

Diffusion Models for Medical Anomaly Detection

Julia Wolleb, Florentin Bieder, Robin Sandkühler, Philippe C. Cattin

Department of Biomedical Engineering, University of Basel, Allschwil, Switzerland
julia.wolleb@unibas.ch

Abstract. In medical applications, weakly supervised anomaly detection methods are of great interest, as only image-level annotations are required for training. Current anomaly detection methods mainly rely on generative adversarial networks or autoencoder models. Those models are often complicated to train or have difficulties to preserve fine details in the image. We present a novel weakly supervised anomaly detection method based on denoising diffusion implicit models. We combine the deterministic iterative noising and denoising scheme with classifier guidance for image-to-image translation between diseased and healthy subjects. Our method generates very detailed anomaly maps without the need for a complex training procedure. We evaluate our method on the BRATS2020 dataset for brain tumor detection and the CheXpert dataset for detecting pleural effusions.

Keywords: Anomaly detection · Diffusion models · Weak supervision.

1 Introduction

In medical image analysis, pixel-wise annotated ground truth is hard to obtain, often unavailable and contains a bias to the human annotators. Weakly supervised anomaly detection has gained a lot of interest in research as an essential tool to overcome the aforementioned issues. Compared to fully supervised methods, weakly supervised models rely only on image-level labels for training. In this paper, we present a novel pixel-wise anomaly detection approach based on Denoising Diffusion Implicit Models (DDIMs) [25]. Figure 1 shows an overview

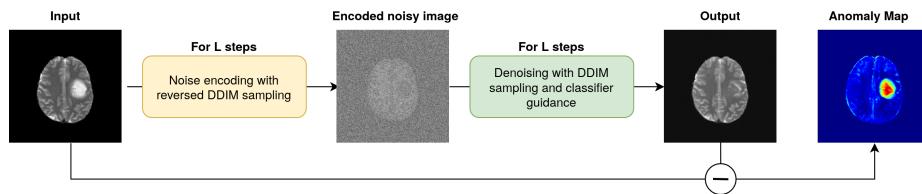


Fig. 1. Proposed sampling scheme for image-to-image translation between a diseased input image and a healthy output image. The anomaly map is defined as the difference between the two.

of the proposed method. We assume two unpaired sets of images for the training, the first containing images of healthy subjects and the second images of subjects affected by a disease. Only the image and the corresponding image-level label (healthy, diseased) are provided during training.

Our method consists of two main parts. In the first part, we train a Denoising Diffusion Probabilistic Models (DDPM)[10] and a binary classifier on a dataset of healthy and diseased subjects. In the second part, we create the actual anomaly map of an unseen image. For this, we first encode the anatomical information of an image with the reversed sampling scheme of DDIMs. This is an iterative noising process. Then, in the denoising process, we use the deterministic sampling scheme proposed in DDIM with classifier guidance to generate an image of a healthy subject. The final pixel-wise anomaly map is the difference between the original and the synthetic image. With this encoding and denoising procedure, our method can preserve many details of the input image that are not affected by the disease while re-painting the diseased part with realistic looking tissue. We apply our algorithm on two different medical datasets, i.e., the BRATS2020 brain tumor challenge [16,2,3], and the CheXpert dataset [11], and compare our method against standard anomaly detection methods. The source code and implementation details are available at <https://gitlab.com/cian.unibas.ch/diffusion-anomaly>.

Related Work In classical anomaly detection, autoencoders [29,13] are trained on data of healthy subjects. Any deviations from the learned distribution then lead to a high anomaly score. This idea has been applied for unsupervised anomaly detection in medical images [30,6,14], where the difference between the healthy reconstruction and the anomalous input image highlight pixels that are perceived as anomalous. Other approaches focus on Generative Adversarial Networks (GANs) [9] for image-to-image translation [24,5,27].

However, training of GANs is challenging and requires a lot of hyperparameter tuning. Furthermore, additional loss terms and changes to the architecture are required to ensure cycle-consistent results. In [19,1], the gradient of a classifier is used to obtain anomaly maps. Recently, transformer networks [21] were also successfully applied on brain anomaly detection [20]. Non-synthesis based methods such as density estimation, feature modeling or self-supervised classification also provide state-of-the-art techniques for anomaly detection [28]. In [15], a new thresholding method is proposed for anomaly segmentation on the BRATS dataset.

Lately, DDPMs were in focus for their ability to beat GANs on image synthesis [8]. In the flow of this success, they were also applied on image-to-image translation [23,7], segmentation [4], reconstruction [22] and registration[12]. As shown in [25], DDIMs are closely related to score-based generative models [26], which can be used for interpolation between images. However, there is no diffusion model for anomaly detection so far to the best of our knowledge.

