Medical Image Generation Using Segmentation-Guided Diffusion

Florian Hunecke and Chin Ju Chen

Technical University of Munich, Germany

Abstract. While deep learning has significantly improved medical image segmentation, generating medical images from segmentation masks remains relatively unexplored. Existing generative approaches often lack precise pixel-wise anatomical constraints, limiting their ability to produce anatomically accurate medical images. To address this challenge, we propose the usage of a novel segmentation-guided diffusion-based generative framework that synthesizes high-fidelity medical images from segmentation masks, applying it to the AMOS 2022 dataset using novel, augmented segmentations. Leveraging modern diffusion architectures and the increasing availability of annotated medical datasets, our approach ensures anatomical consistency in the generated images. By explicitly conditioning the denoising process on segmentation masks, our method preserves the spatial distribution of anatomical landmarks and structural regions, enabling the generation of realistic and clinically meaningful images. Experimental results demonstrate that our framework effectively generates high-quality medical images while maintaining anatomical accuracy. Furthermore, our evaluation highlights modality-specific variations, likely influenced by differences in image intensity distributions. Our code is available at GitHub.

Keywords: diffusion model · image generation · semantic synthesis

1 Introduction

Generative models have become essential in medical imaging, enabling applications such as image synthesis [17], segmentation [26], anomaly detection [15], cross-modality translation, and denoising [7,23]. In particular, diffusion models have demonstrated strong potential in medical image segmentation, aiding clinical tasks such as disease diagnosis and treatment planning [9]. Deep learning-based segmentation models have significantly improved automation, reducing manual annotation efforts and streamlining clinical workflows [28].

Despite these advancements, the inverse problem—generating medical images from segmentation masks—remains underexplored. This capability offers benefits such as data augmentation, counterfactual analysis, and the creation of anatomically registered datasets for training machine learning models. Current GAN-based [24] and diffusion-based [27] approaches lack explicit pixel-wise anatomical constraints, limiting their ability to produce anatomically precise images.

To address this, we propose a segmentation-guided diffusion framework that reverses the segmentation process, generating realistic medical images directly from segmentation masks. By conditioning the denoising process on segmentation masks, our approach ensures anatomical consistency. Our key contributions include:

1. **Segmentation-Aware Constraints:** We integrate anatomical constraints into the diffusion process, ensuring realistic and spatially accurate medical image synthesis.
2. **Bridging Segmentation and Generative Modeling:** Our approach enhances data augmentation and synthetic dataset creation while improving the robustness of deep learning models for medical imaging.
3. **High-Fidelity Image Synthesis:** Experiments show our method enables high-resolution segmentation-to-CT/MRI translation, producing anatomically precise images for diverse medical applications.

2 Related Work

2.1 Image Diffusion Models

Diffusion models, introduced by Sohl-Dickstein et al. [21], have gained prominence in image generation [2]. Latent Diffusion Models (LDMs) [17] improve efficiency by operating in latent space, reducing computational cost while preserving quality. Text-to-image diffusion further enhances performance using pre-trained language models like CLIP [16]. Imagen [19] employs hierarchical diffusion directly in pixel space to further improve image fidelity.

2.2 Conditional Image Synthesis

Conditional synthesis approaches use diverse inputs, including labels [13], text descriptions [8], and images for translation [10]. Diffusion-based synthesis has become a robust alternative to GANs, offering improved stability and diversity, particularly for medical imaging.

SPADE [14] injects segmentation masks into normalization layers to enhance spatial control but struggles with mode collapse. ControlNet [27] improves conditioning by integrating external structural inputs directly into the diffusion process, enhancing flexibility at the cost of added computation.

Segmentation-guided diffusion [12] extends this further by enforcing anatomical accuracy, making it highly suitable for medical image generation. This combination of diffusion robustness and structural guidance ensures realistic and diagnostically relevant outputs, forming the basis for our chosen approach.

3 Methodology

3.1 Diffusion Models

As proposed in [7], Diffusion Diffusion Probabilistic Models (DDPMs) try to learn a dataset distribution $p(x_0)$ with $x_0 \in \mathbb{R}^n$ by defining a forward process

$q(x_t|x_{t-1})$, gradually converting a data point to noise, and a backward process by denoising $p_\theta(x_{t-1}|x_t)$ using a trained model with parameters θ. A sample x_0 is then generated by iteratively sampling from $p_\theta(x_{t-1}|x_t)$, beginning with a noise sample x_T $p(x_T)$ for $t = T-1, \ldots, 0$.

