Image-conditioned Diffusion Models for Medical Anomaly Detection

Matthew Baugh[1], Hadrien Reynaud[1], Sergio Naval Marimont[2], Sarah Cechnicka[1], Johanna P. Müller[3], Giacomo Tarroni[2], and Bernhard Kainz[1,3]

[1] Imperial College London, UK
matthew.baugh17@imperial.ac.uk
[2] City, University of London, UK
[3] Friedrich–Alexander University Erlangen–Nürnberg, DE

Abstract. Generating pseudo-healthy reconstructions of images is an effective way to detect anomalies, as identifying the differences between the reconstruction and the original can localise arbitrary anomalies whilst also providing interpretability for an observer by displaying what the image 'should' look like. All existing reconstruction-based methods have a common shortcoming; they assume that models trained on purely normal data are incapable of reproducing pathologies yet also able to fully maintain healthy tissue. These implicit assumptions often fail, with models either not recovering normal regions or reproducing both the normal and abnormal features. We rectify this issue using image-conditioned diffusion models. Our model takes the input image as conditioning and is explicitly trained to correct synthetic anomalies introduced into healthy images, ensuring that it removes anomalies at test time. This conditioning allows the model to attend to the entire image without any loss of information, enabling it to replicate healthy regions with high fidelity. We evaluate our method across four datasets and define a new state-of-the-art performance for residual-based anomaly detection. Code is available at `https://github.com/matt-baugh/img-cond-diffusion-model-ad` .

Keywords: Anomaly detection · Diffusion model · Self-supervised.

1 Introduction

Supervised machine learning's requirement for large quantities of labelled training data is a barrier, as for it to be effective the training data must comprehensively cover all pathologies that the model could encounter in clinical practice. This is not practical, as there is an ever-growing number of rare or out-of-distribution conditions for which there is insufficient data available. This problem is even more pertinent in pathology segmentation, as although imaging data may be available, the cost of clinicians annotating samples is often prohibitive. Unsupervised anomaly detection offers a solution, aiming to model the distribution of healthy data itself, which enables the identification of any anomalous deviation without the need for manual labels.

The predominant paradigm for unsupervised anomaly detection in medical imaging is to train models to reconstruct images from a normative distribution [12,8,30,11], as taking the reconstruction error at test time provides a natural way to localise anomalies. The resulting pseudo-healthy images can also provide interpretability about the abnormality's features. However, these methods rely on two core assumptions; that models trained on only healthy data are **(A1)** unable to reconstruct anomalous features which were not in their training data, and **(A2)** capable of correctly reconstructing images from the normative distribution. Recently these assumptions have come into question, as such models are often able to reconstruct anomalies unseen during training [37,6], as well as predicting false-positives in healthy regions [5].

One of the most prominent solutions is restoration-based models, which edit the image to increase its likelihood and produce a pseudo-healthy version of it [6,12]. Adding this aspect aims to enforce **(A1)**, that anomalies are not preserved during the reconstruction process. Another recent paradigm is self-supervised anomaly detection, which involves using a proxy task to train a model to identify synthetic anomalies [32,19,4]. By framing the proxy task as identifying an anomaly within otherwise healthy data, models trained with this technique are less sensitive to the natural variation of healthy images, which has resulted in methods of this paradigm winning every iteration of the MICCAI Medical Out-of-Distribution Analysis (MOOD) Challenge (2020-2023)[36].

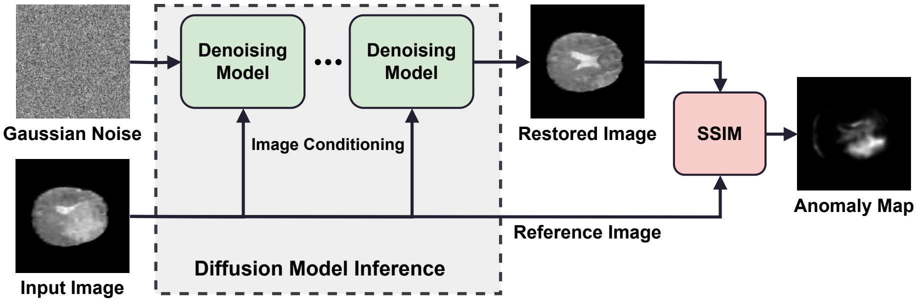

Fig. 1: We condition the diffusion model on the input image to produce a restored pseudo-healthy version of it, and then compare that with the original using the Structural Similarity Index Measure to localise anomalies.

