Diffusion Models with Implicit Guidance for Medical Anomaly Detection

Cosmin I. Bercea[1,2], Benedikt Wiestler[1,3], Daniel Rueckert[1,3,5], and Julia A. Schnabel[1,2,4]

[1] Technical University of Munich, Munich, Germany
[2] Helmholtz AI and Helmholtz Center Munch, Munich, Germany
[3] Klinikum Rechts der Isar, Munich, Germany
[4] Kings College London, London, UK
[5] Imperial College London, London, UK
cosmin.bercea@tum.de

Abstract. Diffusion models have advanced unsupervised anomaly detection by improving the transformation of pathological images into pseudo-healthy equivalents. Nonetheless, standard approaches may compromise critical information during pathology removal, leading to restorations that do not align with unaffected regions in the original scans. Such discrepancies can inadvertently increase false positive rates and reduce specificity, complicating radiological evaluations. This paper introduces Temporal Harmonization for Optimal Restoration (*THOR*), which refines the reverse diffusion process by integrating implicit guidance through intermediate masks. *THOR* aims to preserve the integrity of healthy tissue details in reconstructed images, ensuring fidelity to the original scan in areas unaffected by pathology. Comparative evaluations reveal that *THOR* surpasses existing diffusion-based methods in retaining detail and precision in image restoration and detecting and segmenting anomalies in brain MRIs and wrist X-rays. Code: `https://github.com/compai-lab/2024-miccai-bercea-thor.git`.

Keywords: Generative AI · OoD · Brain MRI · Wrist X-ray

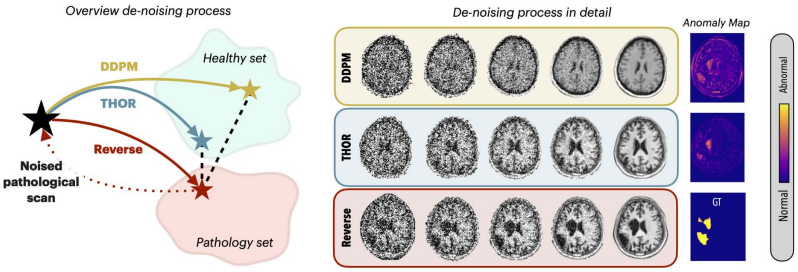

Fig. 1: Denoising diffusion probabilistic models (DDPMs) trend towards a generalized healthy reference, diverging from nuanced details in the original image. *THOR* aims for restoration close to the original, within healthy limits.

1 Introduction

Robust and accurate anomaly detection is vital for early diagnosis and effective treatment, especially in the face of rare and diverse pathologies. The complexity and variability inherent in medical conditions present substantial challenges to conventional diagnostic methods anchored in supervised learning [8,16]. These methods depend heavily on extensive, annotated datasets, which are difficult to obtain for rare conditions, limiting the scope and flexibility of diagnostic tools. In response, unsupervised learning has emerged as a viable alternative, capable of detecting anomalies across a broad spectrum without the need for explicit labels [6,17,11,12]. Among unsupervised techniques, denoising diffusion probabilistic models (DDPMs) [7] have shown substantial promise in enhancing the precision and efficiency of anomaly detection. By adding and subsequently removing noise, DDPMs transform pathological inputs into pseudo-healthy outputs, demonstrating impressive generative potential. Nonetheless, this noise-dependent process can result in significant loss of information, leading restored images to deviate from their original state, including in regions unaffected by pathology [4]. Such deviations risk increasing false positives and decreasing specificity, further complicating the diagnostic process.

