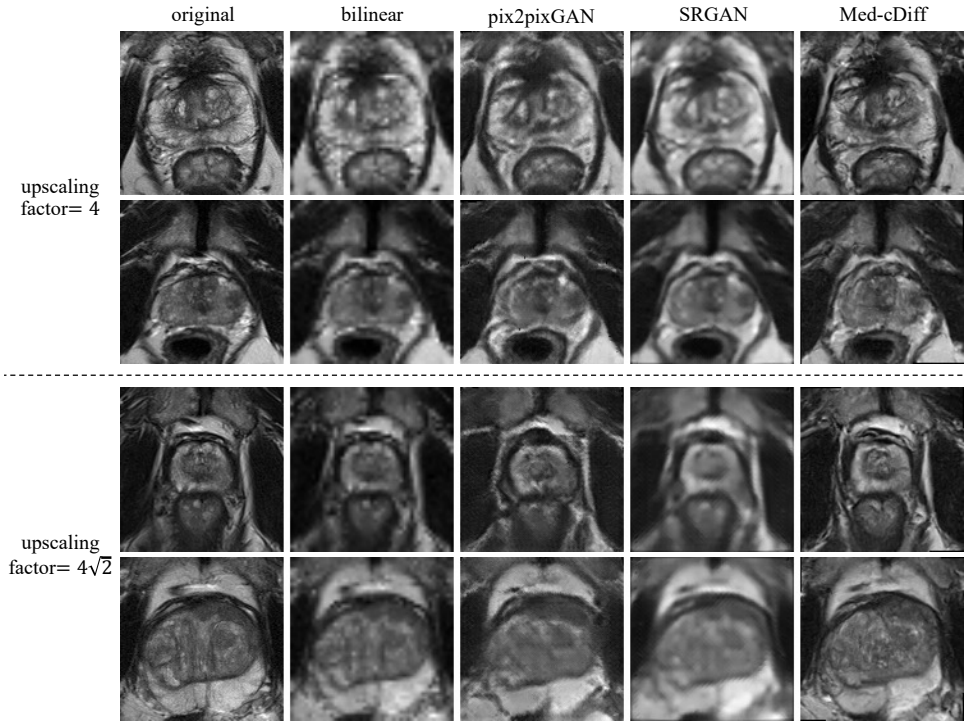

Figure 2. Qualitative comparison of Med-cDiff against other super-resolution methods.

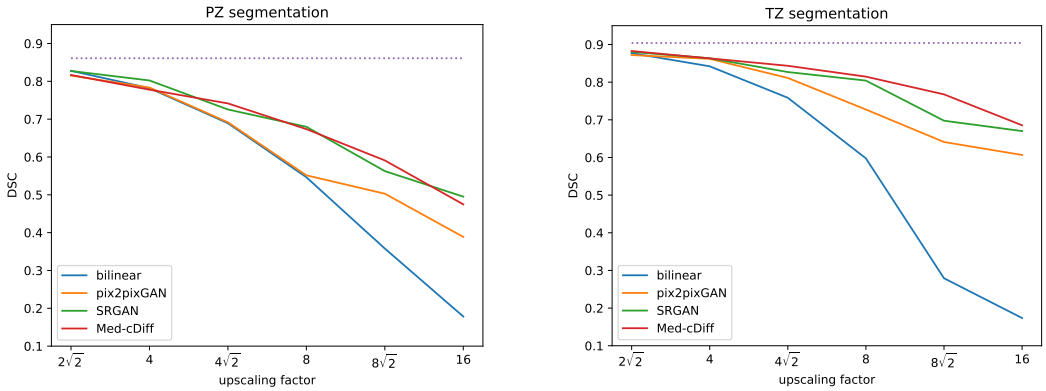

Figure 3. DSC comparison of Med-cDiff against bilinear interpolation, pix2pixGAN, and SRGAN for zonal segmentation. The purple dotted lines indicate scores from the original high-resolution images.