Contributions. Our method uses diffusion models to produce high-fidelity, pseudo-healthy restorations of test images by conditioning on the input image at every timestep of the reverse diffusion process (Fig. 1). To train the diffusion model we take our healthy training data as the target and use the self-supervision tasks developed by [3] to generate anomalous images for conditioning. We guarantee that assumption **(A1)** holds by explicitly training the model to restore the image. Other state-of-the-art methods jeopardise **(A2)** as they lose information by noising or masking the input image. On the other hand, our use of image conditioning solidifies it as the model can attend to the entire image thus maintaining healthy regions. By combining these aspects we achieve beyond state-of-the-art performance, which we demonstrate on a challenging benchmark across multiple image modalities.

Related Work. Reconstruction-based approaches involve training a generative model or autoencoder on a dataset of only normal images. At inference time the model reconstructs the test sample and calculates the residual map between the original and reconstructed image. Originally most methods were based on either Variational Autoencoders [37] or Generative Adversarial Networks [30], although the success of diffusion models has led to them being increasingly favoured [12,20]. Regardless of architecture, by using the reconstruction error to identify anomalies, all of these methods depend on assumptions **(A1)** and **(A2)**. However, the subtle distinction between healthy and unhealthy image features in medical imaging means that autoencoders trained to reproduce normative data can also reconstruct anomalies [6]. [37] found that models with excessive capacity can maintain anomalous regions, requiring tuning with samples from the same distribution as the test set to enable regularisation through limited capacity. Predictions from reconstruction-based methods often predict false positives around complex structures or edges that they fail to accurately replicate [26]. Post-processing can remove these errors, but this requires prior knowledge about the type of anomalies expected at test time [5].

Masking regions of the input guarantees that the model can not use the information in those regions to reconstruct any anomalies that may be present [15,34]. However, this loss of information also means that the fine-grain details of healthy structures are also lost, resulting in false positives. Restoration-based methods have been proposed that aim to explicitly edit an image or its corresponding latent representation to increase its likelihood with respect to the normative dataset distribution [8], which has been shown to consistently produce residual errors that are better at localising anomalies when compared with their reconstruction-based counterparts [5]. [7] combined masking with image restoration, using a two-stage process to identify potentially anomalous regions and then inpaint those to produce a pseudo-healthy restoration.

A fundamental issue with reconstruction-based approaches stems from relying on the comparison of pixel-intensity values. This means that, even when a model successfully corrects anomalies, the difference between the original and reconstructed images will remain minimal if the anomaly exhibits low contrast compared to healthy tissue [5]. [25] found that performing the reconstruction in feature space resulted in a better measure of deviation when comparing the input and output of the model. They also proposed using the Structural Similarity Index Measure (SSIM) [35] in both training and inference, as it accounts for structural and contrast differences between the original and reconstructed image. Using SSIM for inference has since been shown to consistently improve anomaly localisation for residual-based methods [18].

Self-supervised anomaly detection methods have adopted a completely different paradigm, using a proxy task to directly train a model to identify anomalies within a normal image. An important work in this area was Poisson Image Interpolation (PII) [32] which trained a model to identify patches of other samples blended into a target image. By directly identifying anomalies without using residual error it does not suffer from the issues concerning pathologies of low contrast or extreme texture. [28] better aligned this synthetic task with the imaging modality whilst incorporating a probabilistic element to mirror the effect of multiple different annotators. The most recent approaches use a combination of multiple synthetic tasks to perform cross-validation and prevent model overfitting to the synthetic training tasks [3]. Denoising autoencoders [16] combines a proxy task (removing coarse additive noise) with residual-based anomaly detection to train a model to directly restore an image, however recently this has been shown to generalise poorly to other datasets as the scale of the noise was chosen to match the scale of the anomalies it was evaluated on [18]. A not yet fully available approach [24] suggests a patch-blending proxy task as a diffusion process so that the learnt score function better generalises to medical anomalies.

