Patched Diffusion Models for Unsupervised Anomaly Detection in Brain MRI

Author name(s) withheld EMAIL(S) WITHHELD
Address withheld

Editors: Under Review for MIDL 2023

Abstract

The use of supervised deep learning techniques to detect pathologies in brain MRI scans can be challenging due to the diversity of brain anatomy and the need for large, pixel-level annotated data sets. An alternative approach is to use unsupervised anomaly detection, which only requires sample-level labels of healthy brain anatomy to create a reference representation. This reference representation can then be compared to unhealthy brain anatomy in a pixel-wise manner to identify abnormalities. To accomplish this, generative models are needed to create anatomically consistent MRI scans of healthy brains. While recent diffusion models have shown promise in this task, accurately generating the complex structure of the human brain remains a challenge. In this paper, we propose a method that reformulates the generation task of diffusion models as a patch-based estimation of healthy brain anatomy, using spatial context to guide and improve reconstruction. We evaluate our approach on data of tumors and multiple sclerosis lesions and demonstrate a relative improvement of 25.1% in segmentation performance compared to existing baselines.

Keywords: Deep Learning, Unsupervised Anomaly Detection, MRI, Diffusion Models

1. Introduction

Over the last decades, significant effort has been put into developing support tools that can assist radiologists with tasks such as automatic classification, detection, or segmentation of pathologies in medical images (Kawamoto et al., 2005). Convolutional neural networks (CNNs) have proven successful in these tasks due to their ability to process images effectively (Shen et al., 2017). However, supervised approaches that use CNNs have limitations, such as the need for large amounts of expert-annotated training data and the challenge of learning from noisy or imbalanced data (Ellis et al., 2022; Karimi et al., 2020; Johnson and Khoshgoftaar, 2019). Furthermore, by design, deep learning models that are trained in a supervised fashion can only detect pathologies that are existent in the training data distribution.

Unsupervised anomaly detection (UAD) is an alternative approach that can be trained with healthy samples only, eliminating the need for pixel-level annotations or balanced data sets. During training, UAD models typically focus on reconstructing images from a healthy training distribution. When unseen, unhealthy anatomy is encountered at test time, high values in the pixel-wise reconstruction error indicate abnormalities.

Recently, denoising probabilistic diffusion models (DDPM) (Ho et al., 2020) have emerged as a state-of-the-art approach for image generation. As a result, they have also been applied to the problem of unsupervised anomaly detection (UAD) in brain MRI (Wyatt et al., 2022; Pinaya et al., 2022a). DDPMs work by adding noise to an input image, then using a

© 2023 Anonymous.

trained model to remove the noise and estimate or reconstruct the original image. Hence, in contrast to most autoencoder-based approaches, DDPMs preserve spatial information in their hidden representation of the input which is important for the image generation process (Rombach et al., 2022). However, applying noise to the entire image at once can make it difficult to accurately reconstruct the complex structure of the brain. To address this issue, we introduce patched DDPMs (pDDPMs) for UAD in brain MRI. In pDDPMs, we apply the forward diffusion process only on a small part of the input image and use the whole, partly noised image in the backward process to recover the noised patch. At test time, we use the trained pDDPM to sequentially noise and denoise a sliding patch within the input image and then stitch the individual denoised patches to reconstruct the entire image. This patch-wise approach allows us to consider the global context of the image while estimating the individual brain anatomy. We evaluate our method on the public BraTS21 and MSLUB data sets, reporting on reconstruction error, AUPRC, and Dice scores. Our results show that pDDPM significantly ($p < 0.05$) improves the tumor segmentation performance with a Dice score of 49.01% compared to 40.67% for the baseline DDPM.