Using a additive prescheduled noise β_t, $\alpha_t := 1 - \beta_t$ and $\bar{\alpha}_t := \prod_{s=1}^{t} \alpha_t$ the explicit forward process to step t can be written as $x_t = \sqrt{\bar{\alpha}} x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon$. Using an approximation for the Evidence Lower Bound (ELBO) maximization to use a network $\epsilon_\theta(x_t, t)$ to predict the noise ϵ added to each datapoint x_0 for various time steps t, leads to the prominent loss

$$L = \mathbb{E}_{x_0,t,\epsilon}\left[\|\epsilon - \epsilon_\theta(x_t, t)\|^2\right] = \mathbb{E}_{x_0,t,\epsilon}\left[\|\epsilon - \epsilon_\theta(\sqrt{\bar{\alpha}} x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, t)\|^2\right]$$

3.2 Segmentation Guidance

To be able to condition the generations on segmentation masks, we follow the idea of [12]. For a c-channel image $x_0 \in \mathbb{R}^{c \times h \times w}$ and an anatomical mask $m \in \{0, \ldots, C-1\}^{h \times w}$ with C different label classes, this leads to an updated data likelihood $p(x_0|m)$, reverse process $p_\theta(x_{t-1}|x_t, m)$, and noise-predicting model ϵ_θ leading to the new loss

$$L_m = \mathbb{E}_{(x_0,m),t,\epsilon}\left[\|\epsilon - \epsilon_\theta(\sqrt{\bar{\alpha}} x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, t|m)\|^2\right]$$

The conditioning can be performed by mathematical operations or simple channel-wise concatenation of the image-mask pairs, where the latter produced the best results in this work's experiments. We additionally use the sampling approach of Denoising Diffusion Implicit Models (DDIMs) proposed by [22].

4 Experimental Setup

4.1 Data

For evaluation, we used the Multi-Modality Abdominal Multi-Organ Segmentation Challenge (AMOS) dataset from 2022 [11]. A comprehensive abdominal organ segmentation dataset with extensive annotations across multiple modalities, centers, scanners, phases, and disease conditions, covering a total of 15 organs. The dataset comprises Computed Tomography (CT) and Magnetic Resonance Imaging (MRI) scans from 600 patients diagnosed with abdominal tumors or abnormalities at Longgang District People's Hospital. We provide a model working in the 2D space of manually extracted image slices from 3D scans. We sliced along all anatomical axes to obtain 260k CT and 40k MRI slices. Unlike previous works, we augment our data with a whole body segmentation, created with TotalVibeSegmentor [5]. A full-body 3D segmentation model, developed for the NAKO and UK Biobank, providing annotations for 71 anatomical structures within the torso region. This leads to more precise and complete instructions during the generation process.

Data Distribution A challenging aspect of medical imaging is proper data normalization since different imaging techniques show different intensity distributions, as depicted in Figure 1. For CT images, the values do have absolute meaning and background pixels usually appear in a range around the minimum of -1024. To obtain high-contrast images, we limit the values to the meaningful part between the fixed values -256 and 256 before normalizing to image values between 0 and 255. The intensities in MRI do only exhibit relative meaning and typically range from 0 to 1024. It also appears in the used dataset that background pixels do not have the minimum value within a scan. To compensate for low-contrast images resulting from normalizing the whole scan with only a few pixels of actual value zero and most background pixels being normalized to some gray value, we clip the values to the first and 99th percentile beforehand.