2 Method

Our image diffusion model operates within the discrete-time Denoising Diffusion Probabilistic Model (DDPM) framework [14], leveraging the ϵ-prediction objective. This approach models the forward diffusion process as a Markov chain,

$$p(\mathbf{x}_{1:T}|\mathbf{x}_0) = \prod_{t=1}^{T} p(\mathbf{x}_t|\mathbf{x}_{t-1}), \tag{1}$$

where each diffusion step introduces Gaussian noise into the image, defined by $p(\mathbf{x}_t|\mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t\mathbf{I})$. Here, β_t is a pre-defined variance schedule controlling the noise level at each step.

In the reverse process, a specially designed neural network aims to reconstruct the original image by estimating the noise ϵ that was added at each diffusion step. This estimation, $\epsilon_\theta(\mathbf{x}_t, t)$, allows for the recovery of the denoised image through the update equation:

$$\mathbf{x}_{t-1} = \frac{1}{\sqrt{1-\beta_t}} \left(\mathbf{x}_t - \frac{\beta_t}{\sqrt{1-\alpha_t^2}} \epsilon_\theta(\mathbf{x}_t, t) \right), \tag{2}$$

where α_t and β_t represent time-dependent scaling factors, $\mathbf{x}_t$ denotes the noisy image at time t, and ϵ_θ is the model predicting the noise. This reverse mechanism effectively guides the model in reconstructing the signal from its noised state, step by step, until it recovers the original image $\mathbf{x}_0$. Introducing image-conditioning into this changes the noise estimation function to include a conditioning image c as $\epsilon_\theta(\mathbf{x}_t, t, c)$.

In our method we take c to be the result of applying a synthetic anomaly task to the original image $\mathbf{x}_0$, resulting in $\tilde{\mathbf{x}}_0$. By providing it as conditioning during the reverse process (Fig. 2) we encourage the model to intelligently use

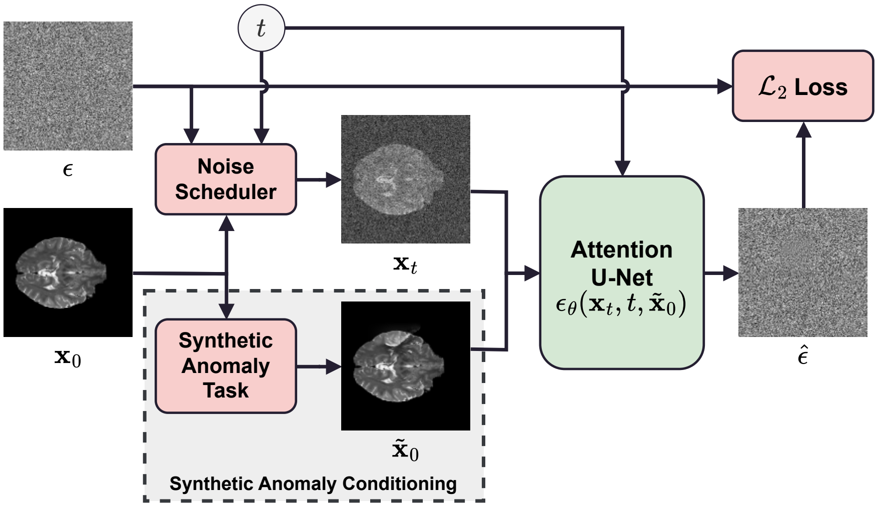

Fig. 2: Training process for our model. At each time step the model is trained to predict the residual which denoises $\mathbf{x}_t$, using the corrupted image $\tilde{\mathbf{x}}_0$ as guidance.

$\tilde{\mathbf{x}}_0$ to restore $\mathbf{x}_t$; by identifying the healthy regions of $\tilde{\mathbf{x}}_0$ it can use them to accurately denoise those areas of $\mathbf{x}_t$, while for the corrupted areas it must use a combination of $\mathbf{x}_t$ and $\tilde{\mathbf{x}}_0$ to predict how to restore $\mathbf{x}_t$ to be fully healthy. The image conditioning is implemented by concatenating $\tilde{\mathbf{x}}_0$ to $\mathbf{x}_t$ as additional image channels. We generate our anomalous images using the tasks from the extension of [3], consisting of 5 tasks covering image blending, deformation and intensity addition. By using a wide range of subtly integrated tasks we ensure the model learns to have a detailed understanding of the appearance of healthy images, covering not just the general intensity distributions of the images but also the intricacies of their structures. To avoid the model learning a prior that every image requires some amount of correction we drop the synthetic task in 20% of training samples, *i.e.*, $\tilde{\mathbf{x}}_0 = \mathbf{x}_0$.