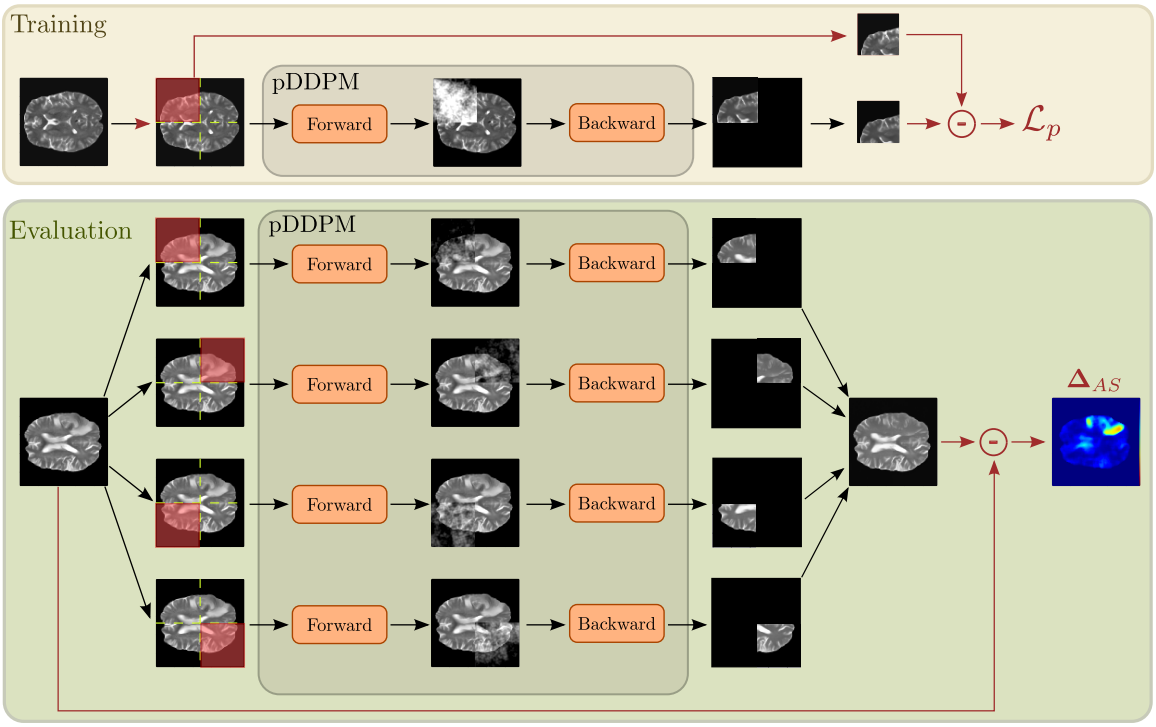

Figure 1: Schematic drawing of our method. From left to right: A patch is sampled within the input image, noise is added to that patch in the forward process and removed in the backward process. During evaluation, we stitch all individual reconstructed patches and calculate the pixel-wise absolute error as anomaly map $\mathbf{\Delta}_{AS}$. In case of overlapping patches, we use the average of overlapping regions.

2. Recent Work

In recent research on UAD in brain MRI, various architectures have been examined. Autoencoders (AE) and variational autoencoders (VAE) have demonstrated reliable training and fast inference, but their blurry reconstructions have hindered their effectiveness in UAD, as noted in (Baur et al., 2021). Therefore, research often focuses on understanding the image context better by adding spatial latent dimensions (Baur et al., 2018), multi-resolution (Baur et al., 2020b), or skip connections together with dropout (Baur et al., 2020a), or a denoising task as regularization (Kascenas et al., 2021). Similarly, modifications to VAEs aim to enforce the use of spatial context by spatial erasing (Zimmerer et al., 2019) or utilizing 3D information (Bengs et al., 2021; Behrendt et al., 2022). Other approaches propose restoration methods (Chen et al., 2020), uncertainty estimation (Sato et al., 2019), adversarial autoencoders (Chen and Konukoglu, 2018) or the use of encoder activation maps (Silva-Rodríguez et al., 2022). Also, vector-quantized VAEs (VQ-VAEs) have been proposed, where the spatial information of the discrete representations are processed as sequences by transformers (Pinaya et al., 2022b). As an alternative to AE-based architectures, generative adversarial networks (GANs) have been applied to the problem of UAD (Schlegl et al., 2019). However, while the adversarial game between generator and discriminator networks shows to produce sharp reconstructions (Baur et al., 2018), the unstable training nature of GANs makes their application very challenging. Furthermore, GANs suffer from mode collapse and often fail to preserve anatomical coherence (Baur et al., 2021). To alleviate this, inpainting approaches have been proposed that use the generator to inpaint erased patches during training (Nguyen et al., 2021). Lately, DDPMs have shown to be a promising approach for the task of UAD in brain MRI as they have scalable and stable training properties while generating sharp images of high quality (Wolleb et al., 2022; Wyatt et al., 2022; Sanchez et al., 2022; Pinaya et al., 2022a). However, these approaches often leverage supervision from labelled unhealthy data (Wolleb et al., 2022; Sanchez et al., 2022) and aim to estimate the entire brain anatomy at once, which remains a challenge.

3. Method

We apply the diffusion process of DDPMs in a patch-wise fashion, meaning that given the input image $\boldsymbol{x} \in \mathbb{R}^{C,W,H}$ with C channels, width W and height H, we add noise to a patch $\boldsymbol{p_k} \in \mathbb{R}^{C,h,w}$ with $h < H, w < W$ and $k = [1, ..., K]$. Subsequently, we remove the noise of the patch region, i.e. reconstruct the patch to achieve a local estimate of the brain anatomy. Hereby, our motivation is a better understanding of image context by denoising image patches based on their surroundings. Furthermore, we hypothesize that this would also lead to better anatomical coherence in the overall reconstruction of individual brains. As at test time anomalies can appear anywhere in the brain, we need to add and remove noise to the whole brain anatomy with our patch-wise approach. Therefore, we use a sliding window approach where we subsequently add noise to and reconstruct noise from individual patches at positions that are evenly spaced across the image. Having covered the entire input image, we stitch all individual patch reconstructions into one image. This strategy allows estimating each local region in the input by using the spatial context of its surroundings which is assumed to be particularly helpful if the patch covers an anomaly. A depiction of our approach is shown in Figure 1.

3.1. DDPMs

In DDPMs, first, the image structure is gradually destroyed by noise and subsequently, the reverse denoising process is learned. During the forward process, adding noise $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ to $\boldsymbol{x}_0$ follows a predefined schedule $\beta_1, ..., \beta_T$:

$$\boldsymbol{x}_t \sim q(\boldsymbol{x}_t|\boldsymbol{x}_0) = \mathcal{N}(\sqrt{\bar{\alpha}_t}\boldsymbol{x}_0, (1-\bar{\alpha}_t)\mathbf{I}), \text{with } \bar{\alpha}_t = \prod_{s=0}^{t}(1-\beta_t). \tag{1}$$

The time step t is sampled from $t \sim Uniform(1, ..., T)$ and controls how much noise is added to $\boldsymbol{x}_0$. For $t = T$ the image is replaced by pure Gaussian noise $\boldsymbol{x}_t = \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and for $t = 0$, $\boldsymbol{x}_t$ becomes $\boldsymbol{x}_0$.
In the backward process, the goal is to reverse the forward process and to recover $\boldsymbol{x}_0$.