4.4. X-ray Denoising

We evaluated the denoising results using the LPIPS and FID metrics, and further evaluated the results by comparing the downstream classification performance, where 3-class classification (normal/bacterial pneumonia/viral pneumonia) was performed using VGG11 [70]. We compared Med-cDiff against pix2pixGAN [23] and UP-GAN [6].

The quantitative results are reported in Table 2. Med-cDiff outperforms the other methods in every metric. Qualitative results are shown in Figure 4, where we see that pix2pixGAN creates new artifacts and distorts the anatomy while UP-GAN creates unrealistic blurry images lacking details. More specifically, in the normal image example in Figure 4, the yellow arrows point to the newly generated artifacts, and the red arrows point to the unusually large spinal cord. By contrast, Med-cDiff generates realistic patterns in those regions. In the viral pneumonia example, pix2pixGAN cannot generate the bright pattern in the original image at the yellow arrow. As for the bacterial pneumonia example, pix2pixGAN cannot generate the spinal cord with the correct shape at the yellow arrow. In both pneumonia examples, pix2pixGAN failed to recover the correct shape of the ribs at the red arrows.

Table 2. Quantitative comparison of Med-cDiff against other denoising methods.

	LPIPS ($\times 10^{-4}$)↓	**FID↓**	**Classification Accuracy (%)↑**
original image	-	-	70.7
noisy image	17.52	1.35	63.6
pix2pixGAN	1.77	1.32	65.1
UP-GAN	3.36	1.33	62.8
Med-cDiff	**1.19**	**1.30**	**65.8**

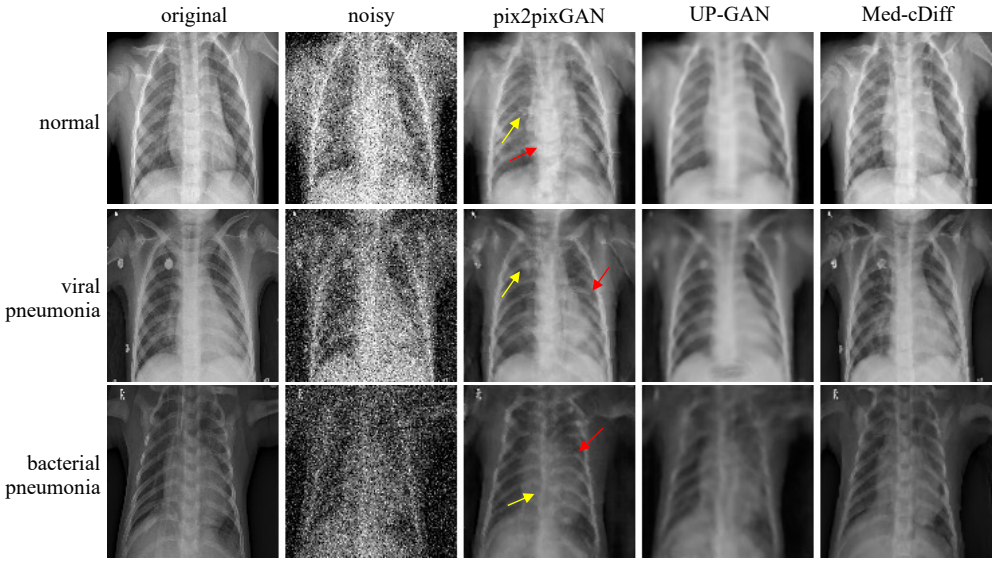

Figure 4. Qualitative comparison of Med-cDiff against other denoising methods. Arrows point to regions that pix2pixGAN cannot correctly generate.

4.5. MRI Inpainting

We compared our method against other inpainting methods such as pix2pixGAN, HPUNet [2], and UP-GAN using the LPIPS and FID metrics. Furthermore, we performed a 2AFC paradigm [65] to measure how well trainees can discriminate real images from the generated ones. We randomly sampled 50 real and generated image pairs from the test set for each method and asked four trainees to perform 2AFC. We averaged the results from the four trainees.