For inference, we identify anomalies by using the test image as conditioning throughout the entire reverse diffusion process, mapping a Gaussian noise sample to a pseudo-healthy version of the test image. To produce an anomaly map we measure the SSIM [35] between the restoration and original, allowing for fair comparison with [18] which uses SSIM in all reported results. In line with [18] the only post-processing we apply is to zero predictions in the background of the MR images. We follow the cross-validation structure of [3], using three tasks to train each model while the remaining two are used to monitor performance, which means ten models are trained per dataset. Similarly, we ensemble our models by averaging their pixel-wise predictions during inference, and take the mean pixel-wise prediction as the sample-level prediction.

Implementation. Our PyTorch [29] code is publically available and integrated into the existing code of the unsupervised pathology detection benchmark [18].

Each model is trained over two A100 GPUs with a batch size of 32. This memory footprint could fit on a single A100 but we opted to split it to speed up training.

3 Experiments

Data. We evaluate in line with the benchmark setting in [18]. Each dataset features anomalies with varying characteristics, offering different perspectives on the adaptability and potential limitations of our suggested approach. All brain MRI models are trained using samples from the Cambridge Centre for Ageing and Neuroscience dataset (CamCAN) [33] which consists of 653 scans of healthy adult patients. Each patient has a T1-weighted and T2-weighted scan which we use to train separate models. To test both T2 and T1 MRI models we use scans containing tumours from the Multimodel Brain Tumor Segmentation (BraTS) Challenge 2020 [27,1,2]. The pathologies in the T2-weighted images are the easiest to identify as they appear as large, hyper-intense lesions. Conversely, identifying anomalies in T1-weighted BraTS scans is more challenging due to the lower contrast between lesions and healthy tissue, despite the anomalies being the same size. To further scrutinise T1 MRI models, we evaluate them on the Anatomical Tracings of Lesions After Stroke (ATLAS) dataset [22] which includes 655 MR scans of stroke patients with small, hypo-intense lesions. We also evaluate our method on retinal fundus images from the DDR dataset [21] which consists of 6243 images of healthy individuals and 745 images taken from patients with Diabetic Retinopathy. The lesions in DDR are both very small and of a similar image intensity distribution to healthy tissue, making them very difficult to detect. Here we follow the same train-test split as in [18], taking 5510 samples for training and validation while a disjoint set of 745 healthy and 745 unhealthy samples are used for testing. We use the same preprocessing pipeline as [18], *i.e.*, our models take 128×128 images as input.

To assess our model's performance at an image level we use the area under the receiver operating characteristic curve (AUC) and image-wise average precision ($\mathrm{AP_i}$) which is the area under the precision-recall curve. Most of the datasets are reasonably balanced so the $\mathrm{AP_i}$ and AUC values are similar, except for the ATLAS dataset where the small size of the lesions means that only 30% of axial slices contain a lesion. For pixel-level evaluation, we use pixel-wise average precision ($\mathrm{AP_p}$) and the maximum Sørensen-Dice index over all possible thresholds ($\lceil \mathrm{Dsc} \rceil$), which both take into account that there is a large data imbalance at a pixel level which favours the normal class.

Results. We achieve state-of-the-art performance across three of the benchmark datasets (Tab. 1). Our models are notably better at localising anomalies regardless of their size. We consistently outperform the best existing method by +0.24 $\mathrm{AP_p}$ on the large anomalies of BraTS-T2 and +0.09 $\mathrm{AP_p}$ on the small, scattered anomalies of DDR. Our restorations also successfully maintain the healthy regions of test samples (Fig. 3), which contributes to our high pixel-wise performance. Other methods using local synthetic anomalies (PII [32] and CutPaste [19]) struggle across all datasets, highlighting that our richer task of

Table 1: Quantative detection and localisation results, comparing against image-reconstruction (**IR**), feature-modeling (**FS**), attention-based (**AB**), and self-supervised anomaly detection (**S-S**) methods. Best scores are bold, second best are underlined, excluding the average result of each fold Ours$_{avg.}$, the best performing individual model Ours$_{max.}$ and standard deviation of folds Ours$_{std.}$.