To overcome the limitations inherent in DDPMs, more sophisticated models have been developed. AnoDDPM proposes to use Simplex noise, which allows the use of lower noise levels [14]. Conditional diffusion models blend the capabilities of autoencoders with diffusion techniques to incorporating semantic information such as tissue intensity into the de-noising process [3]. Patch-based DDPMs (pDDPMs) extend these advancements by applying the diffusion process to localized patches of the image, using adjacent areas as contextual anchors in a sliding-window technique [2]. AutoDDPMs build upon this foundation with a unique approach that involves masking, stitching, and re-sampling, utilizing dual de-noising processes at different levels of noise to seamlessly integrate context into the reconstructions [4]. While these innovations represent substantial progress, they also introduce complexities. The task of determining an optimal patch size that can adapt to the multiple scales of diseases is challenging due to the diversity of pathological presentations. Additionally, the complexity of orchestrating dual de-noising processes across different noise levels requires precise calibration. These challenges could potentially limit their practicability.

Diffusion models enhanced with classifier guidance use weakly supervised classifiers for anomaly detection, leveraging gradients to refine the identification of anomalous regions [13]. However, the effectiveness of this approach depends on the accuracy of classifiers, potentially limiting its capability to detect diseases independently by biasing it towards known pathologies.

In this work, we introduce *THOR* (Temporal Harmonization for Optimal Restoration), a novel approach designed to enhance unsupervised anomaly detection in medical imaging, as illustrated in Figure 1. *THOR* incorporates implicit guidance into diffusion models through the use of intermediate masks, aiming to preserve the original image context while achieving accurate anomaly detection and segmentation. Our key contributions are as follows:

- The development of *THOR*, leveraging implicit guidance within diffusion models to facilitate optimal image restorations and improve the accuracy of anomaly segmentation.
- The application of *THOR* to two challenging medical datasets, where it demonstrates its capability in accurately segmenting stroke lesions on brain scans and localizing pathology in pediatric wrist X-rays, thereby enhancing performance in essential diagnostic tasks.
- An analysis of the sensitivity of critical hyper-parameters such as different noise types and levels.

2 Background

2.1 Anomaly Detection Setup

In medical imaging anomaly detection, the objective is to detect deviations from normal anatomical structures without explicit pathological labels. We define X as the domain of all medical images, where each image $x \in X$ includes regions of both normal and abnormal tissue. The aim is to assign an anomaly score S to each pixel (or voxel), using a function $f : X \to S$.

Considering a dataset of medical images $x_{i=1}^{N}$ for training, these images are presumed to represent a healthy tissue distribution, denoted as $P(x)$. The challenge lies in accurately modeling $P(x)$. By doing so, we can project any input image into the $P(x)$ space, creating a pseudo-healthy reconstruction. If an input image has pathology, this method produces a version where pathological features are replaced with those typical of healthy tissue according to $P(x)$. This approach enables anomaly detection by contrasting the original image with its pseudo-healthy counterpart to identify deviations.

2.2 Denoising Diffusion Probabilistic Models (DDPMs)

The forward diffusion process in DDPMs [7] transforms the original data x_0 into a sequence of increasingly noisy versions $x_1, x_2, \ldots, x_T$. This process can be understood as an approximate posterior: $q(x_t|x_0) = \mathcal{N}(x_t; \sqrt{\alpha_t}x_0, (1-\alpha_t)I)$, where $\alpha_t = \prod_{s=1}^{t}(1-\beta_s)$ with β_t representing the variance schedule that dictates the noise level at each step. Using the reparameterization trick, we can sample x_t directly from x_0:

$$x_t = \sqrt{\alpha_t}x_0 + \sqrt{1-\alpha_t}\epsilon, \text{ with } \epsilon \sim \mathcal{N}(0, I). \tag{1}$$

The reverse diffusion process aims to reconstruct the original data x_0 from the noisy data x_T. This is achieved by learning a parameterized model $p_\theta(x_{t-1}|x_t)$ that approximates the reverse of the forward process:

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \sigma_\theta(x_t, t)^2 I) \tag{2}$$

The mean $\mu_\theta(x_t, t)$ and variance $\sigma_\theta(x_t, t)^2$ are learned by minimizing the variational lower bound. During sampling, the reverse process starts with a sample from the Gaussian prior $p(x_T)$ and iteratively applies $p_\theta(x_{t-1}|x_t)$ to generate samples that resemble the original data distribution $P(x)$.