The quantitative results in Table 3 reveal that Med-cDiff can generate the most realistic images. The 2AFC values convey that it is difficult to determine that images generated by Med-cDiff are not real, while it is easy to discern the inauthenticity of images generated by competing methods. The visual results in Figure 5 further confirm that Med-cDiff generates the most authentic images. More specifically, in the masked regions, pix2pixGAN generates unrealistic patterns that are clear indicators of images generated by GANs, while HPUNet can generate somewhat realistic patterns, although the generated patches are still relatively blurry. HPUNet was designed for ultrasound image inpainting, but the performance is unimpressive when applied to MRI images. This shows the difficulties in applying some methods to cross-imaging modalities. As for UP-GAN, the generated patches were blurry, while Med-cDiff generated realistic patterns and contents.

Table 3. Quantitative comparison of Med-cDiff against other inpainting methods.

	LPIPS ($\times 10^{-6}$)↓	**FID↓**	**2AFC Accuracy (%)↓**
pix2pixGAN	7.62	1.010	98.0
HPUNet	5.39	0.995	95.0
UP-GAN	3.17	0.897	94.5
Med-cDiff	**2.96**	**0.582**	**64.0**

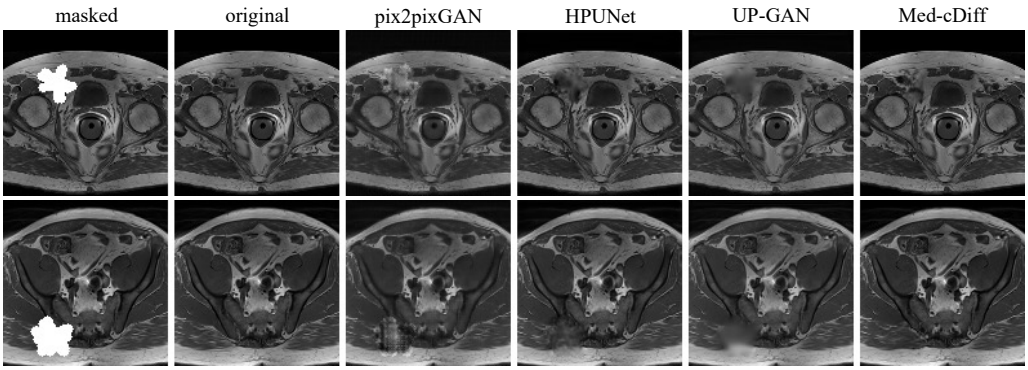

Figure 5. Qualitative comparison of Med-cDiff against other inpainting methods.

5. Conclusions

We have introduced Med-cDiff, a conditional diffusion model for medical image generation, and shown that Med-cDiff is effective in several medical image generation tasks, including MRI super-resolution, X-ray image denoising, and MRI image inpainting. We have demonstrated that Med-cDiff can generate high-fidelity images, both quantitatively and qualitatively superior to those generated by other GAN- and VAE-based methods. The images generated by Med-cDiff were also tested in downstream tasks such as organ segmentation and disease classification, and we showed that these tasks can benefit from the images generated by Med-cDiff.

More importantly, Med-cDiff was not designed for any specific application yet it outperforms models designed for specific applications. For example, SRGAN is specifically designed to generate high-resolution images from low-resolution images as it upsamples the low-resolution images within the network, while HPUNet is mainly used for inpainting ultrasound images to generate realistic ultrasound noise patterns. By contrast, since conditional diffusion models can generate highly realistic images, Med-cDiff can learn to generate various medical images with different characteristics and patterns.

In future work, we will apply Med-cDiff to other downstream tasks; e.g., anomaly detection and faster image reconstruction. Conditional medical image generation is not limited to these tasks. Other applications, such as inter-modality image translation and image enhancement, are also worthy of exploration.