		BraTS-T2				BraTS-T1				ATLAS				DDR			
		image-level		pixel-level		image-level		pixel-level		image-level		pixel-level		image-level		pixel-level	
	Method	AP_i	AUC	AP_p	$\lceil$Dsc$\rceil$	AP_i	AUC	AP_p	$\lceil$Dsc$\rceil$	AP_i	AUC	AP_p	$\lceil$Dsc$\rceil$	AP_i	AUC	AP_p	$\lceil$Dsc$\rceil$
	Random	0.48	0.50	0.06	0.11	0.48	0.50	0.06	0.11	0.30	0.50	0.02	0.03	0.50	0.50	0.004	0.01
IR	VAE [17]	0.68	0.73	0.28	0.33	0.64	0.70	0.13	0.19	0.57	0.76	0.11	0.20	0.48	0.48	0.02	0.05
	r-VAE [8]	0.75	0.77	0.36	0.40	0.70	0.76	0.13	0.19	**0.60**	**0.78**	0.09	0.17	0.52	0.50	0.03	0.09
	f-AnoGAN [30]	0.56	0.61	0.15	0.21	0.48	0.53	0.06	0.12	0.26	0.46	0.02	0.06	0.44	0.45	0.01	0.01
	H-TAE-S [11]	0.68	0.70	0.21	0.12	0.54	0.57	0.06	0.12	0.29	0.49	0.01	0.03	0.51	0.51	0.01	0.01
FM	FAE [25]	0.87	0.87	0.51	0.52	**0.86**	**0.85**	**0.42**	**0.45**	0.50	0.73	0.08	0.18	0.64	0.63	0.07	0.15
	PaDiM [9]	0.66	0.68	0.34	0.38	0.60	0.65	0.21	0.28	0.34	0.56	0.05	0.13	0.55	0.55	0.02	0.07
	CFLOW-AD [13]	0.71	0.72	0.31	0.35	0.65	0.69	0.16	0.24	0.40	0.62	0.04	0.10	0.51	0.51	0.03	0.08
	RD [10]	0.85	0.85	0.47	0.50	0.81	0.83	0.36	0.42	0.55	0.77	0.11	0.22	0.66	0.64	0.10	0.19
AB	ExpVAE [23]	0.63	0.66	0.12	0.18	0.56	0.56	0.07	0.13	0.37	0.57	0.01	0.03	0.53	0.54	0.004	0.01
	AMCons [31]	0.78	0.78	0.35	0.40	0.61	0.64	0.05	0.12	0.32	0.53	0.01	0.03	0.49	0.49	0.004	0.01
S-S	PII [32]	0.57	0.62	0.13	0.22	0.54	0.64	0.13	0.22	0.37	0.60	0.03	0.07	0.62	0.63	0.01	0.01
	DAE [16]	0.81	0.80	0.47	0.49	0.70	0.74	0.13	0.20	0.44	0.65	0.05	0.13	0.54	0.55	0.01	0.03
	CutPaste [19]	0.59	0.63	0.22	0.26	0.61	0.65	0.07	0.13	0.37	0.58	0.03	0.06	0.64	0.60	0.02	0.06
	Ours$_{ens.}$	**0.89**	**0.88**	**0.74**	**0.69**	0.75	0.77	0.30	0.36	0.55	0.75	**0.25**	**0.34**	**0.73**	**0.73**	**0.19**	**0.27**
	Ours$_{avg.}$	0.88	0.87	0.69	0.66	0.72	0.74	0.26	0.34	0.54	0.74	0.18	0.28	0.71	0.69	0.14	0.24
	Ours$_{max.}$	0.89	0.89	0.77	0.71	0.75	0.76	0.34	0.39	0.60	0.78	0.31	0.38	0.74	0.72	0.17	0.27
	Ours$_{std.}$	0.01	0.01	0.05	0.04	0.03	0.02	0.05	0.03	0.04	0.03	0.06	0.05	0.01	0.02	0.02	0.02

restoring synthetic anomalies gives more consistent performance than training models to purely detect synthetic anomalies. Comparing against DAE [16], which also combines self-supervised and reconstruction-based techniques, DAE fails to effectively localise the anomalies of ATLAS and DDR (0.05 and 0.01 AP_p vs. 0.02 and 0.004 AP_p for a random classifier) while our method sets a new state-of-the-art performance (0.25 and 0.19 AP_p).