2 Method

A typical example for image-to-image translation in medicine is the transformation of an image of a patient to an image without any pathologies. For anomaly detection it is crucial that only pathological regions are changed, and the rest of the image is preserved. Then, the difference between the original and the translated image defines the anomaly map. Our detail-preserving image-to-image translation is based on diffusion models. We follow the formulation of DDPMs given in [10,17]. In Algorithm 1, we present the workflow of our approach.

The general idea of diffusion models is that for an input image x , we generate a series of noisy images $\{x_0, x_1, \dots, x_T\}$ by adding small amounts of noise for many timesteps T . The noise level t of an image x_t is steadily increased from 0 to T . A U-Net ϵ_θ is trained to predict x_{t-1} from x_t according to (5), for any step $t \in \{1, \dots, T\}$. During training, we know the ground truth for x_{t-1} , and the model is trained with an MSE loss. During evaluation, we start from $x_T \sim \mathcal{N}(0, \mathbf{I})$ and predict x_{t-1} for $t \in \{T, \dots, 1\}$. With this iterative denoising process, we can generate a fake image x_0 . The forward noising process q with variances β_1, \dots, β_T is defined by

$$q(x_t|x_{t-1}) := \mathcal{N}(x_t; \sqrt{1 - \beta_t}x_{t-1}, \beta_t\mathbf{I}). \quad (1)$$

This recursion can be written explicitly as

$$x_t = \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1 - \bar{\alpha}_t}\epsilon, \quad \text{with } \epsilon \sim \mathcal{N}(0, \mathbf{I}). \quad (2)$$

with $\alpha_t := 1 - \beta_t$ and $\bar{\alpha}_t := \prod_{s=1}^t \alpha_s$. The denoising process p_θ is learned by optimizing the model parameters θ and is given by

$$p_\theta(x_{t-1}|x_t) := \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t)). \quad (3)$$

The output of the U-Net is denoted as ϵ_θ , and the MSE loss used for training is

$$\mathcal{L} := \|\epsilon - \epsilon_\theta(\sqrt{\bar{\alpha}_t}x_0 + \sqrt{1 - \bar{\alpha}_t}\epsilon, t)\|_2^2, \quad \text{with } \epsilon \sim \mathcal{N}(0, \mathbf{I}). \quad (4)$$

As shown in [25], we use the DDPM formulation to predict x_{t-1} from x_t with

$$x_{t-1} = \sqrt{\bar{\alpha}_{t-1}} \left(\frac{x_t - \sqrt{1 - \bar{\alpha}_t}\epsilon_\theta(x_t, t)}{\sqrt{\bar{\alpha}_t}} \right) + \sqrt{1 - \bar{\alpha}_{t-1} - \sigma_t^2}\epsilon_\theta(x_t, t) + \sigma_t\epsilon, \quad (5)$$

with $\sigma_t = \sqrt{(1 - \bar{\alpha}_{t-1})/(1 - \bar{\alpha}_t)}\sqrt{1 - \bar{\alpha}_t/\bar{\alpha}_{t-1}}$. DDPMs have a stochastic element ϵ in each sampling step (5). In DDIMs however, we set $\sigma_t = 0$, which results in a deterministic sampling process. As derived in [25], (5) can be viewed as the Euler method to solve an ordinary differential equation (ODE). Consequently, we can reverse the generation process by using the reversed ODE. Using enough discretization steps, we can encode x_{t+1} given x_t with

$$x_{t+1} = x_t + \sqrt{\bar{\alpha}_{t+1}} \left[\left(\sqrt{\frac{1}{\bar{\alpha}_t}} - \sqrt{\frac{1}{\bar{\alpha}_{t+1}}} \right) x_t + \left(\sqrt{\frac{1}{\bar{\alpha}_{t+1}} - 1} - \sqrt{\frac{1}{\bar{\alpha}_t} - 1} \right) \epsilon_\theta(x_t, t) \right]. \quad (6)$$

By applying (6) for $t \in \{0, \dots, T-1\}$, we can encode an image x_0 in a noisy image x_T . Then, we recover the identical x_0 from x_T by using (5) with $\sigma_t = 0$ for $t \in \{T, \dots, 1\}$.