$$\boldsymbol{x}_0 \sim p_\theta(\boldsymbol{x}_t) \prod_{t=1}^{T} p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t), \text{with } p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t) = \mathcal{N}(\boldsymbol{\mu}_\theta(\boldsymbol{x}_t, t), \boldsymbol{\Sigma}_\theta(\boldsymbol{x}_t, t)). \tag{2}$$

$\boldsymbol{\mu}_\theta$ and $\boldsymbol{\Sigma}_\theta$ are estimated by a neural network with parameters θ. Following (Ho et al., 2020), we use an Unet (Ronneberger et al., 2015) for this task and keep $\boldsymbol{\Sigma}_\theta(\boldsymbol{x}_t, t) = \frac{1-\alpha_{t-1}}{1-\alpha_t}\beta_t\mathbf{I}$ fixed. To calculate a tractable loss function, the variational lower bound is derived as

$$\mathcal{L}_{VLB} = -log(p_\theta(\boldsymbol{x}_0)) + D_{KL}(q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)||p_\theta(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)). \tag{3}$$

By applying reformulations, Bayes rule and by conditioning $q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)$ on $\boldsymbol{x}_0$, minimizing $\mathcal{L}_{VLB}$ can be approximated by the simpler loss derivation $\mathcal{L}_{simple} = ||\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, t)||^2$. Hence, DDPMs can be trained by predicting the noise $\boldsymbol{\epsilon}$ that is added to $\boldsymbol{x}_0$ to achieve $\boldsymbol{x}_t$ during the forward step. In this work, we utilize the $l1$-error and change the objective to directly estimate $\boldsymbol{x}_0^{rec} \sim p_\theta(\boldsymbol{x}_0|\boldsymbol{x}_t, t)$

$$\mathcal{L}_{rec} = |\boldsymbol{x}_0 - p_\theta(\boldsymbol{x}_0|\boldsymbol{x}_t, t)| = |\boldsymbol{x}_0 - \boldsymbol{x}_0^{rec}|. \tag{4}$$

For sampling images with DDPMs at test time, typically step-wise denoising is applied for all time steps starting from $t = T$. As this comes at the cost of long sampling times, in this work we directly estimate $\boldsymbol{x}_0^{rec} \sim p_\theta(\boldsymbol{x}_0|\boldsymbol{x}_t, t)$ at a fixed time step t_{test}. This simplification is possible since we do not aim to generate new images from noise but are interested in reconstructing a given image.

3.2. Patched DDPMs

As aforementioned, patched DDPMs (pDDPM) apply the forward and backward process in a patched fashion. During training, we sample the patches either at random positions or from a fixed grid defined as follows. We partition $\boldsymbol{x}$ into K patch regions that are evenly spaced across $\boldsymbol{x}$. The number of possible patch regions in $\boldsymbol{x}$ is derived as $K = \lceil \frac{W-w}{w} \rceil + \lceil \frac{H-h}{h} \rceil + 2$, where $\lceil . \rceil$ denotes the ceiling operation. From this grid, we uniformly sample an index k.
During the forward step of the diffusion process, i.e., the noising step we sample the noised image $\boldsymbol{x}_t$ only at the given patch position $\boldsymbol{p_k}$. Consider $\boldsymbol{M_p} \in \mathbb{R}^{C,H,W}$ a binary mask where the pixels that overlap with $\boldsymbol{p_k}$ are set to one and pixels that do not overlap with $\boldsymbol{p_k}$ are set to zero. We obtain the partly noised image as

$$\tilde{\boldsymbol{x}}_{\boldsymbol{t}} = \boldsymbol{x}_t \odot \boldsymbol{M_p} + \boldsymbol{x}_0 \odot \neg \boldsymbol{M_p} \tag{5}$$

where $\odot$ denotes element-wise multiplication. In the backward process, $\tilde{\boldsymbol{x}}_{\boldsymbol{t}}$ is fed to the denoising network to estimate the given noise area. The denoised image is derived as $\tilde{\boldsymbol{x}}_0^{rec} \sim p_\theta(\boldsymbol{x}_0|\tilde{\boldsymbol{x}}_{\boldsymbol{t}}, t)$. To train the patch-wise denoising task, we optionally use an objective function $\mathcal{L}_p$ adapted from $\mathcal{L}_{rec}$. We calculate $\mathcal{L}_p$ based on the noised region within $\boldsymbol{p_k}$, ignoring the surrounding area

$$\mathcal{L}_p = |(\boldsymbol{x}_0 - \tilde{\boldsymbol{x}}_0^{rec}) \odot \boldsymbol{M_p}|. \tag{6}$$

During Evaluation, for every $k \in [0, ..., K]$, we subsequently perform the diffusion process based on the patch $\boldsymbol{p_k}$. After the reconstruction of all patch regions, we use the reconstructed patches $[\boldsymbol{p}_0^{rec}, ..., \boldsymbol{p}_K^{rec}]$ and stitch them with respect to their original position in the input image to retain the full reconstruction of $\boldsymbol{x}_0$. In the case of overlapping patches, we average the overlapping regions of the reconstructed patches.

4. Experimental setup

4.1. Data

We use the publicly available IXI data set as healthy reference distribution for training. The IXI data set consists of 560 pairs of T1 and T2-weighted brain MRI scans, acquired in three different hospital sites. From the training data, we use 158 samples for testing and partition the remaining data set into 5 folds of 358 training samples and 44 validation samples for cross-validation and stratify the sampling by the age of the patients.