Discussion. The performance on BraTS-T1 can be largely explained by the low contrast between anomalies and healthy tissue. This is a limitation of all residual-based methods [18]. Nevertheless, we are the best-performing method in that category (0.30 AP_p vs. joint second-best method DAE's 0.13). Fig. 3 illustrates that while our models effectively correct sample anatomy, the similar pixel intensities of anomalies and healthy tissue limit the clarity of anomaly maps produced by comparing restored images to the originals. Feature-modelling methods [25,10] perform better in this case, as the difference in feature space is more pronounced. However these methods do not produce a pseudo-healthy restoration, hence lack an element of interpretability. Our method's image-level performance on ATLAS slightly trails the leading approaches, with a narrow margin of 0.05 AP_i / 0.03 AUC from the top method. This discrepancy could stem from the domain shift between the CamCAN training dataset and ATLAS, attributed to differences in population samples and scanner types. Consequently, all ATLAS images might be perceived as anomalous, diluting the distinction in

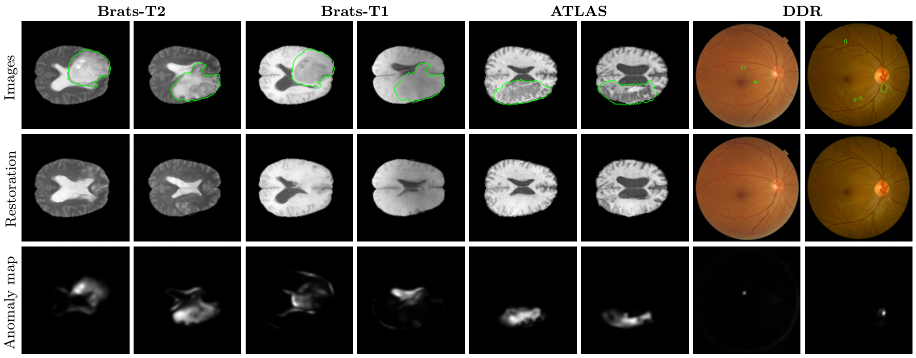

Fig. 3: Qualitative examples of anomaly restorations along with their corresponding anomaly maps. For BraTS datasets we take the same samples for each modality to highlight the difference in anomaly contrast between the modalities. The ground truth segmentation is illustrated as a green contour on the input image, except for DDR where we use an ellipse as the anomalies are so small (1-4 pixels) that the boundary obscures the anomaly.

anomaly scores triggered by actual anomalies. Interestingly, the base VAE and r-VAE models, despite lagging behind the state-of-the-art in other datasets, show commendable performance on ATLAS. This suggests that factors other than the model's capability may be constraining the effectiveness across all methods.

To compare our final ensemble (Ours$_{ens.}$ in Tab. 1) with the performance of each individual model, the ensemble consistently outperforms the average metrics of each fold (Ours$_{avg.}$). This shows that different models are learning to correct different aspects of the images, otherwise the performance gain would be negligible when ensembling the predictions. The maximum performance of any individual model (Ours$_{max.}$) is marginally higher than the ensemble, however we cannot know which models perform best apriori in a realistic setting. This further shows that our ensemble Ours$_{ens.}$ approximates the maximum potential performance across various scenarios. Ours$_{std.}$ shows the standard deviation of the performance of each model, where the image-level performance is more consistent than the pixel-level, which is expected as the pixel-level task is notably more challenging.

4 Conclusion

We train image-conditioned diffusion models to correct synthetic anomalies and demonstrate that these models can produce realistic, pseudo-healthy restorations of images containing real-world, unseen pathologies. By comparing the restored image with the original we can localise these anomalies with state-of-the-art performance, outperforming all methods across three datasets covering various modalities and anomaly types. By training our models to explicitly restore anomalous regions and using image conditioning to avoid loss of information, we guarantee the core assumptions of reconstruction-based anomaly detection to ensure our models maintain healthy regions whilst successfully replacing anomalies.