Author Contributions: Conceptualization, A.L.Y.H. and K.S.; Methodology, A.L.Y.H., K.Z. and H.Z.; Software, A.L.Y.H. and K.Z.; Validation, A.L.Y.H., K.Z., H.Z., R.Y., S.S.R. and K.S.; Formal analysis, A.L.Y.H. and K.S.; Investigation, A.L.Y.H. and K.S.; Resources, S.S.R. and K.S.; Data curation, A.L.Y.H. and S.S.R.; Writing—original draft preparation, A.L.Y.H. and K.S.; Writing—review and editing, A.L.Y.H., K.Z., H.Z., R.Y., S.S.R., D.T. and K.S.; Visualization, A.L.Y.H.; Supervision, D.T. and K.S.; Project administration, K.S.; Funding acquisition, K.S. All authors contributed to the article and approved the submitted version.

Funding: This research was funded in part by the National Institutes of Health under grants R01-CA248506 and R01-CA272702, and by the Integrated Diagnostics Program of the Departments of Radiological Sciences and Pathology in the UCLA David Geffen School of Medicine.

Institutional Review Board Statement: The study was conducted according to the guidelines of the Declaration of Helsinki and approved by the Institutional Review Board of UCLA.

Informed Consent Statement: The ethics committee waived the requirement of written informed consent for participation.

Data Availability Statement: The raw data supporting the conclusions of this study will be made available by the authors in accordance with UCLA's institutional management and sharing policy.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study, in the collection, analyses, or interpretation of data, in the writing of the manuscript, or in the decision to publish the results.

References

1. Havaei, M.; Mao, X.; Wang, Y.; Lao, Q. Conditional generation of medical images via disentangled adversarial inference. *Med. Image Anal.* **2021**, *72*, 102106. [CrossRef]
2. Hung, A.L.Y.; Sun, Z.; Chen, W.; Galeotti, J. Hierarchical probabilistic ultrasound image inpainting via variational inference. In Proceedings of the First MICCAI Workshop on Deep Generative Models, and Data Augmentation, Labelling, and Imperfections, Strasbourg, France, 1 October 2021; Springer: Berlin/Heidelberg, Germany, 2021; pp. 83–92.
3. Mahapatra, D.; Bozorgtabar, B.; Garnavi, R. Image super-resolution using progressive generative adversarial networks for medical image analysis. *Comput. Med. Imaging Graph.* **2019**, *71*, 30–39. [CrossRef]
4. Zhao, C.; Dewey, B.E.; Pham, D.L.; Calabresi, P.A.; Reich, D.S.; Prince, J.L. SMORE: A self-supervised anti-aliasing and super-resolution algorithm for MRI using deep learning. *IEEE Trans. Med. Imaging* **2020**, *40*, 805–817.
5. Armanious, K.; Jiang, C.; Fischer, M.; Küstner, T.; Hepp, T.; Nikolaou, K.; Gatidis, S.; Yang, B. MedGAN: Medical image translation using GANs. *Comput. Med. Imaging Graph.* **2020**, *79*, 101684.
6. Upadhyay, U.; Chen, Y.; Hepp, T.; Gatidis, S.; Akata, Z. Uncertainty-guided progressive GANs for medical image translation. In Proceedings of the 24th International Conference on Medical Image Computing and Computer Assisted Intervention, Strasbourg, France, 27 September–1 October 2021; Springer: Berlin/Heidelberg, Germany, 2021; pp. 614–624.
7. Cetin, I.; Stephens, M.; Camara, O.; Ballester, M.A.G. Attri-VAE: Attribute-based interpretable representations of medical images with variational autoencoders. *Comput. Med. Imaging Graph.* **2023**, *104*, 102158.
8. Ehrhardt, J.; Wilms, M. Autoencoders and variational autoencoders in medical image analysis. In *Biomedical Image Synthesis and Simulation*; Elsevier: Amsterdam, The Netherlands 2022; pp. 129–162.
9. Hong, S.; Marinescu, R.; Dalca, A.V.; Bonkhoff, A.K.; Bretzner, M.; Rost, N.S.; Golland, P. 3D-StyleGAN: A style-based generative adversarial network for generative modeling of three-dimensional medical images. In Proceedings of the First MICCAI Workshop on Deep Generative Models, and Data Augmentation, Labelling, and Imperfections, Strasbourg, France, 1 October 2021; Springer: Berlin/Heidelberg, Germany, 2021; pp. 24–34.
10. AlAmir, M.; AlGhamdi, M. The Role of generative adversarial network in medical image analysis: An in-depth survey. *ACM Comput. Surv.* **2022**, *55*, 1–36.
11. van de Schaft, V.; van Sloun, R.J. Ultrasound speckle suppression and denoising using MRI-derived normalizing flow priors. *arXiv* **2021**, arXiv:2112.13110.
12. Hajij, M.; Zamzmi, G.; Paul, R.; Thukar, L. Normalizing flow for synthetic medical images generation. In Proceedings of the IEEE Healthcare Innovations and Point of Care Technologies, Houston, TX, USA, 10–11 March 2022; IEEE: Piscataway, NJ, USA, 2022; pp. 46–49.
13. Ho, J.; Jain, A.; Abbeel, P. Denoising diffusion probabilistic models. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 6840–6851.
14. Dhariwal, P.; Nichol, A. Diffusion models beat GANS on image synthesis. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 8780–8794.
15. Saharia, C.; Chan, W.; Chang, H.; Lee, C.; Ho, J.; Salimans, T.; Fleet, D.; Norouzi, M. Palette: Image-to-image diffusion models. In Proceedings of the ACM SIGGRAPH 2022 Conference, Vancouver, BC, Canada, 8–11 August 2022; Association for Computing Machinery: New York, NY, USA, 2022; pp. 1–10.