For anomaly detection, we train a DDPM on a dataset containing images of healthy and diseased subjects. For evaluation, we define a noise level $L \in \{1, \dots, T\}$ and a gradient scale s . Given an input image x , we encode it to a noisy image x_L using (6) for $t \in \{0, \dots, L-1\}$. With this iterative noising process, we can induce anatomical information of the input image. During the denoising process, we follow (5) with $\sigma_t = 0$ for $t \in \{L, \dots, 1\}$. We apply classifier guidance as introduced in [8] to lead the image generation to the desired healthy class h . For this, we pretrain a classifier network C on the noisy images x_t for $t \in \{1, \dots, T\}$, to predict the class label of x . During the denoising process, the scaled gradient $s\nabla_{x_t} \log C(h|x_t, t)$ of the classifier is used to update $\epsilon_\theta(x_t, t)$. This iterative noising and denoising scheme is presented in Algorithm 1. We generate an image x_0 of the desired class h that preserves the basic structure of x . The anomaly map is then defined by the difference between x and x_0 . The choice of the noise level L and the gradient scale s is crucial for the trade-off between detail-preserving image reconstruction and freedom for translation to a healthy subject.

Algorithm 1 Anomaly detection using noise encoding and classifier guidance

Input: input image x , healthy class label h , gradient scale s , noise level L
Output: synthetic image x_0 , anomaly map a

```

for all  $t$  from 0 to  $L-1$  do
     $x_{t+1} \leftarrow x_t + \sqrt{\bar{\alpha}_{t+1}} \left[ \left( \sqrt{\frac{1}{\bar{\alpha}_t}} - \sqrt{\frac{1}{\bar{\alpha}_{t+1}}} \right) x_t + \left( \sqrt{\frac{1}{\bar{\alpha}_{t+1}} - 1} - \sqrt{\frac{1}{\bar{\alpha}_t} - 1} \right) \epsilon_\theta(x_t, t) \right]$ 
end for
for all  $t$  from  $L$  to 1 do
     $\hat{\epsilon} \leftarrow \epsilon_\theta(x_t, t) - s\sqrt{1 - \bar{\alpha}_t} \nabla_{x_t} \log C(h|x_t, t)$ 
     $x_{t-1} \leftarrow \sqrt{\bar{\alpha}_{t-1}} \left( \frac{x_t - \sqrt{1 - \bar{\alpha}_t} \hat{\epsilon}}{\sqrt{\bar{\alpha}_t}} \right) + \sqrt{1 - \bar{\alpha}_{t-1}} \hat{\epsilon}$ 
end for
 $a \leftarrow \sum_{\text{channels}} |x - x_0|$ 
return  $x_0, a$ 

```

3 Experiments

The DDPM is trained as proposed in [17] without data augmentation. We choose the hyperparameters for the DDPM model as described in the appendix of [8], for $T = 1000$ sampling steps. The model is trained with the Adam optimizer and the hybrid loss objective described in [17], with a learning rate of 10^{-4} , and a batch size of 10. By choosing the number of channels in the first layer as 128, and using one attention head at resolution 16, the total number of parameters is 113,681,160 for the diffusion model and 5,452,962 for the classifier. We train the class-conditional DDPM model for 50,000 iterations and the classifier network

for 20,000 iterations, which takes about one day on an NVIDIA Quadro RTX 6000 GPU. We used Pytorch 1.7.1 as software framework. The CheXpert and the BRATS2020 dataset are used for the evaluation of our method.

CheXpert This dataset contains lung X-ray images. For training, we choose 14,179 subjects of the healthy control group, as well as 16,776 subjects suffering from pleural effusions. The images are of size 256×256 and normalized to values between 0 and 1. The test set comprises 200 images of each class.

BRATS2020 This dataset contains 3D brain Magnetic Resonance (MR) images of subjects with a brain tumor, as well as pixel-wise ground truth labels. Every subject is scanned with four different MR sequences, namely, T1-weighted, T2-weighted, FLAIR, and T1-weighted with contrast enhancement. Since we focus on a 2D approach, we only consider axial slices. Each slice contains the aforementioned four channels, is padded to a size of 256×256 , and normalized to values between 0 and 1. Since tumors mostly occur in the middle of the brain, we exclude the lowest 80 slices and the uppermost 26 slices. A slice is considered healthy if no tumor is found on the ground truth label mask. All other slices get the image-level label *diseased*. Our training set includes 5,598 healthy slices, and 10,607 diseased slices. The test set consists of 1,082 slices containing a tumor, and 705 slices without.