Acknowledgments: We received support by an EPSRC DTP award (MB), the ERC project MIA-NORMAL 101083647, the State of Bavaria (HTA), and DFG 512819079. HPC resources were provided by NHR@FAU of FAU Erlangen-Nürnberg under the NHR projects b180dc and b143dc. NHR funding is provided by federal and Bavarian state authorities. NHR@FAU hardware is partially funded by DFG - 440719683.

References

1. Bakas, S., Akbari, H., Sotiras, A., Bilello, M., Rozycki, M., et al.: Advancing the cancer genome atlas glioma mri collections with expert segmentation labels and radiomic features. Scientific data **4**(1), 1–13 (2017)
2. Bakas, S., Reyes, M., Jakab, A., Bauer, S., Rempfler, M., et al.: Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the BRATS challenge. arXiv:1811.02629 (2018)
3. Baugh, M., Tan, J., Müller, J.P., Dombrowski, M., Batten, J., Kainz, B.: Many tasks make light work: Learning to localise medical anomalies from multiple synthetic tasks. In: MICCAI'23. pp. 162–172. Springer (2023)
4. Baugh, M., Tan, J., Vlontzos, A., Müller, J.P., Kainz, B.: nnOOD: A framework for benchmarking self-supervised anomaly localisation methods. In: Unsure@MICCAI'22. pp. 103–112. Springer (2022)
5. Baur, C., Denner, S., Wiestler, B., Navab, N., Albarqouni, S.: Autoencoders for unsupervised anomaly segmentation in brain mr images: a comparative study. Medical Image Analysis **69**, 101952 (2021)
6. Bercea, C.I., Rueckert, D., Schnabel, J.A.: What do AEs learn? challenging common assumptions in unsupervised anomaly detection. In: MICCAI'23. pp. 304–314. Springer (2023)
7. Bercea, C.I., Wiestler, B., Rueckert, D., Schnabel, J.A.: Reversing the abnormal: Pseudo-healthy generative networks for anomaly detection. In: MICCAI'23. pp. 293–303. Springer (2023)
8. Chen, X., You, S., Tezcan, K.C., Konukoglu, E.: Unsupervised lesion detection via image restoration with a normative prior. Medical image analysis **64**, 101713 (2020)
9. Defard, T., Setkov, A., Loesch, A., Audigier, R.: Padim: a patch distribution modeling framework for anomaly detection and localization. In: ICPR. pp. 475–489. Springer (2021)
10. Deng, H., Li, X.: Anomaly detection via reverse distillation from one-class embedding. In: CVPR'22. pp. 9737–9746 (2022)
11. Ghorbel, A., Aldahdooh, A., Albarqouni, S., Hamidouche, W.: Transformer based models for unsupervised anomaly segmentation in brain MR images. arXiv:2207.02059 (2022)
12. Graham, M.S., Pinaya, W.H.L., Wright, P., Tudosiu, P.D., Mah, Y.H., et al.: Unsupervised 3d out-of-distribution detection with latent diffusion models. In: MICCAI'22. pp. 446–456. Springer (2023)