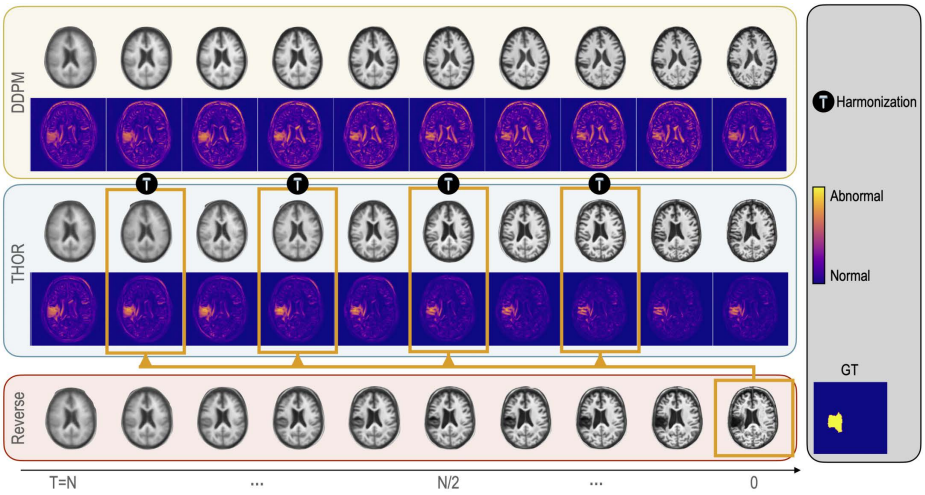

Fig. 2: The top row displays the traditional DDPM denoising sequence, progressively reducing noise to enhance image features. In contrast, the middle row features *THOR*, which adapts the reverse process using unsupervised intermediate masks for 'harmonization' (marked by orange boxes)—maintaining normal tissue integrity while reducing anomalies. The bottom rows shows the deterministic pathological reverse process, gradually revealing anomalies. The ground truth (GT) image where the anomaly is clearly delineated is provided by experts.

3 Method: THOR

THOR advances the de-noising process in DDPMs, offering guidance during inference through the application of implicit intermediate masks, without necessitating retraining. Typically, DDPMs necessitate high noise levels (T) to effectively obscure anomalies, a practice that can compromise the integrity of non-pathological tissue details. Such an approach may result in the loss of critical anatomical information, thereby elevating the potential for false positives. The innovation of *THOR* lies in its ability to guide the restoration process by strategically reintegrating healthy tissue information, a technique we refer to as "harmonization." This method starts at the same elevated noise levels but diverges by using implicit intermediate masks to inform the denoising trajectory. Such guidance aims to selectively restore the image, focusing on preserving the fidelity of non-pathological regions while reducing the anomalies. Details of our procedural approach are delineated in Figure 2.

Implicit Guidance via Intermediate Masks. Intermediate masks play an essential role in the unsupervised "harmonization" process of *THOR*. These masks critically compare the predictive reconstructions x_0^t with the actual input image x_0^{input}, highlighting discrepancies that indicate anomalies and distinguishing regions that are likely healthy. Intermediate masks m combine residual differ-

ences with the Learned Perceptual Image Patch Similarity (LPIPS) metric [15], enhancing the identification of subtle pathological changes [6]:

$$m(x, y) = |x - y| \cdot S_{\text{LPIPS}}(x, y). \tag{3}$$

To avoid incorporating anomalous regions in the denoising process, we normalize the values of m between 0 and 1 and apply morphological operations, specifically a sequence of closing followed by dilation (denoted as cd).