For evaluation, we utilize two publicly available data sets, namely the Multimodal Brain Tumor Segmentation Challenge 2021 (BraTS21) data set (Baid et al.; Bakas et al., 2017; Menze et al., 2014), and the multiple sclerosis data set from the University Hospital of Ljubljana (MSLUB) (Lesjak et al., 2018).

The BraTS21 data set consists of 1251 brain MRI scans of four different weightings (T1, T1-CE, T2, FLAIR). We split the data set into an unhealthy validation set of 100 samples and an unhealthy test set of 1151 samples. The MSLUB data set consists of brain MRI scans of 30 patients with multiple sclerosis (MS). For each patient T1, T2, and FLAIR-weighted scans are available. We split the data into an unhealthy validation set of 10 samples and an unhealthy test set of 20 samples. For both evaluation data sets, expert annotations in form of pixel-wise segmentation maps are available.

Across our experiments, we utilize T2-weighted images from all data sets. To align all MRI scans we register the brain scans to the SRI24-Atlas (Rohlfing et al., 2010) by affine transformations. Next, we apply skull stripping with HD-BET (Isensee et al., 2019). Note that these steps are already applied to the BraTS21 data set by default. Subsequently, we remove black borders, leading to a fixed resolution of $[192 \times 192 \times 160]$ voxels. Lastly, we perform a bias field correction. To save computational resources, we reduce the volume resolution by a factor of two resulting in $[96 \times 96 \times 80]$ voxels and remove 15 top and bottom slices parallel to the transverse plane.

4.2. Implementation Details

We evaluate our proposed method *pDDPM*, against multiple established baselines for UAD in brain MRI. These include *AE*, *VAE* (Baur et al., 2021), their sequential extension *SVAE*

(Behrendt et al., 2022), and denoising AEs *DAE* (Kascenas et al., 2021). We also compare simple thresholding *Thresh* (Meissen et al., 2022), and the GAN-based *f-AnoGAN* (Schlegl et al., 2019). Additionally, we chose *DDPM* (Wyatt et al., 2022) as a counterpart to our proposed method. We implement all baselines based on their original publications with the following individual adaptations that have been shown to improve training stability and performance. For *VAE* and *SVAE*, we set the value of β_{VAE} to 0.001. For *f-AnoGAN*, we set the latent size to 128 and the learning rate to $1e-04$.
For *DDPM* and *pDDPM*, we utilize structured simplex noise, rather than Gaussian noise, as it is known to better capture the natural frequency distribution of MRI images (Wyatt et al., 2022). For training, we uniformly sample $t \in [1, T]$ with $T = 1000$, and at test time, we choose a fixed value of $t_{test} = \frac{T}{2} = 500$. We also choose a linear schedule for β_t, which ranges from 10^{-4} to 0.02 and use an Unet similar to (Dhariwal and Nichol, 2021) as a denoising network. For each channel dimension $C_f \in [128, 128, 256]$, the Unet consists of a stack of 3 residual layers and downsampling convolutions. This structure is mirrored in the upsampling path with transposed convolutions. Skip connections connect the layers at each resolution. In each residual block, groupnorm is used for normalization and SiLU (Elfwing et al., 2018) acts as activation function before convolution. For time step conditioning, the time step is first encoded using a sinusoidal position embedding and then projected to a vector that matches the channel dimension. This is added to the feature representation using scale-shift-norm (Perez et al., 2018) in each residual block. Unless specified otherwise, all models are trained for a maximum of 1600 epochs, and the best model checkpoint, as determined by performance on the healthy validation set, is used for testing. We process the volumes in a slice-wise fashion, randomly sampling slices during training and iterating over all slices to reconstruct the full volume at test time. The models were trained on NVIDIA V100 GPUs (32GB) using Adam as the optimizer, a learning rate of $1e-5$, and a batch size of 32. The code for this work is available at `https://github.com/withheld`.

4.3. Post-Processing and Anomaly Scoring

During training, all models aim to minimize the $l1$ error between the input and its reconstruction. At test time, we use the reconstruction error as a pixel-wise anomaly score $\mathbf{\Delta}_{AS} = |\boldsymbol{x}_0 - \boldsymbol{x}_0^{rec}|$, where high values indicate larger reconstruction errors and vice versa. Given the hypothesis that the models will fail to reconstruct unhealthy brain anatomy, we assume that anomalies are located at regions of high reconstruction errors. We apply several post-processing steps that are commonly used in the literature (Baur et al., 2021; Zimmerer et al., 2019). Before binarizing $\mathbf{\Delta}_{AS}$, we use a median filter with kernel size $K_M = 5$ to smooth $\mathbf{\Delta}_{AS}$ and perform brain mask eroding for 3 iterations. Having binarized $\mathbf{\Delta}_{AS}$, we apply a connected component analysis, removing segments with less than 7 voxels. To achieve a threshold for binarizing $\mathbf{\Delta}_{AS}$, we perform a greedy search based on the unhealthy validation set where the threshold is determined by iteratively calculating Dice scores for different thresholds. The best threshold is then used to calculate the average Dice score on the unhealthy test set (DICE). Furthermore, we report the average Area Under Precision-Recall Curve (AUPRC) and report the mean absolute reconstruction error ($l1$) of the test split from our healthy IXI data set. For statistic significance tests, we use a permutation test with $k = 10000$ rounds of permutations and a significance level of $\alpha = 5\%$.

Table 1: Comparison of the evaluated models with the best results highlighted in bold. *fixed sampling* denotes that patch positions are sampled from a fixed grid, in contrast to *random sampling*, where patch positions are randomly sampled. $\mathcal{L}_p$ denotes calculating the reconstruction loss only on the patch region whereas $\mathcal{L}_{rec}$ denotes calculating the reconstruction loss for the whole image. For all metrics, mean ± standard deviation across the different folds are reported.