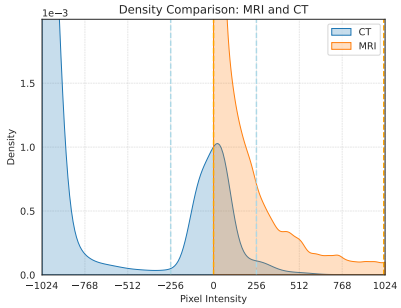


Fig. 1: Density Comparison and clipping

4.2 Architecture and Training

Following [12] and as described in subsection 3.2, a normalized and resized 256×256 image-mask pair is taken from the dataset. Different transformations like rotations, zooms, and crops yielded different stable convergences for the model. A simple center crop was used for our experiments to obtain a stable network quickly. Normalizing the mask channel and transforming using a bilinear instead of a nearest-neighbor approach yielded better-denoised images. For a random step $n \in 1, \ldots, N$ with $N = 1000$, the image is noised using the DDIM scheduler from hugging face's diffusers library. We tried different operations to add the segmentation information to the images. With addition or multiplication, the model could not extract the necessary information and generate the correct distribution of the original images. Channel-wise concatenation proved to be the most successful augmentation strategy. Subsequent, a U-Net model [18] using ResNet down-sampling blocks with channels 128, 128, 256, 256, 512, 512 and spatial self-attention and a reverse decoder was used to predict the added noise.

An auxiliary 2D segmentation model was trained on the train split of the augmented AMOS dataset to evaluate the generations. We used a standard U-Net with output channels 16, 32, 64, 128, 256, 512, kernel size of 3, and strides of 2 on images resized to 256 pixels. A batch size of 64 and a learning rate of 1e-2 were used. We also examined any differences in the model's performance when being trained on a subset of only one axis instead of all. The data was split according to the AMOS challenge, consisting of 240 train, 120 validation, and 240 test scans. The diffusion or segmentation model did not operate on the test set until sampling and evaluation.

4.3 Evaluation

This section presents the evaluation methods used for assessing the generated images. As exact reconstruction is not the objective, we focus on comparing data distributions and semantic segmentations of the generated images.

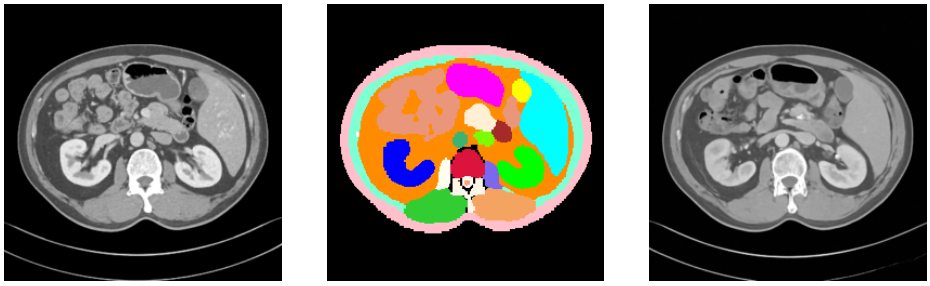
1. **Peak Signal-to-Noise Ratio (PSNR)** [4] quantifies the similarity between a processed signal and its original source. It helps assess the fidelity of the processed signal and detect potential noise or distortions introduced during processing. Essentially, it measures how much a signal degrades after undergoing modifications. The PSNR value is expressed in decibels (dB), with higher values indicating better image quality and lower distortion levels.
2. **Structural Similarity Index (SSIM)** [25] is used for assessing the perceptual similarity between two images. SSIM incorporates structural information, luminance, and contrast, providing a more perceptually relevant evaluation. SSIM values range from -1 to 1, where a value of 1 indicates identical images and lower values suggest increasing levels of distortion.
3. **Fréchet Inception Distance (FID)** [6] is based on the Fréchet distance between two distributions, one representing the original images and the other representing the generated images, in a feature space extracted by a pretrained Inception network. The key idea is to use the activations of the Inception network to represent the images in a high-dimensional space and compare the distributions of these representations. FID is lower when the generated image distribution is closer to the original image distribution, indicating higher quality in the generated images.
4. **Dice Similarity Coefficient (DSC)** is a suitable metric for semantic assessment, first introduced by [3] to measure the association between different species. While the metrics mentioned above give valuable insights into image similarity, the goal of our experiment is not to create identical reconstructions of the original image or image set. Instead, we would want an independently created image following the anatomical guidelines given by the segmentation, on which the generation process is conditioned.
   To evaluate this anatomical consistency, we compute the DSC between predicted segmentations of real and generated images. Using a pretrained segmentation model, masks are obtained for both, and DSC quantifies structural overlap. Additionally, we compute the DSC between generated masks and the corresponding real image masks, providing an upper-bound reference. This segmentation-based approach follows prior work [1,20], offering a robust metric for structural fidelity.