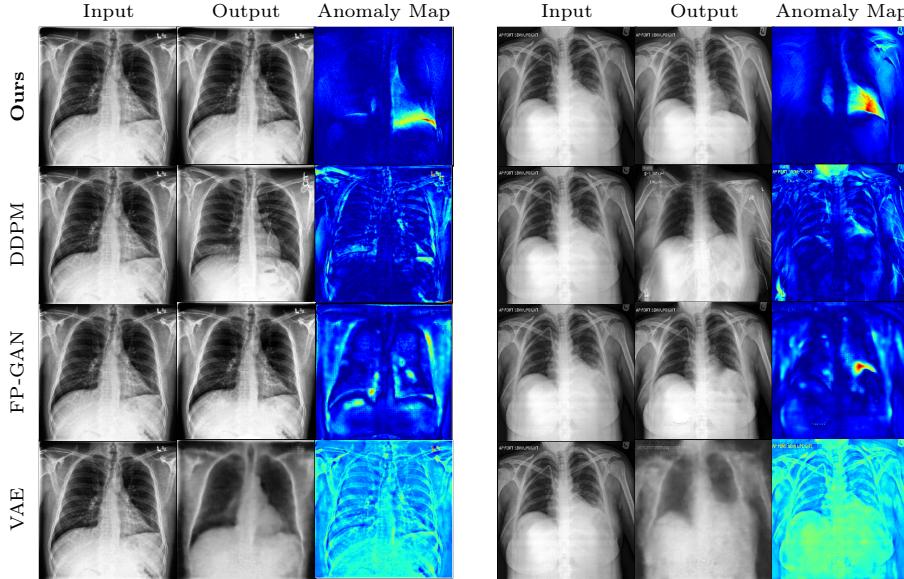


Fig. 2. Results for two X-ray images of the CheXpert dataset for $L = 500$ and $s = 100$.

4 Results and Discussion

For the evaluation of our method, we compare our method to the Fixed-Point GAN (FP-GAN) [24], and the variational autoencoder (VAE) proposed in [6]. As an ablation study, we add random noise for L steps to the input image using and perform the sampling using the DDPM sampling scheme with classifier guidance. In all experiments, we set $s = 100$ and $L = 500$. In Figure 2, we show two exemplary patient images of the CheXpert dataset, and apply all comparing methods to generate the corresponding healthy image. We observe that compared to the other methods, our approach generates realistic looking images and preserves all the details of the input image, which leads to a very detailed anomaly map. The other methods either change other parts image, or are not able to find an anomaly. Figure 3 shows the results for all four MR sequences for an exemplary image of the BRATS2020 dataset. More examples can be found in

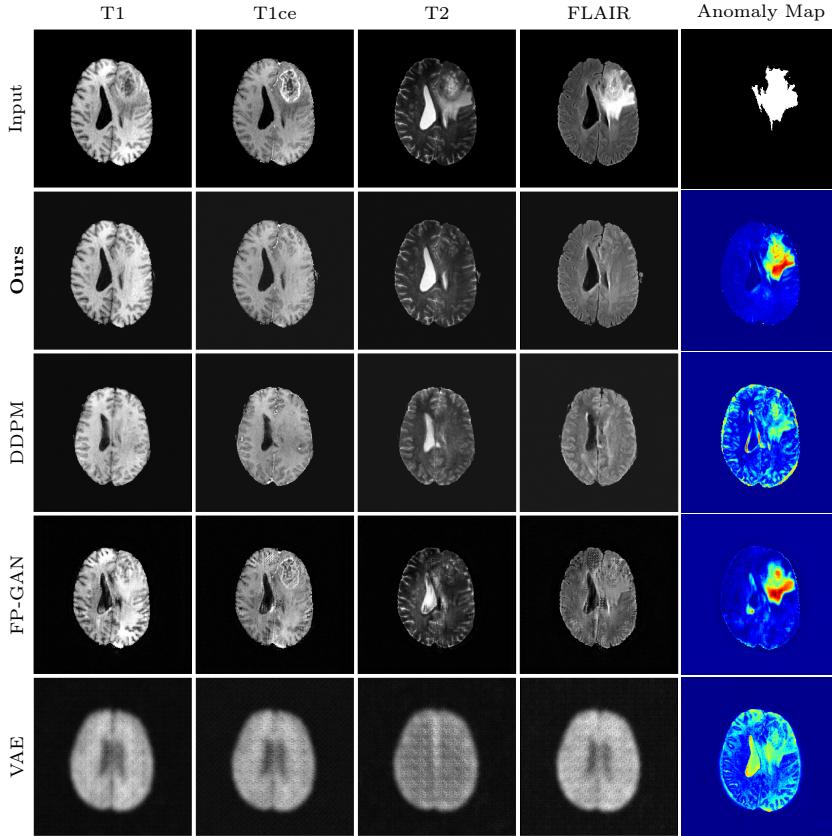


Fig. 3. Results for an image of the BRATS2020 dataset for $L = 500$ and $s = 100$.

the supplementary material. Of all methods, only the VAE tries to reconstruct the right ventricle. Comparing our results to the results of DDPM, we see that encoding information in noise using the deterministic noising process of DDIM brings the advantage that all details of the input image can be reconstructed. In contrast, we see that sampling with the DDPM approach changes the basic anatomy of the input image. The computation of a complete image translation takes about 158s. This longish running time is mainly due to the iterative image generation process. We could speed up this process by choosing a smaller L , or by skipping timesteps in the DDIM sampling scheme. However, we observed that this degrades the image quality.