These intermediate masks are then utilized in the "harmonization" process to adjust the interpolation between the pseudo-healthy predictions and the actual inputs. This adjustment aims to producing reconstructions that not only closely resemble the original images but also conform to the healthy profile:

$$x_t = cd(m(x_0^t, x_0^{\text{input}})) \cdot x_0^{\text{prediction}} + (1 - cd(m(x_0^t, x_0^{\text{input}}))) \cdot x_0^{\text{input}}. \tag{4}$$

The final anomaly score, S, is calculated using the harmonic mean of the intermediate masks at the different harmonization timesteps:

$$S = n \Big/ \sum_{t \in \text{harmonization steps}} \frac{1}{m(x_0^t, x_0^{\text{input}})}, \tag{5}$$

where n is the total number of harmonization steps.

4 Experiments

4.1 Ischemic Stroke Lesion Segmentation in Brain MRI

This experiment evaluates the effectiveness of *THOR* and other recent diffusion-based anomaly detection methods in segmenting ischemic stroke lesions. Stroke represents a major cause of disability and mortality worldwide, with its early and accurate detection being paramount for effective intervention and treatment planning. The variability in stroke lesions, in terms of size, location, and affected brain tissue, adds layers of complexity to their identification in neuroimaging.

Datasets. The IXI [1] dataset include 581 healthy T1-weighted MRI scans (465 for training, 58 for validation, and 58 for testing). ATLAS 2.0 [9], comprising 655 T1-weighted MRI scans with expert-segmented lesion masks, is used primarily for testing. Out of 655 scans, 217 do not contain visible pathologies and are used to augment the healthy training set. All scans come from different patients, ensuring no overlap between the training and evaluation sets. Anomalies were stratified into small (< 71 pixels), medium, and large (≥ 570 pixels) lesions. We excluded 20 slices with hypo-intense artifacts that were not annotated to maintain data quality. For pre-processing, we selected the mid-axial slices, normalized all images to the 98th percentile, and pad and resized them to a 128×128 resolution. Lesion segmentation was quantified using the maximum Dice ($\lceil \text{Dice} \rceil$).

Table 1: **Performance on Brain MRI Stroke Segmentation.** *THOR*, our proposed method, considerably outperforms other methods (DDPM, AutoDDPM, AnoDDPM, pDDPM) across different lesion sizes, marked by the ***bold*** numbers and percentage improvements (▲ x) compared to the best baseline.

Noise	Method	Pathology ⌈*Dice*⌉ ↑ Average	Small	Medium	Large
Gauss	THOR (ours)	**20.41** ▲ 20%	**9.14** ▲ 103%	**26.34** ▲ 19%	41.26 ▼ 5%
	DDPM [7]	8.05 ▼ 61%	1.37 ▼ 85%	9.53 ▼ 64%	25.65 ▼ 38%
	AutoDDPM [4]	16.95 ▼ 17%	4.55 ▼ 50%	22.07 ▼ 16%	**43.47** ▲ 5%
Simplex	THOR (ours)	**29.74** ▲ 33%	**11.54** ▲ 44%	**39.20** ▲ 30%	**63.64** ▲ 34%
	AnoDDPM [14]	18.07 ▼ 39%	4.82 ▼ 58%	23.45 ▼ 40%	46.65 ▼ 27%
	pDDPM [2]	22.28 ▼ 25%	8.02 ▼ 31%	30.16 ▼ 23%	47.66 ▼ 25%

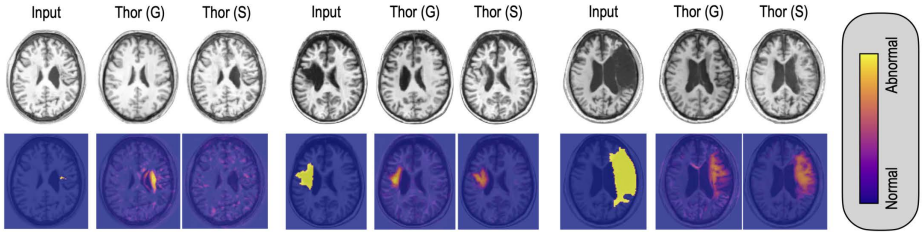

Fig. 3: Anomaly detection in brain MRI scans processed by *THOR* using Gaussian (G) and Simplex (S) noise. From left to right, the lesions increase in size, with the smallest representing a challenging case.