5 Results

5.1 Qualitative Interpretation

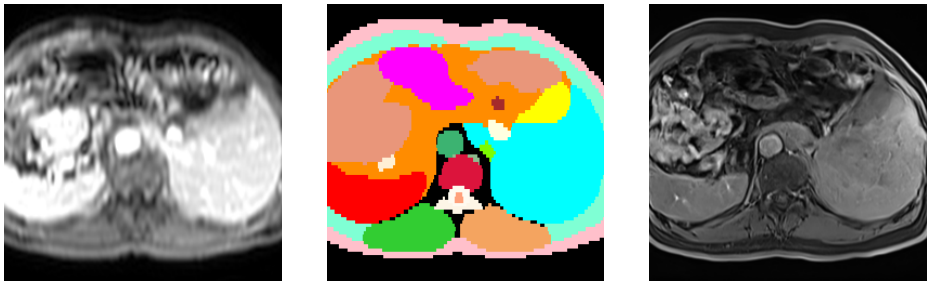
Original image and segmentation paired with an example generation can be seen for axial slices in Figure 2 for CT and Figure 3 for MRI.

We see that a complete and authentic generation of new images is possible. The conditioning on the segmentation works perfectly as all organs follow the exact outlines given by the mask, and the most important structures for an anatomically correct image are captured. Concrete details are assumed by the learned data distribution. Especially for the MRI image, we can see large quality differences to the original, where - according to the mask - a detailed image of average quality was generated without any blur as prevalent in the original image.



(a) Original (b) Segmentation (c) Generation

Fig. 2: Generation of a CT slice



(a) Original (b) Segmentation (c) Generation

Fig. 3: Generation of a MRI slice

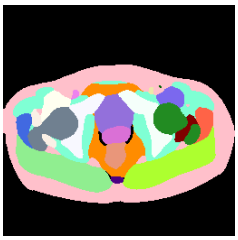
We generally observe that the reconstructions of CT images are better learned by the model, and the images look a lot clearer and more precise.

A more extensive comparison of originals and generations for the different imaging modalities can be found in Appendix A.

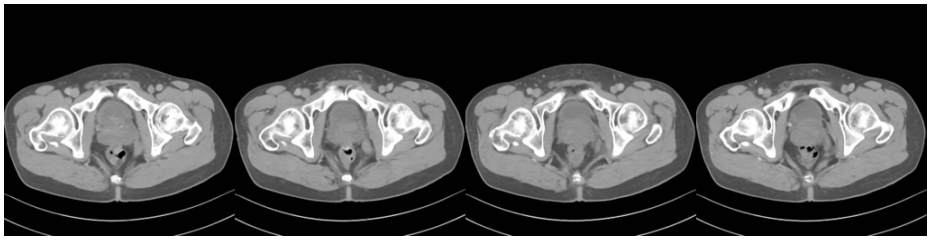
5.2 Further Application

Generation Diversity To evaluate the diversity of the generated images, we also sampled multiple images from different noise but conditioned on the same mask as can be seen in Figure 4. Whereas most structural information is kept equal, differences, for example, in tissue and air distribution, can be seen since these are not encoded into the segmentation mask.

(a) Original



(b) Segmentation



(c) Generations

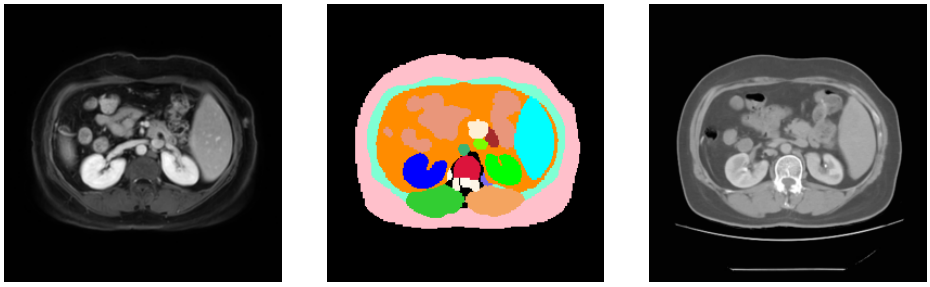
Fig. 4: Multiple CT image generations on the same segmentation

Generating out of Dataset Another interesting application of the generated images is to generate a different modality than the original one. Figure 5 shows the generation of a MRI where the segmentation was based on a CT image.