16. Saharia, C.; Ho, J.; Chan, W.; Salimans, T.; Fleet, D.J.; Norouzi, M. Image super-resolution via iterative refinement. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *45*, 4713–4726. [CrossRef]
17. Pinaya, W.H.; Tudosiu, P.D.; Dafflon, J.; Da Costa, P.F.; Fernandez, V.; Nachev, P.; Ourselin, S.; Cardoso, M.J. Brain imaging generation with latent diffusion models. In Proceedings of the Second MICCAI Workshop on Deep Generative Models, Singapore, 22 September 2022; Springer: Berlin/Heidelberg, Germany, 2022; pp. 117–126.
18. Özbey, M.; Dalmaz, O.; Dar, S.U.; Bedel, H.A.; Özturk, Ş.; Güngör, A.; Çukur, T. Unsupervised medical image translation with adversarial diffusion models. *IEEE Trans. Med. Imaging* 2023, *early access*. [CrossRef]
19. Wolleb, J.; Sandkühler, R.; Bieder, F.; Valmaggia, P.; Cattin, P.C. Diffusion models for implicit image segmentation ensembles. In Proceedings of the International Conference on Medical Imaging with Deep Learning, Zurich, Switzerland, 6–8 July 2022; pp. 1336–1348.
20. Chung, H.; Ye, J.C. Score-based diffusion models for accelerated MRI. *Med. Image Anal.* **2022**, *80*, 102479. [CrossRef]
21. Chung, H.; Lee, E.S.; Ye, J.C. MR Image Denoising and Super-Resolution Using Regularized Reverse Diffusion. *IEEE Trans. Med. Imaging* **2022**, *42*, 922–934. [CrossRef]
22. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial nets. *Adv. Neural Inf. Process. Syst.* **2014**, 2672–2680.
23. Isola, P.; Zhu, J.Y.; Zhou, T.; Efros, A.A. Image-to-image translation with conditional adversarial networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1125–1134.
24. Tang, Y.B.; Oh, S.; Tang, Y.X.; Xiao, J.; Summers, R.M. CT-realistic data augmentation using generative adversarial network for robust lymph node segmentation. In Proceedings of the Medical Imaging 2019: Computer-Aided Diagnosis, San Diego, CA, USA, 17–20 February 2019; SPIE: Bellingham, WA, USA, 2019; Volume 10950, pp. 976–981.
25. Popescu, D.; Deaconu, M.; Ichim, L.; Stamatescu, G. Retinal blood vessel segmentation using Pix2Pix GAN. In Proceedings of the 29th Mediterranean Conference on Control and Automation, Puglia, Italy, 22–25 June 2021; IEEE: Piscataway, NJ, USA, 2021; pp. 1173–1178.
26. Aljohani, A.; Alharbe, N. Generating synthetic images for healthcare with novel deep Pix2Pix GAN. *Electronics* **2022**, *11*, 3470.
27. Sun, J.; Du, Y.; Li, C.; Wu, T.H.; Yang, B.; Mok, G.S. Pix2Pix generative adversarial network for low dose myocardial perfusion SPECT denoising. *Quant. Imaging Med. Surg.* **2022**, *12*, 3539. [CrossRef]
28. 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, Venice, Italy, 22–29 October 2017; pp. 2223–2232.