	BraTS21		MSLUB		IXI
Model	**DICE** [%]	**AUPRC** [%]	**DICE** [%]	**AUPRC** [%]	$l1$ (10^{-3})
Thresh (Meissen et al., 2022)	19.69	20.27	6.21	4.23	145.12
AE (Baur et al., 2021)	32.87±1.25	31.07±1.75	7.10±0.68	5.58±0.26	30.55±0.27
VAE (Baur et al., 2021)	31.11±1.50	28.80±1.92	6.89±0.09	5.00±0.40	31.28±0.71
SVAE (Behrendt et al., 2022)	33.32±0.14	33.14±0.20	5.76±0.44	5.04±0.13	28.08±0.02
DAE (Kascenas et al., 2021)	37.05±1.42	44.99±1.72	3.56±0.91	5.35±0.45	**10.12±0.26**
f-AnoGAN (Schlegl et al., 2019)	24.16±2.94	22.05±3.05	4.18±1.18	4.01±0.90	45.30±2.98
DDPM (Wyatt et al., 2022)	40.67±1.21	49.78±1.02	6.42±1.60	7.44±0.52	13.46±0.65
pDDPM + random sampling + $\mathcal{L}_{rec}$	44.47±2.34	48.84±2.71	9.41±0.96	9.13±1.13	14.08±0.77
pDDPM + fixed sampling + $\mathcal{L}_{rec}$	47.81±1.15	52.38±1.17	**10.47±1.27**	**10.58±0.85**	12.12±0.76
pDDPM + fixed sampling + $\mathcal{L}_p$	**49.00±0.84**	**54.07±1.06**	10.35±0.69	9.79±0.4	11.05±0.15

5. Results

Unless stated otherwise, for *pDDPM*, we use patch dimensions of $h = w = \frac{H}{2} = \frac{W}{2} = 48$. The comparison of our pDDPM with the baseline models is shown in Table 1. Like *DAE*, the *DDPM* shows relatively high performance on the BraTS21 data set, but its performance on the MSLUB data set is moderate. In contrast, our *pDDPM* outperforms all baselines on both data sets regarding DICE and AUPRC, with statistical significance for the BraTS21 data set ($p < 0.05$). Considering the reconstruction quality by means of $l1$ error on healthy data, the *DAE* shows the lowest reconstruction error, followed by *pDDPM*.
Qualitatively, we observe smaller reconstruction errors from *pDDPMs* compared to *DDPMs* for healthy brain anatomy as shown in Figure 4. Examples of reconstructions from other baseline models can be found in Appendix A. As seen in Figure 3, a patch size of 60×60 pixels results in the best performance. Additionally, there is a peak in performance when the noise level at test time is $t_{test} = 400$. A visualization of different noise levels is provided in Appendix B and ablation studies for the MSLUB data set are available in Appendix C.

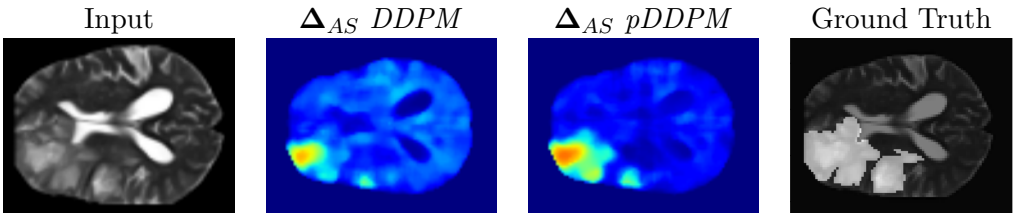

Figure 2: Vizalization of input, errormap and the ground truth for *DDPM* and *pDDPM*.

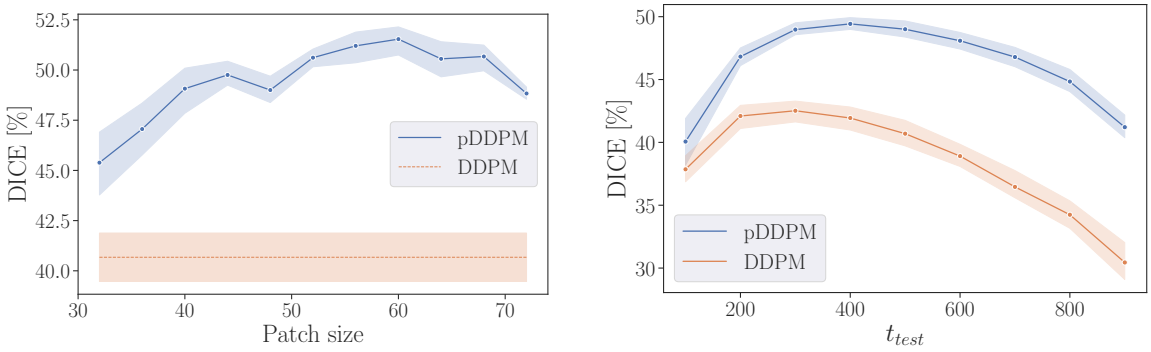

Figure 3: DICE for different patch sizes (left) and noise levels at test time t_{test} (right) for the BraTS21 data set. We report the average DICE across the 5 cross validation folds. Standard deviations are visualized as enveloping intervals.