5.3 Quantitative Evaluation

For a more quantitative evaluation, we use the above-mentioned metrics and apply them to a random subset of the test dataset of at least 1024 image, mask, and generation pairs. The results are shown in Table 1.

For CT images, the axial subset achieves a PSNR of 19.06 dB, while the all-axis subset leads to slightly better results of 20.26 dB, indicating enhanced signal fidelity. While reduced complexity of datasets normally ease the fitting of the model, incorporating additional slice orientations seems to enhance anatomical



(a) Original MRI (b) Segmentation (c) Generated CT image

Fig. 5: Generation of an out of dataset MRI image

Table 1: Evaluation of test set generations on different sliced subsets

	CT axial	CT all axis	MRI axial	MRI all axis
PSNR (dB) ↑	19.06	20.26	9.49	9.80
SSIM ↑	0.637	0.618	0.062	0.170
FID ↓	47.66	61.50	200.31	202.26
DICE (m, m_{real}^{pred}) ↑	0.925	0.927	0.795	0.843
DICE $(m_{real}^{pred}, m_{gen}^{pred})$ ↑	0.923	0.927	0.712	0.766

correctness. Similarly, high DSCs coefficients (0.925 for axial and 0.927 for all-axis) reflect robust anatomical overlap between generated and original segmentation masks. In contrast, MRI results demonstrate lower PSNR values (9.49–9.80 dB) and considerably lower SSIM scores (0.062 for axial, increasing to 0.170 for all-axis), implying reduced perceptual similarity and image quality. This is very likely due to less training data and more variance in the imaging quality.

6 Conclusion

Our segmentation-guided diffusion model synthesizes high-fidelity medical images from segmentation masks. By leveraging modern diffusion architectures and the increasing availability of annotated medical datasets, our approach ensures anatomical consistency in the generated images. Explicitly conditioning the denoising process on segmentation masks preserves the spatial distribution of anatomical landmarks, enabling the generation of realistic and clinically meaningful images. This method presents a promising direction for medical image synthesis, with applications in data augmentation, synthetic dataset generation, and the development of more robust deep learning models for medical imaging analysis. However, this study has certain limitations. Future work could explore further impacts and interplay of data preprocessing and transformation techniques with diffusion noise scheduling to further refine the synthesis process. There also exist other progressive ideas such as including the dice coefficient - as used in the evaluation - directly within the loss calculation of the model.