29. Du, M.; Liang, K.; Xing, Y. Reduction of metal artefacts in CT with Cycle-GAN. In Proceedings of the IEEE Nuclear Science Symposium and Medical Imaging Conference, Sydney, NSW, Australia, 10–17 November 2018; IEEE: Piscataway, NJ, USA, 2018; pp. 1–3.
30. Yang, H.; Sun, J.; Carass, A.; Zhao, C.; Lee, J.; Xu, Z.; Prince, J. Unpaired brain MR-to-CT synthesis using a structure-constrained CycleGAN. In Proceedings of the International Workshops on Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support, Granada, Spain, 20 September 2018; Springer: Berlin/Heidelberg, Germany, 2018; pp. 174–182.
31. Liu, Y.; Chen, A.; Shi, H.; Huang, S.; Zheng, W.; Liu, Z.; Zhang, Q.; Yang, X. CT synthesis from MRI using multi-cycle GAN for head-and-neck radiation therapy. *Comput. Med. Imaging Graph.* **2021**, *91*, 101953. [CrossRef]
32. Harms, J.; Lei, Y.; Wang, T.; Zhang, R.; Zhou, J.; Tang, X.; Curran, W.J.; Liu, T.; Yang, X. Paired cycle-GAN-based image correction for quantitative cone-beam computed tomography. *Med. Phys.* **2019**, *46*, 3998–4009. [CrossRef]
33. Karras, T.; Laine, S.; Aila, T. A style-based generator architecture for generative adversarial networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 4401–4410.
34. Fetty, L.; Bylund, M.; Kuess, P.; Heilemann, G.; Nyholm, T.; Georg, D.; Löfstedt, T. Latent space manipulation for high-resolution medical image synthesis via the StyleGAN. *Z. Med. Phys.* **2020**, *30*, 305–314.
35. Su, K.; Zhou, E.; Sun, X.; Wang, C.; Yu, D.; Luo, X. Pre-trained StyleGAN based data augmentation for small sample brain CT motion artifacts detection. In Proceedings of the International Conference on Advanced Data Mining and Applications, Foshan, China, 12–14 November 2020; Springer: Berlin/Heidelberg, Germany, 2020, pp. 339–346.
36. Kingma, D.P.; Welling, M. Auto-encoding variational Bayes. *arXiv* **2013**, arXiv:1312.6114.
37. Vahdat, A.; Kautz, J. NVAE: A deep hierarchical variational autoencoder. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 19667–19679.
38. Cui, J.; Xie, Y.; Gong, K.; Kim, K.; Yang, J.; Larson, P.; Hope, T.; Behr, S.; Seo, Y.; Liu, H.; et al. PET denoising and uncertainty estimation based on NVAE model. In Proceedings of the IEEE Nuclear Science Symposium and Medical Imaging Conference, virtual, 16–23 October 2021; IEEE: Piscataway, NJ, USA, 2021; pp. 1–3.
39. Grover, A.; Chute, C.; Shu, R.; Cao, Z.; Ermon, S. AlignFlow: Cycle consistent learning from multiple domains via normalizing flows. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; Volume 34, pp. 4028–4035.
40. Bui, T.D.; Nguyen, M.; Le, N.; Luu, K. Flow-based deformation guidance for unpaired multi-contrast MRI image-to-image translation. In Proceedings of the 23rd International Conference on Medical Image Computing and Computer Assisted Intervention, Lima, Peru, 4–8 October 2020; Springer: Berlin/Heidelberg, Germany, 2020; pp. 728–737.