6. Discussion & Conclusion

Our approach frames the reconstruction of healthy brain anatomy as patch-based denoising, allowing to incorporate context information about individual brain structure and appearance when estimating brain anatomy. We show that *pDDPMs* outperform both their non-patched counterparts and various baseline methods with significant differences for the BraTS21 data set ($p < 0.05$).

Our results indicate that the image context around the noised patch can be used effectively by the model to replace potential anomalies covered by noise patches with estimates of healthy anatomy. From the performance improvements resulting from selecting patches from fixed positions and minimizing $\mathcal{L}_p$ rather than $\mathcal{L}_{rec}$, we conclude that it is helpful to focus on pre-defined local patches during training. By stitching the individual patches, we achieve sharp reconstructions without the downside of reconstructing too much unhealthy anatomy. Note that this trade-off is influenced by both, the noise level t_{test} and the patch size as shown in Figure 3. Evaluating the reconstruction quality by means of $l1$ error, *DAE* shows superior results to *pDDPM*. However, *DAE* is able to reconstruct unhealthy anatomy which increases false negative predictions and thus decreases the UAD performance. We observe that accurately identifying MS lesions in T2-weighted MRI scans is challenging, and the limited number of samples makes it hard to achieve statistically significant results. However, our *pDDPMs* show promising improvements on the MSLUB data set, suggesting that it could be useful to address the challenges of detecting MS lesions.

Overall, our proposed *pDDPMs* demonstrate an improvement in the quality of reconstructed healthy brain anatomy, outperforming all compared baselines in the UAD task. In future research, we plan to condition the diffusion process by spatial context and counterfactual features such as age information.

Acknowledgments

Acknowledgments withheld.

References

Ujjwal Baid, Satyam Ghodasara, Suyash Mohan, Michel Bilello, Evan Calabrese, Errol Colak, Keyvan Farahani, Jayashree Kalpathy-Cramer, Felipe C Kitamura, Sarthak Pati, et al. The rsna-asnr-miccai brats 2021 benchmark on brain tumor segmentation and radiogenomic classification. *arXiv preprint arXiv:2107.02314*.

Spyridon Bakas, Hamed Akbari, Aristeidis Sotiras, Michel Bilello, Martin Rozycki, Justin S Kirby, John B Freymann, Keyvan Farahani, and Christos Davatzikos. Advancing the cancer genome atlas glioma mri collections with expert segmentation labels and radiomic features. *Scientific data*, 4(1):1–13, 2017.

Christoph Baur, Benedikt Wiestler, Shadi Albarqouni, and Nassir Navab. Deep autoencoding models for unsupervised anomaly segmentation in brain mr images. In *International MICCAI brainlesion workshop*, pages 161–169. Springer, 2018.

Christoph Baur, Benedikt Wiestler, Shadi Albarqouni, and Nassir Navab. Bayesian skip-autoencoders for unsupervised hyperintense anomaly detection in high resolution brain mri. In *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, pages 1905–1909. IEEE, 2020a.

Christoph Baur, Benedikt Wiestler, Shadi Albarqouni, and Nassir Navab. Scale-space autoencoders for unsupervised anomaly segmentation in brain mri. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 552–561. Springer, 2020b.

Christoph Baur, Stefan Denner, Benedikt Wiestler, Nassir Navab, and Shadi Albarqouni. Autoencoders for unsupervised anomaly segmentation in brain mr images: a comparative study. *Medical Image Analysis*, page 101952, 2021.

Finn Behrendt, Marcel Bengs, Debayan Bhattacharya, Julia Krüger, Roland Opfer, and Alexander Schlaefer. Capturing inter-slice dependencies of 3d brain mri-scans for unsupervised anomaly detection. In *Medical Imaging with Deep Learning*, 2022.

Marcel Bengs, Finn Behrendt, Julia Krüger, Roland Opfer, and Alexander Schlaefer. Three-dimensional deep learning with spatial erasing for unsupervised anomaly segmentation in brain mri. *International journal of computer assisted radiology and surgery*, 16(9): 1413–1423, 2021.

Xiaoran Chen and Ender Konukoglu. Unsupervised Detection of Lesions in Brain MRI using Constrained Adversarial Auto-encoders. In *International Conference on Medical Imaging with Deep Learning (MIDL)*, Proceedings of Machine Learning Research. PMLR, 2018.

Xiaoran Chen, Suhang You, Kerem Can Tezcan, and Ender Konukoglu. Unsupervised lesion detection via image restoration with a normative prior. *Medical image analysis*, 64:101713, 2020.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34:8780–8794, 2021.

Stefan Elfwing, Eiji Uchibe, and Kenji Doya. Sigmoid-weighted linear units for neural network function approximation in reinforcement learning. *Neural Networks*, 107:3–11, 2018.

Randall J. Ellis, Ryan M. Sander, and Alfonso Limon. Twelve key challenges in medical machine learning and solutions. *Intelligence-Based Medicine*, 6:100068, 2022. ISSN 2666-5212.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

Fabian Isensee, Marianne Schell, Irada Pflueger, Gianluca Brugnara, David Bonekamp, Ulf Neuberger, Antje Wick, Heinz-Peter Schlemmer, Sabine Heiland, Wolfgang Wick, et al. Automated brain extraction of multisequence mri using artificial neural networks. *Human brain mapping*, 40(17):4952–4964, 2019.

Justin M. Johnson and Taghi M. Khoshgoftaar. Survey on deep learning with class imbalance. *Journal of Big Data*, 6(1):1–54, 2019.

Davood Karimi, Haoran Dou, Simon K Warfield, and Ali Gholipour. Deep learning with noisy labels: Exploring techniques and remedies in medical image analysis. *Medical Image Analysis*, 65:101759, 2020.