Results. Table 1 shows quantitative results and explores two key diffusion noise scenarios: Gaussian and Simplex. This examination is vital for assessing the performance of *THOR* in comparison with leading diffusion models. *THOR* is proficient with both types of noise, illustrating its broad applicability.

Gaussian noise is the conventional choice for DDPMs but introduces challenges in anomaly detection. Due to the partial denoising strategy employed for anomaly detection, a high noise level (here T=350) is essential to effectively conceal anomalies [14]. Yet, deploying Gaussian noise at such high iterations frequently results in false positives due to inaccuracies in restoring healthy tissue. This limitation is reflected in the diminished segmentation scores for DDPM. Conversely, our harmonization process navigates the de-noising towards more precise restorations. Consequently, *THOR* addresses the challenge of false positives and significantly refines the accuracy of anomaly segmentation, as evidenced both numerically in Table 1 and visually in Figure 2 and Figure 3.

Simplex noise provides a notable advantage with its coarse noise patterns, allowing for the de-noising process to commence at lower levels (T=250) as demonstrated in [14]. This characteristic is beneficial, preserving more of the original image context and laying a stronger groundwork for restoration. The utility of Simplex noise becomes apparent when observing the improved performance of

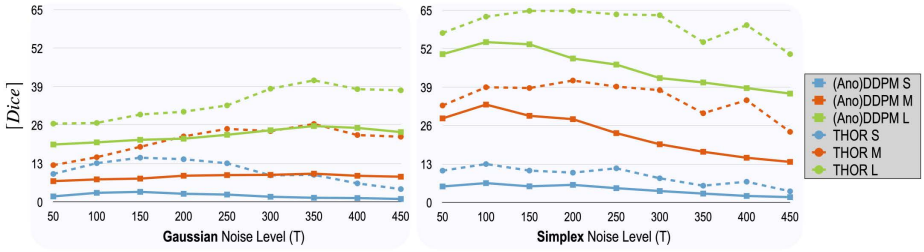

Fig. 4: **Noise Level Ablation.** *THOR* outperforms the diffusion counterparts under both Gaussian and Simplex noise types across different noise levels T.

models like AnoDDPM, which exhibit significant enhancements over the traditional DDPM. Leveraging the capability of Simplex noise, *THOR* advances the restoration process further. Its harmonization process meticulously refines the output, ensuring restorations more faithfully represent the original healthy tissue and thereby outperforming AnoDDPM and similar models (see Table 1).

Sensitivity analysis of noise levels T is shown in Figure 4. Increasing noise levels in the Gaussian setting enhances the detection of larger lesions, highlighting the role of higher noise in their effective masking. In contrast, performance abruptly declines with elevated Simplex noise levels. *THOR* excels across different noise intensities, showcasing particular robustness at higher levels. This robustness minimizes the need for finely tuned noise adjustments for specific applications or anomaly sizes, underscoring *THOR*'s adaptability and efficacy.

4.2 Anomaly Localization in Pediatric Wrist X-rays

In this section, we evaluate the localization of anomalies in pediatric wrist X-rays. Bone fractures are notably prevalent in children, with their detection through X-rays being a critical step for timely medical interventions. In this experiment, we opted for Gaussian noise due to its broader applicability. Simplex noise, although not specifically designed for brain MRI, was ineffective in wrist X-ray experiments. The DDPMs utilizing Simplex noise did not adequately address visible anomalies such as fractures or metal implants, as detailed in the Supplementary Material. This outcome is consistent with findings from other studies on brain MRI [5], where DDPMs trained with Simplex noise failed to detect anomalies outside the targeted distribution of coarse intensity-based patterns.