41. Wang, R.; Chaudhari, P.; Davatzikos, C. Harmonization with flow-based causal inference. In Proceedings of the 24th International Conference on Medical Image Computing and Computer Assisted Intervention, Strasbourg, France, 27 September–1 October 2021; Springer: Berlin/Heidelberg, Germany, 2021; pp. 181–190.
42. Beizaee, F.; Desrosiers, C.; Lodygensky, G.A.; Dolz, J. Harmonizing Flows: Unsupervised MR harmonization based on normalizing flows. In Proceedings of the International Conference on Information Processing in Medical Imaging, San Carlos de Bariloche, Argentina, 18–23 June 2023; Springer: Berlin/Heidelberg, Germany, 2023; pp. 347–359.
43. Kadkhodaie, Z.; Simoncelli, E.P. Solving linear inverse problems using the prior implicit in a denoiser. *arXiv* **2020**, arXiv:2007.13640.
44. Meng, C.; He, Y.; Song, Y.; Song, J.; Wu, J.; Zhu, J.Y.; Ermon, S. SDEdit: Guided image synthesis and editing with stochastic differential equations. *arXiv* **2021**, arXiv:2108.01073.
45. Sinha, A.; Song, J.; Meng, C.; Ermon, S. D2C: Diffusion-decoding models for few-shot conditional generation. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 12533–12548.
46. Sasaki, H.; Willcocks, C.G.; Breckon, T.P. UNIT-DDPM: Unpaired image translation with denoising diffusion probabilistic models. *arXiv* **2021**, arXiv:2104.05358.
47. Wolleb, J.; Bieder, F.; Sandkühler, R.; Cattin, P.C. Diffusion models for medical anomaly detection. In Proceedings of the 25th International Conference on Medical Image Computing and Computer Assisted Intervention, Singapore, 18–22 September 2022; Springer: Berlin/Heidelberg, Germany, 2022; pp. 35–45.
48. Behrendt, F.; Bhattacharya, D.; Krüger, J.; Opfer, R.; Schlaefer, A. Patched diffusion models for unsupervised anomaly detection in brain MRI. *arXiv* **2023**, arXiv:2303.03758.
49. Wyatt, J.; Leach, A.; Schmon, S.M.; Willcocks, C.G. AnoDDPM: Anomaly detection with denoising diffusion probabilistic models using simplex noise. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 650–656.
50. Rahman, A.; Valanarasu, J.M.J.; Hacihaliloglu, I.; Patel, V.M. Ambiguous medical image segmentation using diffusion models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, Canada, 18–22 June 2023; pp. 11536–11546.
51. Wu, J.; Fang, H.; Zhang, Y.; Yang, Y.; Xu, Y. MedSegDiff: Medical image segmentation with diffusion probabilistic model. *arXiv* **2022**, arXiv:2211.00611.
52. Zbinden, L.; Doorenbos, L.; Pissas, T.; Sznitman, R.; Márquez-Neila, P. Stochastic segmentation with conditional categorical diffusion models. *arXiv* **2023**, arXiv:2303.08888.
53. Chen, T.; Wang, C.; Shan, H. BerDiff: Conditional Bernoulli diffusion model for medical image segmentation. *arXiv* **2023**, arXiv:2304.04429.
54. Cao, C.; Cui, Z.X.; Liu, S.; Liang, D.; Zhu, Y. High-frequency space diffusion models for accelerated MRI. *arXiv* **2022**, arXiv:2208.05481.