Abbreviations

AMOS Multi-Modality Abdominal Multi-Organ Segmentation Challenge

CT Computed Tomography

DDPM Diffusion Diffusion Probabilistic Model

DDIM Denoising Diffusion Implicit Model

DSC Dice Similarity Coefficient

ELBO Evidence Lower Bound

FID Fréchet Inception Distance

LDMs Latent Diffusion Models

MRI Magnetic Resonance Imaging

PSNR Peak Signal-to-Noise Ratio

SSIM Structural Similarity Index

A Additional Images

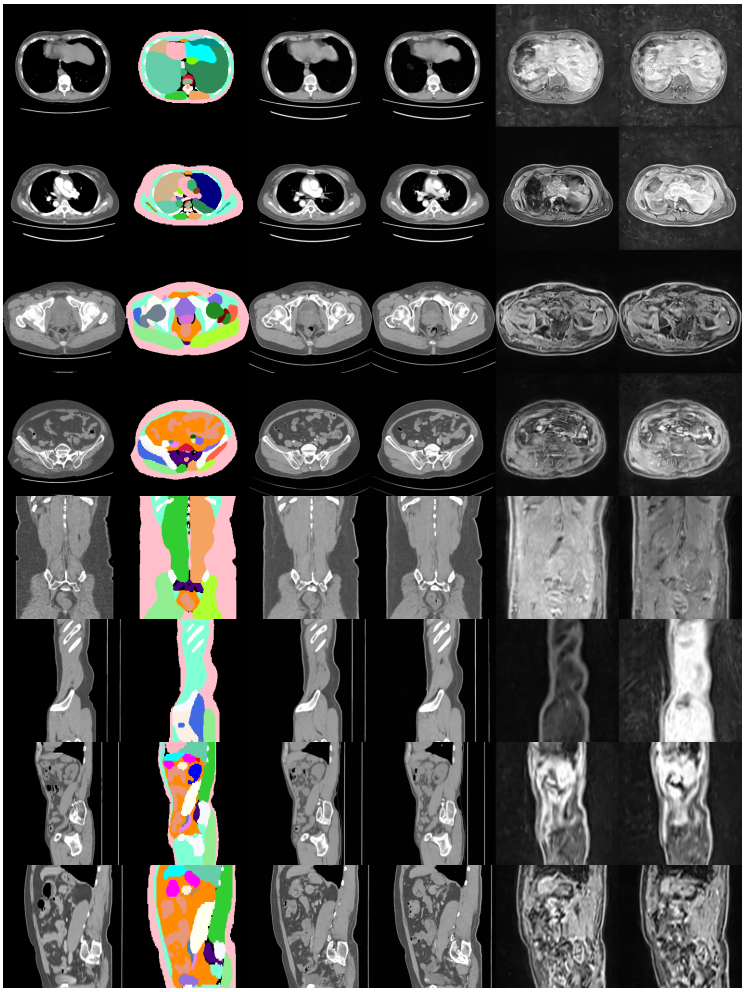


Fig. 6: Additional generations. From left to right: Original CT, segmentation, two generated CT images, two generated MRI images

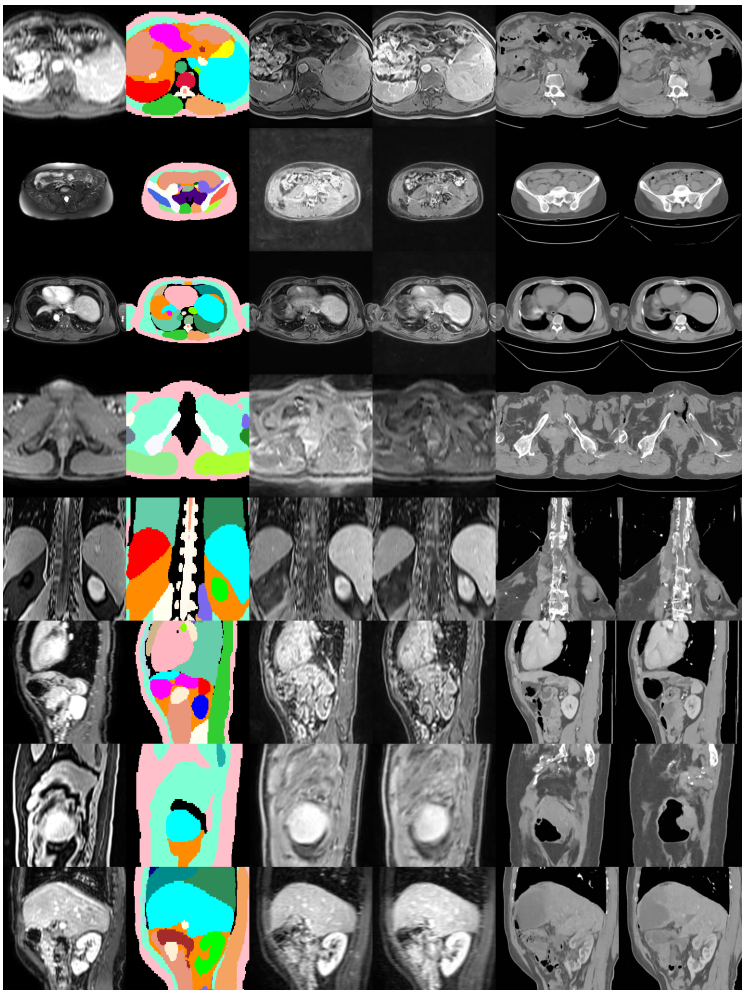


Fig. 7: Additional generations. From left to right: Original MRI, segmentation, two generated MRI images, two generated CT images