Dataset. We utilize the comprehensive GRAZPEDWRI-DX dataset [10], encompassing 10,643 (L1,R1) X-rays of pediatric wrist injuries from 6,091 patients. It includes a wide array of anomalies annotated with bounding boxes by certified radiologists. This includes bone anomalies (BA), foreign bodies (FB), fractures (Frac.), metal implants, periosteal reactions (PR), and soft tissue conditions

Table 2: Anomaly detection and localization results in pediatric wrist X-rays.

Noise	Method	BA		FB		Frac.		Metal		Pr.		Soft	
		Recall	F1	Recall	F1	Recall	F1	Recall	F1	Recall	F1	Recall	F1
Gauss	THOR (ours)	**83.33**	23.76	**75.00**	25.00	**75.39**	**16.46**	**99.76**	**73.76**	**76.42**	16.64	26.32	10.77
	DDPM [7]	32.22	6.35	**75.00**	29.83	28.53	5.10	86.47	39.66	52.25	9.79	23.68	8.89
	AutoDDPM [4]	63.89	**23.93**	**75.00**	**58.33**	45.56	15.84	95.89	72.05	62.29	**29.00**	**31.58**	**16.45**

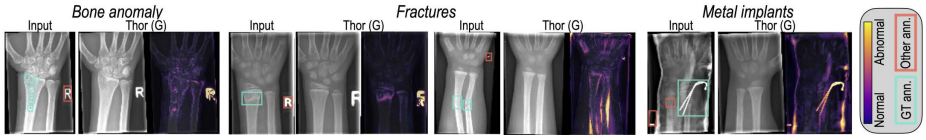

Fig. 5: Anomaly detection in pediatric wrist X-rays processed by *THOR* using Gaussian noise. False positives arise from unannotated indirect pathological changes like unnatural bone positions following fractures or the presence of casts.

(Soft). We report the recall and F1 scores as detailed in [6].

Results. Table 2 and Figure 5 summarize quantitative and qualitative results. *THOR* outperforms SOTA diffusion models, considerably improving the number of anomalies detected by up to 65% in case of fractures. The applicaiton of wrist X-ray anomaly detection poses some challenges for unsupervised methods such as the rise of false positives due to unannotated indirect pathological changes shown in Figure 5. These are correctly identified as anomalies, but not annotated by the radiologists. Furthermore, some conditions like soft tissue anomalies are subtle and remain difficult to spot on X-rays and small resolutions.

5 Conclusion

In this study, we introduced Temporal Harmonization for Optimal Restoration *(THOR)*, a novel approach that enhances the utility of diffusion models for unsupervised anomaly detection in medical imaging. Our key innovation lies in refining the reverse diffusion process by incorporating intermediate masks, which implicitly guide the generation of pseudo-healthy restorations and ensure the preservation of healthy tissue integrity. We rigorously tested *THOR* in two challenging scenarios—detecting stroke lesions in brain MRIs and identifying pediatric wrist injuries in X-rays. Our results show that *THOR* considerably outperforms state-of-the-art diffusion-based methods.

Despite these advancements, unsupervised anomaly detection still faces challenges, including a higher rate of false positives from unannotated, indirect pathological changes, and difficulties in detecting subtle or small anomalies. While our method shows notable improvements for small lesions, it also highlights the need for enhanced precision in these areas.

Future work will focus on developing clinically relevant metrics to evaluate the detection accuracy of (small) lesions, conducting comprehensive clinical validation, optimizing the computation of intermediate masks, and expanding *THOR*'s application to various pathologies and organs. These efforts aim to enhance *THOR*'s diagnostic accuracy and broaden its clinical applicability.