55. Luo, G.; Blumenthal, M.; Heide, M.; Uecker, M. Bayesian MRI reconstruction with joint uncertainty estimation using diffusion models. *Magn. Reson. Med.* **2023**, *90*, 295–311. [CrossRef]
56. Turkbey, B.; Rosenkrantz, A.B.; Haider, M.A.; Padhani, A.R.; Villeirs, G.; Macura, K.J.; Tempany, C.M.; Choyke, P.L.; Cornud, F.; Margolis, D.J.; et al. Prostate imaging reporting and data system version 2.1: 2019 update of prostate imaging reporting and data system version 2. *Eur. Urol.* **2019**, *76*, 340–351. [CrossRef] [PubMed]
57. Kermany, D.S.; Goldbaum, M.; Cai, W.; Valentim, C.C.; Liang, H.; Baxter, S.L.; McKeown, A.; Yang, G.; Wu, X.; Yan, F.; et al. Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell* **2018**, *172*, 1122–1131. [CrossRef] [PubMed]
58. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Munich, Germany, 5–9 October 2015; Springer: Berlin/Heidelberg, Germany, 2015; pp. 234–241.
59. Wu, Y.; He, K. Group normalization. In Proceedings of the European Conference on Computer Vision, Munich, Germany, 8–14 September 2018; Springer: Berlin/Heidelberg, Germany, 2018; pp. 3–19.
60. Zhang, R.; Isola, P.; Efros, A.A.; Shechtman, E.; Wang, O. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 586–595.
61. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. ImageNet classification with deep convolutional neural networks. *Commun. ACM* **2017**, *60*, 84–90. [CrossRef]
62. 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. *Adv. Neural Inf. Process. Syst.* **2017**, 6629–6640.
63. Higgins, G.; Jones, L. The nature and evaluation of the sharpness of photographic images. *J. Soc. Motion Pict. Telev. Eng.* **1952**, *58*, 277–290. [CrossRef]
64. Dice, L.R. Measures of the amount of ecologic association between species. *Ecology* **1945**, *26*, 297–302. [CrossRef]
65. Fechner, G.T. *Elemente der Psychophysik*; Breitkopf u. Härtel: Leipzig, Germany, 1860; Volume 2.
66. Nam, H.; Lee, H.; Park, J.; Yoon, W.; Yoo, D. Reducing domain gap by reducing style bias. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 19–25 June 2021; pp. 8690–8699.

67. Romera, E.; Bergasa, L.M.; Yang, K.; Alvarez, J.M.; Barea, R. Bridging the day and night domain gap for semantic segmentation. In Proceedings of the IEEE Intelligent Vehicles Symposium, Paris, France, 9–12 June 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 1312–1318.
68. Ledig, C.; Theis, L.; Huszár, F.; Caballero, J.; Cunningham, A.; Acosta, A.; Aitken, A.; Tejani, A.; Totz, J.; Wang, Z.; et al. Photo-realistic single image super-resolution using a generative adversarial network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; IEEE: Piscataway, NJ, USA, 2017; pp. 4681–4690.
69. Hung, A.L.Y.; Zheng, H.; Miao, Q.; Raman, S.S.; Terzopoulos, D.; Sung, K. CAT-Net: A cross-slice attention transformer model for prostate zonal segmentation in MRI. *IEEE Trans. Med. Imaging* **2022**, *42*, 291–303. [CrossRef]
70. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556 .