In [15], using the reconstruction error as anomaly score has received some criticism. It was shown that a simple method based on histogram equalization could outperform neural networks and state that reconstruction quality does not correlate well with the Dice score. As an alternative, anomaly scores of other types of methods, i.e., the log-likelihood of density estimation models, will be explored in future work.

Hyperparameter Sensitivity Our method has two major hyperparameters, the classifier gradient scale s and the noise level L . We performed experiments to evaluate the sensitivity of our method to changes of s and L . On the BRATS2020 dataset, we have pixel-wise ground truth labels, which enable us to calculate the Dice score and the Area under the receiver operating statistics (AUROC) for diseased slices. For the Dice score, we use the average Otsu thresholding [18] on the anomaly maps. In Figure 4, we show the average Dice and AUROC scores on the test set with respect to the gradient scale s for different noise levels L . The scores for the comparing methods FP-GAN and VAE are shown in horizontal bars in Figure 4.

Figure 5 shows an exemplary FLAIR image. We fix $L = 500$ and show the sampled results for various values of s . If we choose s too small, the tumor cannot be removed. However, if we choose s too large, additional artefacts are introduced

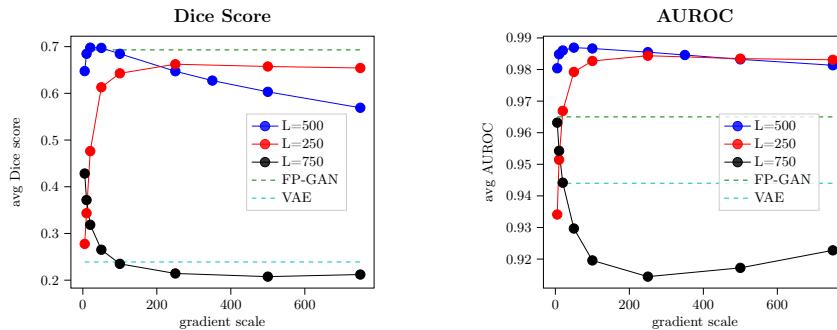


Fig. 4. Average Dice and AUROC scores on the test set for different s and L .

to the image. Those artefacts are mainly at the border of the brain, and lead to a decrease in the Dice score. In Figure 6, we fix $s = 100$, and show the sampled results for the same image for varying noise levels L . If L is chosen too large, this results in a destruction of the images. If L is chosen too small, the model does not have enough freedom to remove the tumor from the image.

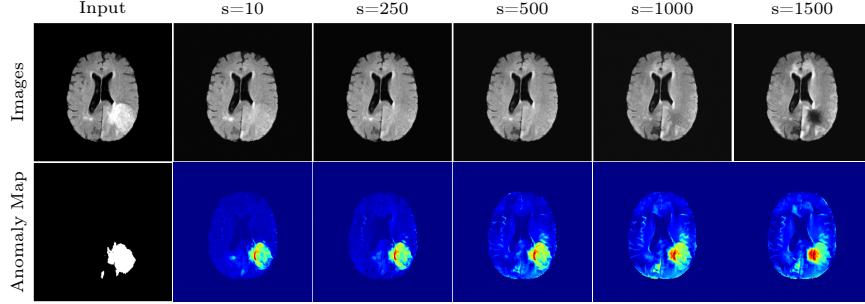


Fig. 5. Illustration of the effect of the gradient scale s for a fixed noise level $L = 500$.

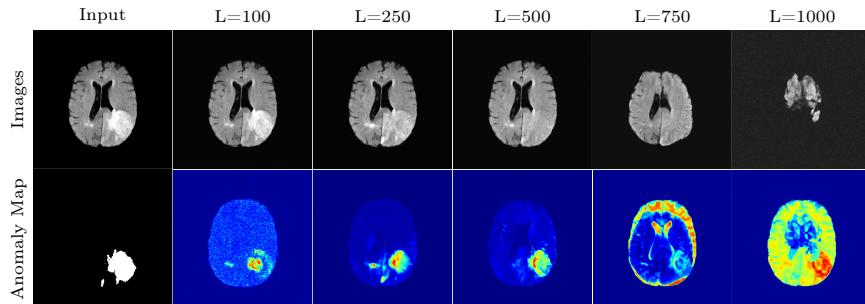


Fig. 6. Illustration of the effect of the noise level L , for a fixed gradient scale $s = 100$.

Translation of a Healthy Subject If an input image shows a healthy subject, our method should not make any changes to this image. In Figure 7, we evaluate our approach on a healthy slice of the BRATS dataset. We get a very detailed reconstruction of the image, resulting in an anomaly map close to zero.