Acknowledgments. C.I.B. is funded via the EVUK program ("Next-generation Al for Integrated Diagnostics") of the Free State of Bavaria and partially supported by the Helmholtz Association under the joint research school 'Munich School for Data Science'.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Ixi dataset. `https://brain-development.org/ixi-dataset/`, accessed: 2023-02-15
2. Behrendt, F., Bhattacharya, D., Krüger, J., Opfer, R., Schlaefer, A.: Patched diffusion models for unsupervised anomaly detection in brain mri. In: Medical Imaging with Deep Learning (2023)
3. Behrendt, F., Bhattacharya, D., Mieling, R., Maack, L., Krüger, J., Opfer, R., Schlaefer, A.: Guided reconstruction with conditioned diffusion models for unsupervised anomaly detection in brain mris. arXiv preprint arXiv:2312.04215 (2023)
4. Bercea, C.I., Neumayr, M., Rueckert, D., Schnabel, J.A.: Mask, stitch, and re-sample: Enhancing robustness and generalizability in anomaly detection through automatic diffusion models. In: ICML 3rd Workshop on Interpretable Machine Learning in Healthcare (IMLH) (2023)
5. Bercea, C.I., Wiestler, B., Rueckert, D., Schnabel, J.A.: Reversing the abnormal: Pseudo-healthy generative networks for anomaly detection. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 293–303. Springer (2023)
6. Bercea, C.I., Wiestler, B., Rueckert, D., Schnabel, J.A.: Generalizing unsupervised anomaly detection: Towards unbiased pathology screening. In: Medical Imaging with Deep Learning. pp. 39–52. PMLR (2024)
7. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)
8. Kamnitsas, K., Ferrante, E., Parisot, S., Ledig, C., Nori, A.V., Criminisi, A., Rueckert, D., Glocker, B.: DeepMedic for brain tumor segmentation. In: Medical Image Computing and Computer Assisted Intervention BrainLes Workshop. pp. 138–149 (2016)
9. Liew, S.L., Lo, B.P., ., Miarnda R. Donnelly, e.a.: A large, curated, open-source stroke neuroimaging dataset to improve lesion segmentation algorithms. Scientific Data **9** (2022)
10. Nagy, E., Janisch, M., Hržić, F., et al.: A pediatric wrist trauma x-ray dataset (grazpedwri-dx) for machine learning. Scientific Data **9**, 222 (2022)
11. Pinaya, W.H., Tudosiu, P.D., Gray, R., Rees, G., Nachev, P., Ourselin, S., Cardoso, M.J.: Unsupervised brain imaging 3d anomaly detection and segmentation with transformers. Medical Image Analysis **79**, 102475 (2022)

12. Tan, J., Hou, B., Day, T., Simpson, J., Rueckert, D., Kainz, B.: Detecting outliers with poisson image interpolation. In: Medical Image Computing and Computer Assisted Intervention. pp. 581–591. Springer (2021)
13. Wolleb, J., Bieder, F., Sandkühler, R., Cattin, P.C.: Diffusion models for medical anomaly detection. In: Medical Image Computing and Computer Assisted Intervention. pp. 35–45. Springer (2022)
14. Wyatt, J., Leach, A., Schmon, S.M., Willcocks, C.G.: Anoddpm: Anomaly detection with denoising diffusion probabilistic models using simplex noise. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops pp. 650–656 (2022)
15. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 586–595 (2018)
16. Zhou, Y., Chia, M.A., Wagner, S.K., Ayhan, M.S., Williamson, D.J., Struyven, R.R., Liu, T., Xu, M., Lozano, M.G., Woodward-Court, P., et al.: A foundation model for generalizable disease detection from retinal images. Nature pp. 1–8 (2023)
17. Zimmerer, D., Isensee, F., Petersen, J., Kohl, S., Maier-Hein, K.: Unsupervised anomaly localization using variational auto-encoders. In: Medical Image Computing and Computer Assisted Intervention. pp. 289–297. Springer (2019)