13. Gudovskiy, D., Ishizaka, S., Kozuka, K.: Cflow-ad: Real-time unsupervised anomaly detection with localization via conditional normalizing flows. In: WACV. pp. 98–107 (2022)
14. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. NeurIPS'20 **33**, 6840–6851 (2020)
15. Iqbal, H., Khalid, U., Chen, C., Hua, J.: Unsupervised anomaly detection in medical images using masked diffusion model. In: MLMI@MICCAI'23. pp. 372–381. Springer (2023)
16. Kascenas, A., Pugeault, N., O'Neil, A.Q.: Denoising autoencoders for unsupervised anomaly detection in brain mri. In: MICCAI'22. pp. 653–664. PMLR (2022)
17. Kingma, D.P., Welling, M.: Auto-encoding variational Bayes. In: ICLR'14, (2014)
18. Lagogiannis, I., Meissen, F., Kaissis, G., Rueckert, D.: Unsupervised pathology detection: A deep dive into the state of the art. IEEE Trans. Med. Imaging. **43**(1), 241–252 (2024)
19. Li, C.L., Sohn, K., Yoon, J., Pfister, T.: Cutpaste: Self-supervised learning for anomaly detection and localization. In: CVPR'21. pp. 9664–9674 (2021)
20. Li, J., Cao, H., Wang, J., Liu, F., Dou, Q., Chen, G., Heng, P.A.: Fast non-markovian diffusion model for weakly supervised anomaly detection in brain MR images. In: MICCAI'23. pp. 579–589. Springer (2023)
21. Li, T., Gao, Y., Wang, K., Guo, S., Liu, H., Kang, H.: Diagnostic assessment of deep learning algorithms for diabetic retinopathy screening. Information Sciences **501**, 511–522 (2019)
22. Liew, S.L., Anglin, J.M., Banks, N.W., Sondag, M., Ito, K.L., et al.: A large, open source dataset of stroke anatomical brain images and manual lesion segmentations. Scientific data **5**(1), 1–11 (2018)
23. Liu, W., Li, R., Zheng, M., Karanam, S., Wu, Z., , et al.: Towards visually explaining variational autoencoders. In: CVPR'20. pp. 8642–8651 (2020)
24. Marimont, S.N., Baugh, M., Siomos, V., Tzelepis, C., Kainz, B., Tarroni, G.: Disyre: Diffusion-inspired synthetic restoration for unsupervised anomaly detection. arXiv preprint arXiv:2311.15453 (2023)
25. Meissen, F., Paetzold, J., Kaissis, G., Rueckert, D.: Unsupervised anomaly localization with structural feature-autoencoders. arXiv:2208.10992 (2022)
26. Meissen, F., Wiestler, B., Kaissis, G., Rueckert, D.: On the pitfalls of using the residual error as anomaly score. In: MIDL'22. pp. 914–928. PMLR (2022)
27. Menze, B.H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., et al.: The multimodal brain tumor image segmentation benchmark (brats). IEEE Trans. Med. Imaging. **34**(10), 1993–2024 (2015)
28. P. Müller, J., Baugh, M., Tan, J., Dombrowski, M., Kainz, B.: Confidence-aware and self-supervised image anomaly localisation. In: Unsure@MICCAI'23. pp. 177–187. Springer (2023)
29. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., et al.: Pytorch: An imperative style, high-performance deep learning library. NeurIPS'19 **32** (2019)
30. Schlegl, T., Seeböck, P., Waldstein, S.M., Langs, G., Schmidt-Erfurth, U.: f-AnoGAN: Fast unsupervised anomaly detection with generative adversarial networks. Medical image analysis **54**, 30–44 (2019)
31. Silva-Rodríguez, J., Naranjo, V., Dolz, J.: Constrained unsupervised anomaly segmentation. Medical Image Analysis **80**, 102526 (2022)
32. Tan, J., Hou, B., Day, T., Simpson, J., Rueckert, D., Kainz, B.: Detecting outliers with poisson image interpolation. In: MICCAI'21. pp. 581–591 (2021)

33. Taylor, J.R., Williams, N., Cusack, R., Auer, T., Shafto, M.A., et al.: The Cambridge Centre for Ageing and Neuroscience (Cam-CAN) data repository: Structural and functional MRI, MEG, and cognitive data from a cross-sectional adult lifespan sample. NeuroImage **144**, 262–269 (2017)
34. Tian, Y., Pang, G., Liu, Y., Wang, C., Chen, Y., et al.: Unsupervised anomaly detection in medical images with a memory-augmented multi-level cross-attentional masked autoencoder. In: MLMI@MICCAI'23. pp. 11–21. Springer (2023)
35. 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)
36. Zimmerer, D., Full, P.M., Isensee, F., Jäger, P., Adler, T., et al.: Mood 2020: A public benchmark for out-of-distribution detection and localization on medical images. IEEE Trans. Med. Imaging. **41**(10), 2728–2738 (2022)
37. Zimmerer, D., Isensee, F., Petersen, J., Kohl, S., Maier-Hein, K.: Unsupervised anomaly localization using variational auto-encoders. In: MICCAI'19. pp. 289–297. Springer (2019)