5 Conclusion

In this paper, we presented a novel weakly supervised anomaly detection method by combining the iterative DDIM noising and denoising schemes, and classifier

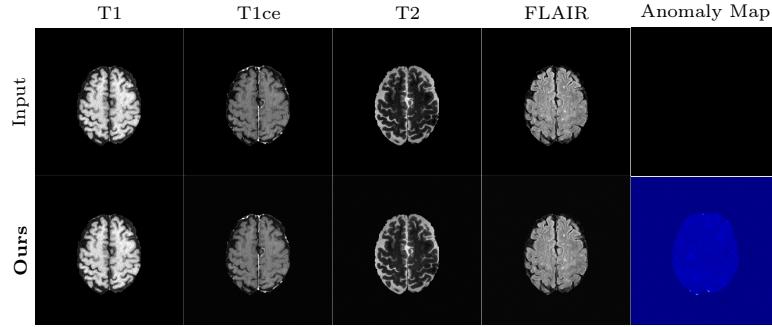


Fig. 7. Results of the presented method for an image without a tumor. The difference between the input image and the synthetic image is close to zero.

guidance. No changes were made to the loss function or the training scheme of the original implementations, making the training on other datasets straightforward. We applied our method for anomaly detection on two different medical datasets and successfully translated images of patients to images without pathologies. Our method only performs changes in the anomalous regions of the image to achieve the translation to a healthy subject. This improves the quality of the anomaly maps. We point out that we achieve a detail-consistent image-to-image translation without the need of changing the architecture or training procedure. We achieve excellent results on the BRATS2020 and the CheXpert dataset.

Acknowledgements This research was supported by the Novartis FreeNnovation initiative and the Uniscientia Foundation (project #147-2018).

References

1. Arun, N.T., Gaw, N., Singh, P., Chang, K., Hoebel, K.V., Patel, J., Gidwani, M., Kalpathy-Cramer, J.: Assessing the validity of saliency maps for abnormality localization in medical imaging. arXiv preprint arXiv:2006.00063 (2020)
2. Bakas, S., Akbari, H., Sotiras, A., Bilello, M., Rozycki, M., Kirby, J.S., Freymann, J.B., Farahani, K., Davatzikos, C.: Advancing the cancer genome atlas glioma MRI collections with expert segmentation labels and radiomic features. Scientific data **4**(1), 1–13 (2017)
3. Bakas, S., Reyes, M., Jakab, A., Bauer, S., Rempfler, M., Crimi, A., Shinohara, R.T., Berger, C., Ha, S.M., Rozycki, M., et al.: Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the BRATS challenge. arXiv preprint arXiv:1811.02629 (2018)
4. Baranchuk, D., Voynov, A., Rubachev, I., Khrulkov, V., Babenko, A.: Label-efficient semantic segmentation with diffusion models. In: International Conference on Learning Representations (2022)

5. Baumgartner, C.F., Koch, L.M., Tezcan, K.C., Ang, J.X., Konukoglu, E.: Visual feature attribution using wasserstein gans. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 8309–8319 (2018)
6. Chen, X., Konukoglu, E.: Unsupervised detection of lesions in brain mri using constrained adversarial auto-encoders. arXiv preprint arXiv:1806.04972 (2018)
7. Choi, J., Kim, S., Jeong, Y., Gwon, Y., Yoon, S.: Ilvr: Conditioning method for denoising diffusion probabilistic models. arXiv preprint arXiv:2108.02938 (2021)
8. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. Advances in Neural Information Processing Systems **34** (2021)
9. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. Advances in neural information processing systems **27** (2014)
10. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in Neural Information Processing Systems **33**, 6840–6851 (2020)
11. Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., Marklund, H., Haghgoo, B., Ball, R., Shpanskaya, K., et al.: Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In: Proceedings of the AAAI conference on artificial intelligence. vol. 33, pp. 590–597 (2019)
12. Kim, B., Han, I., Ye, J.C.: Diffusemorph: Unsupervised deformable image registration along continuous trajectory using diffusion models. arXiv preprint arXiv:2112.05149 (2021)
13. Kingma, D.P., Welling, M.: An introduction to variational autoencoders. arXiv preprint arXiv:1906.02691 (2019)
14. Marimont, S.N., Tarroni, G.: Anomaly detection through latent space restoration using vector quantized variational autoencoders. In: 2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI). pp. 1764–1767. IEEE (2021)
15. Meissen, F., Kaassis, G., Rueckert, D.: Challenging current semi-supervised anomaly segmentation methods for brain mri. arXiv preprint arXiv:2109.06023 (2021)
16. Menze, B.H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., Burren, Y., Porz, N., Slotboom, J., Wiest, R., et al.: The multimodal brain tumor image segmentation benchmark (BRATS). IEEE transactions on medical imaging **34**(10), 1993–2024 (2014)
17. Nichol, A.Q., Dhariwal, P.: Improved denoising diffusion probabilistic models. In: Proceedings of the 38th International Conference on Machine Learning. vol. 139, pp. 8162–8171. PMLR (2021)
18. Otsu, N.: A threshold selection method from gray-level histograms. IEEE transactions on systems, man, and cybernetics **9**(1), 62–66 (1979)
19. Panwar, H., Gupta, P., Siddiqui, M.K., Morales-Menendez, R., Bhardwaj, P., Singh, V.: A deep learning and grad-cam based color visualization approach for fast detection of covid-19 cases using chest x-ray and ct-scan images. Chaos, Solitons & Fractals **140**, 110190 (2020)
20. Pinaya, W.H.L., Tudosi, P.D., Gray, R., Rees, G., Nachev, P., Ourselin, S., Cardoso, M.J.: Unsupervised brain anomaly detection and segmentation with transformers. arXiv preprint arXiv:2102.11650 (2021)
21. Pirnay, J., Chai, K.: Inpainting transformer for anomaly detection. arXiv preprint arXiv:2104.13897 (2021)
22. Saharia, C., Chan, W., Chang, H., Lee, C.A., Ho, J., Salimans, T., Fleet, D.J., Norouzi, M.: Palette: Image-to-image diffusion models. arXiv preprint arXiv:2111.05826 (2021)