Antanas Kascenas, Nicolas Pugeault, and Alison Q O'Neil. Denoising autoencoders for unsupervised anomaly detection in brain mri. In *International Conference on Medical Imaging with Deep Learning (MIDL)*, Proceedings of Machine Learning Research. PMLR, 2021.

Kensaku Kawamoto, Caitlin A Houlihan, E Andrew Balas, and David F Lobach. Improving clinical practice using clinical decision support systems: a systematic review of trials to identify features critical to success. *Bmj*, 330(7494):765, 2005.

Žiga Lesjak, Alfiia Galimzianova, Aleš Koren, Matej Lukin, Franjo Pernuš, Boštjan Likar, and Žiga Špiclin. A novel public mr image dataset of multiple sclerosis patients with lesion segmentations based on multi-rater consensus. *Neuroinformatics*, 16(1):51–63, 2018.

Felix Meissen, Georgios Kaissis, and Daniel Rueckert. Challenging current semi-supervised anomaly segmentation methods for brain mri. In *International MICCAI brainlesion workshop*, pages 63–74. Springer, 2022.

Bjoern H Menze, Andras Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahani, Justin Kirby, Yuliya Burren, Nicole Porz, Johannes Slotboom, Roland Wiest, et al. The multimodal brain tumor image segmentation benchmark (brats). *IEEE transactions on medical imaging*, 34(10):1993–2024, 2014.

Bao Nguyen, Adam Feldman, Sarath Bethapudi, Andrew Jennings, and Chris G Willcocks. Unsupervised region-based anomaly detection in brain mri with adversarial image inpainting. In *2021 IEEE 18th international symposium on biomedical imaging (ISBI)*, pages 1127–1131. IEEE, 2021.

Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

Walter HL Pinaya, Mark S Graham, Robert Gray, Pedro F Da Costa, Petru-Daniel Tudosiu, Paul Wright, Yee H Mah, Andrew D MacKinnon, James T Teo, Rolf Jager, et al. Fast unsupervised brain anomaly detection and segmentation with diffusion models. *arXiv preprint arXiv:2206.03461*, 2022a.

Walter HL Pinaya, Petru-Daniel Tudosiu, Robert Gray, Geraint Rees, Parashkev Nachev, Sebastien Ourselin, and M Jorge Cardoso. Unsupervised brain imaging 3d anomaly detection and segmentation with transformers. *Medical Image Analysis*, 79:102475, 2022b.

Torsten Rohlfing, Natalie M Zahr, Edith V Sullivan, and Adolf Pfefferbaum. The sri24 multichannel atlas of normal adult human brain structure. *Human brain mapping*, 31(5): 798–819, 2010.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

Pedro Sanchez, Antanas Kascenas, Xiao Liu, Alison Q O'Neil, and Sotirios A Tsaftaris. What is healthy? generative counterfactual diffusion for lesion localization. In *MICCAI Workshop on Deep Generative Models*, pages 34–44. Springer, 2022.

Kazuki Sato, Kenta Hama, Takashi Matsubara, and Kuniaki Uehara. Predictable uncertainty-aware unsupervised deep anomaly segmentation. In *2019 International Joint Conference on Neural Networks (IJCNN)*, pages 1–7, 2019. doi: 10.1109/IJCNN.2019.8852144.

Thomas Schlegl, Philipp Seeböck, Sebastian M. Waldstein, Georg Langs, and Ursula Schmidt-Erfurth. f-AnoGAN: Fast unsupervised anomaly detection with generative adversarial networks. *Medical image analysis*, 54:30–44, 2019.

Dinggang Shen, Guorong Wu, and Heung-Il Suk. Deep learning in medical image analysis. *Annual review of biomedical engineering*, 19:221, 2017.

Julio Silva-Rodríguez, Valery Naranjo, and Jose Dolz. Constrained unsupervised anomaly segmentation. *Medical Image Analysis*, 80:102526, 2022.

Julia Wolleb, Florentin Bieder, Robin Sandkühler, and Philippe C Cattin. Diffusion models for medical anomaly detection. *arXiv preprint arXiv:2203.04306*, 2022.

Julian Wyatt, Adam Leach, Sebastian M Schmon, and Chris G Willcocks. Anoddpm: Anomaly detection with denoising diffusion probabilistic models using simplex noise. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 650–656, 2022.

David Zimmerer, Simon Kohl, Jens Petersen, Fabian Isensee, and Klaus Maier-Hein. Context-encoding variational autoencoder for unsupervised anomaly detection. In *International Conference on Medical Imaging with Deep Learning–Extended Abstract Track*, 2019.