References

1. Bowles, C., Chen, L., Guerrero, R., Bentley, P., Gunn, R., Hammers, A., Dickie, D.A., Hernández, M.V., Wardlaw, J., Rueckert, D.: Gan augmentation: Augmenting training data using generative adversarial networks. arXiv preprint arXiv:1810.10863 (2018)
2. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. Advances in neural information processing systems **34**, 8780–8794 (2021)
3. Dice, L.R.: Measures of the amount of ecologic association between species. Ecology **26**(3), 297–302 (1945)
4. Fardo, F.A., Conforto, V.H., de Oliveira, F.C., Rodrigues, P.S.: A formal evaluation of psnr as quality measurement parameter for image segmentation algorithms. arXiv preprint arXiv:1605.07116 (2016)
5. Graf, R., Platzek, P.S., Riedel, E.O., Ramschütz, C., Starck, S., Möller, H.K., Atad, M., Völzke, H., Bülow, R., Schmidt, C.O., et al.: Totalvibesegmentator: Full body mri segmentation for the nako and uk biobank. arXiv preprint arXiv:2406.00125 (2024)
6. 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)
7. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)
8. Hong, S., Yang, D., Choi, J., Lee, H.: Inferring semantic layout for hierarchical text-to-image synthesis. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7986–7994 (2018)
9. Isensee, F., Petersen, J., Klein, A., Zimmerer, D., Jaeger, P.F., Kohl, S., Wasserthal, J., Koehler, G., Norajitra, T., Wirkert, S., et al.: nnu-net: Self-adapting framework for u-net-based medical image segmentation. arXiv preprint arXiv:1809.10486 (2018)
10. 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. pp. 1125–1134 (2017)
11. Ji, Y., Bai, H., Ge, C., Yang, J., Zhu, Y., Zhang, R., Li, Z., Zhanng, L., Ma, W., Wan, X., et al.: Amos: A large-scale abdominal multi-organ benchmark for versatile medical image segmentation. Advances in neural information processing systems **35**, 36722–36732 (2022)
12. Konz, N., Chen, Y., Dong, H., Mazurowski, M.A.: Anatomically-controllable medical image generation with segmentation-guided diffusion models. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 88–98. Springer (2024)
13. Mirza, M., Osindero, S.: Conditional generative adversarial nets. arXiv preprint arXiv:1411.1784 (2014)
14. Park, T., Liu, M.Y., Wang, T.C., Zhu, J.Y.: Semantic image synthesis with spatially-adaptive normalization. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2337–2346 (2019)
15. Pinaya, W.H., Graham, M.S., Gray, R., Da Costa, P.F., Tudosiu, P.D., Wright, P., Mah, Y.H., MacKinnon, A.D., Teo, J.T., Jager, R., et al.: Fast unsupervised brain anomaly detection and segmentation with diffusion models. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 705–714. Springer (2022)

16. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021)
17. 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)
18. 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)
19. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E.L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., et al.: Photorealistic text-to-image diffusion models with deep language understanding. Advances in neural information processing systems **35**, 36479–36494 (2022)
20. Shin, H.C., Tenenholtz, N.A., Rogers, J.K., Schwarz, C.G., Senjem, M.L., Gunter, J.L., Andriole, K.P., Michalski, M.: Medical image synthesis for data augmentation and anonymization using generative adversarial networks. In: Simulation and Synthesis in Medical Imaging: Third International Workshop, SASHIMI 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Proceedings 3. pp. 1–11. Springer (2018)
21. Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., Ganguli, S.: Deep unsupervised learning using nonequilibrium thermodynamics. In: International conference on machine learning. pp. 2256–2265. pmlr (2015)
22. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502 (2020)
23. Song, Y., Ermon, S.: Improved techniques for training score-based generative models. Advances in neural information processing systems **33**, 12438–12448 (2020)
24. Wang, Y., Zhang, Z., Hao, W., Song, C.: Multi-domain image-to-image translation via a unified circular framework. IEEE Transactions on Image Processing **30**, 670–684 (2020)
25. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing **13**(4), 600–612 (2004)
26. Zaman, F.A., Jacob, M., Chang, A., Liu, K., Sonka, M., Wu, X.: Denoising diffusions in latent space for medical image segmentation. arXiv preprint arXiv:2407.12952 (2024)
27. 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)
28. Zhou, Z., Rahman Siddiquee, M.M., Tajbakhsh, N., Liang, J.: Unet++: A nested u-net architecture for medical image segmentation. In: Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, Proceedings 4. pp. 3–11. Springer (2018)