23. Sasaki, H., Willcocks, C.G., Breckon, T.P.: Unit-ddpm: Unpaired image translation with denoising diffusion probabilistic models. arXiv preprint arXiv:2104.05358 (2021)
24. Siddiquee, M.M.R., Zhou, Z., Tajbakhsh, N., Feng, R., Gotway, M.B., Bengio, Y., Liang, J.: Learning fixed points in generative adversarial networks: From image-to-image translation to disease detection and localization. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 191–200 (2019)
25. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502 (2020)
26. Song, Y., Sohl-Dickstein, J., Kingma, D.P., Kumar, A., Ermon, S., Poole, B.: Score-based generative modeling through stochastic differential equations. arXiv preprint arXiv:2011.13456 (2020)
27. Wolleb, J., Sandkühler, R., Cattin, P.C.: Descargan: Disease-specific anomaly detection with weak supervision. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 14–24. Springer (2020)
28. Yang, J., Xu, R., Qi, Z., Shi, Y.: Visual anomaly detection for images: A survey. arXiv preprint arXiv:2109.13157 (2021)
29. Zhou, C., Paffenroth, R.C.: Anomaly detection with robust deep autoencoders. In: Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining. pp. 665–674 (2017)
30. Zimmerer, D., Kohl, S.A., Petersen, J., Isensee, F., Maier-Hein, K.H.: Context-encoding variational autoencoder for unsupervised anomaly detection. arXiv preprint arXiv:1812.05941 (2018)

Supplementary Material

1 Additional Results on the CheXpert Dataset

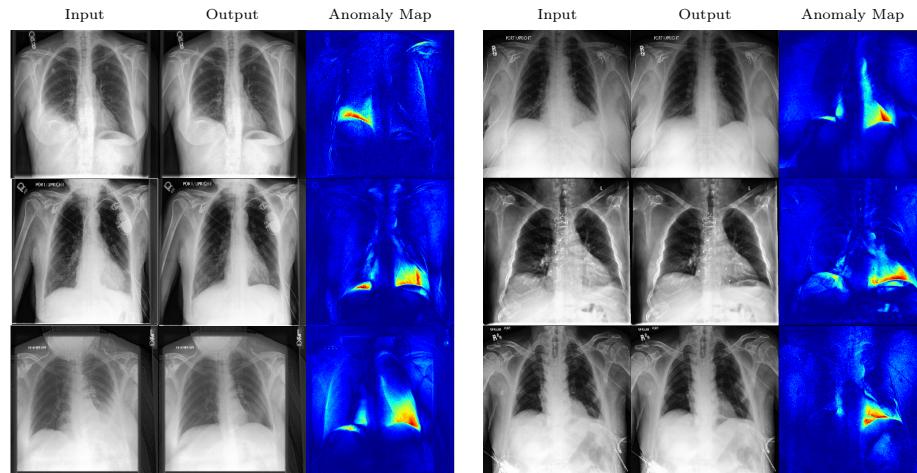


Fig. 1. Additional results of our method for diseased subjects, for $L = 500$ and $s = 100$.