Appendix A. Exemplary reconstructions for all Baselines

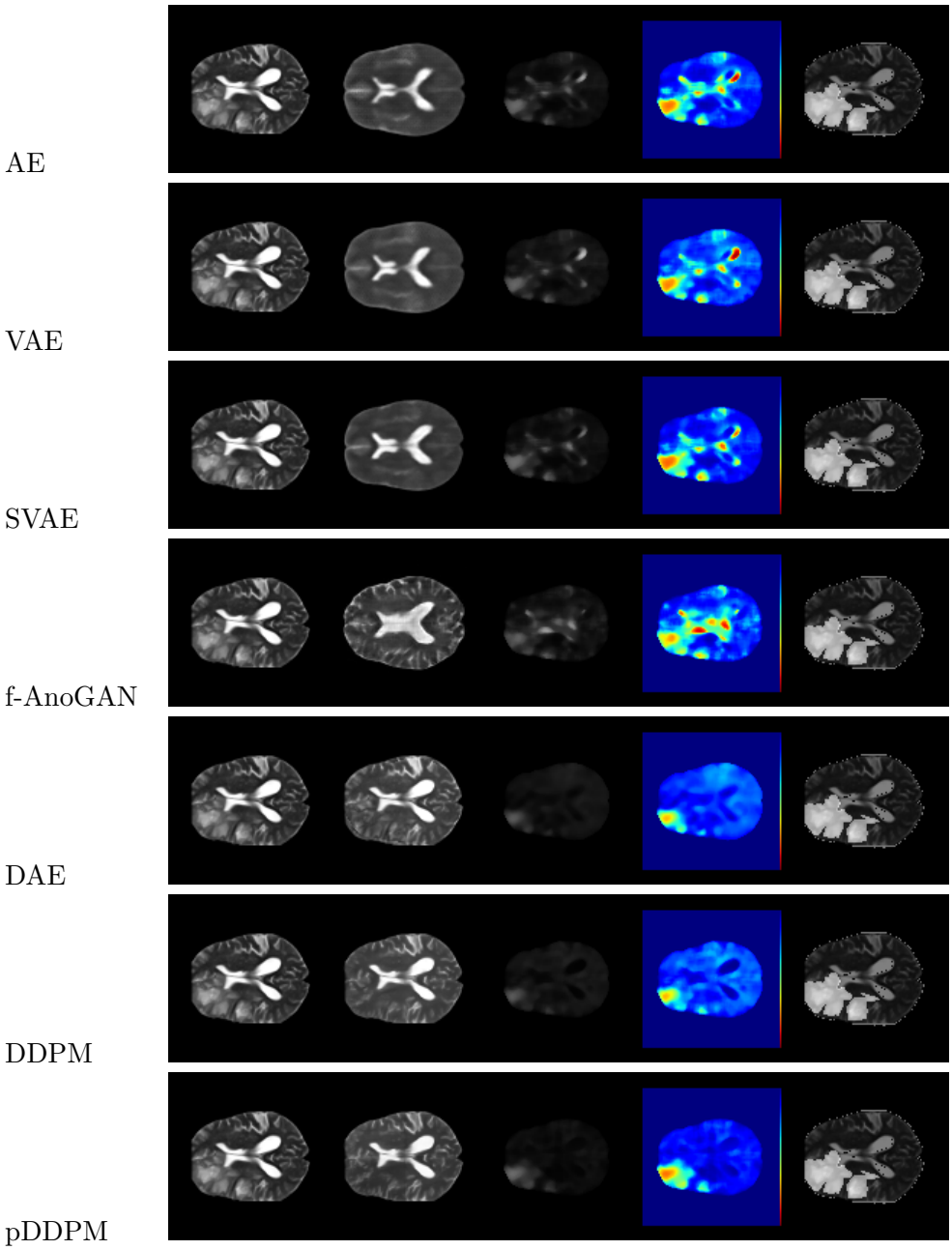

Figure 4: Qualitative evaluation of reconstructions from different models. From top to bottom: AE, VAE, SVAE, f-AnoGAN, DAE, DDPM and pDDPM are presented. From left to right, input, reconstruction, errormap, a heatmap of the errormap and the ground truth annotation is shown

Appendix B. Visualization of different noise levels

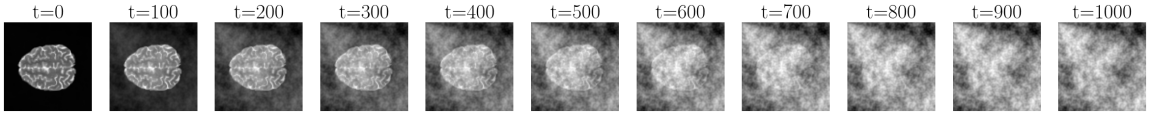

Figure 5: Training image from the IXI data set pertubed by simplex noise for different time steps $t = 0, 100, ..., 1000$

Appendix C. Ablation Studies for MSLUB

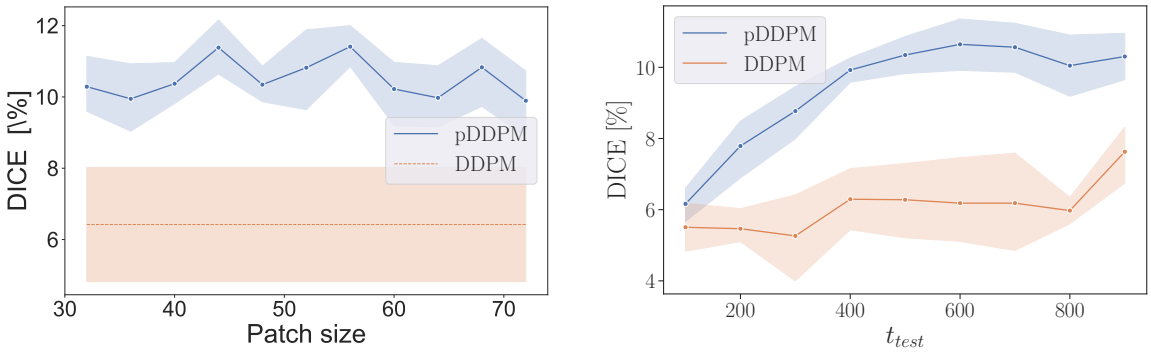

Figure 6: DICE for different patch sizes (left) and noise levels at test time t_{test} (right) for the MSLUB data set. We report the average DICE across the 5 cross validation folds. Standard deviations are visualized as enveloping intervals.