2 Additional Results on the Brats2020 Dataset

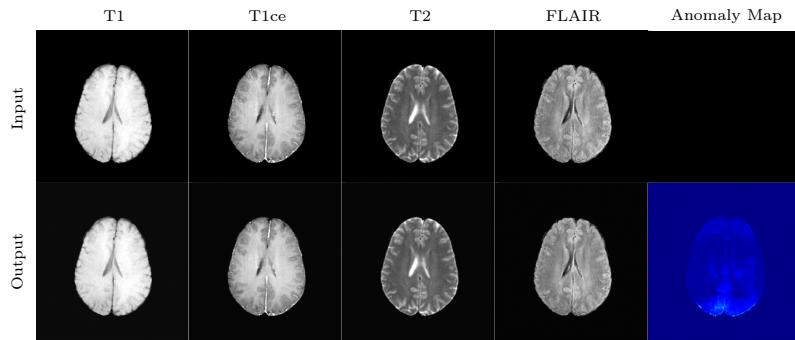


Fig. 2. Results of our method for a healthy subject, for $L = 500$ and $s = 100$.

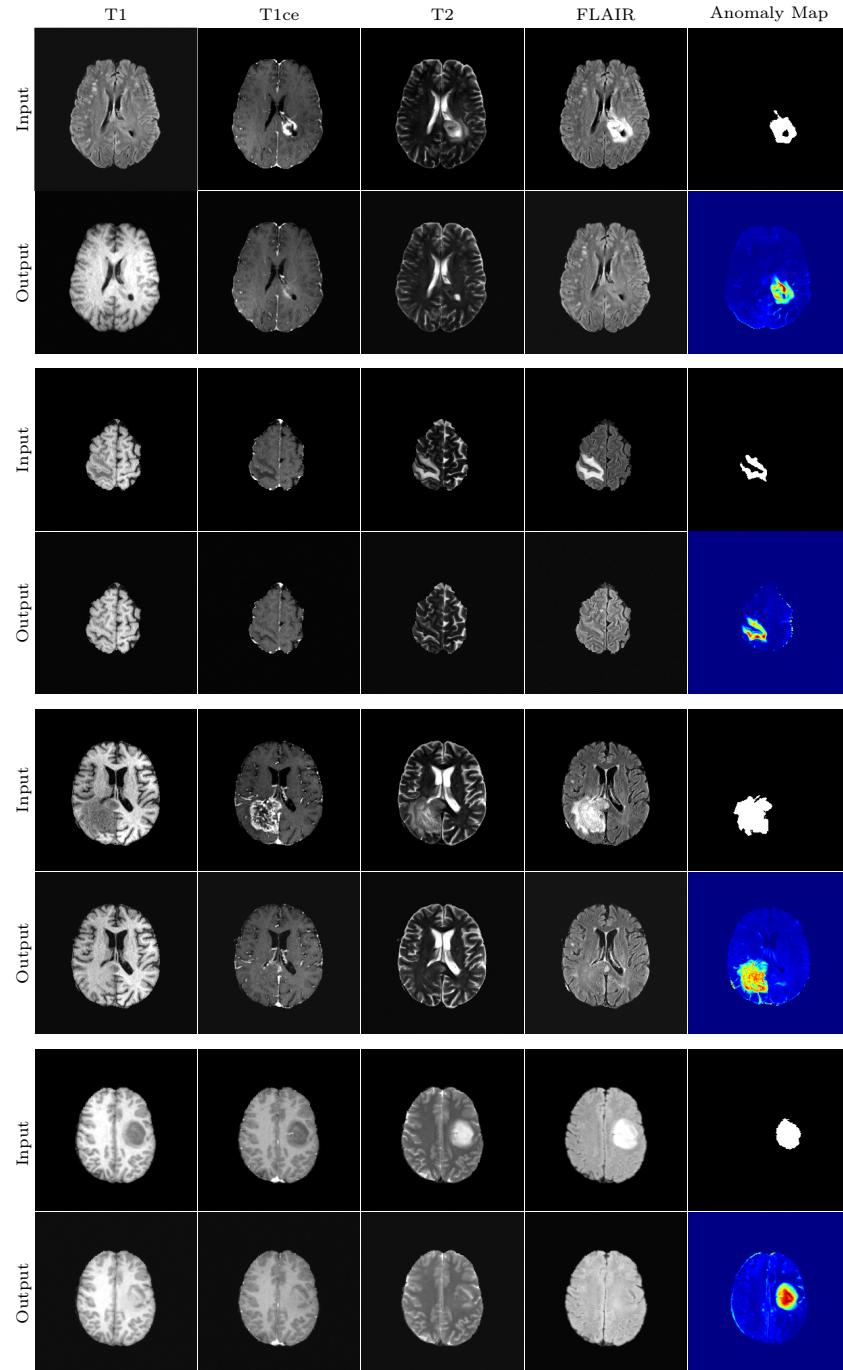


Fig. 3. Additional results of our method for diseased subjects, for $L = 500$ and $s = 100$.