MatchGen: Detecting Medical Abnormal Region by Generating Matched Normal Regions

Xinyu Ma[1], Jinhui Ma[1], Shiqi He[2], Xin Che[1], Hon Yiu So[3], and Lingyang Chu[1]*

[1] McMaster University
{ma209,maj26,chex5,chul9}@mcmaster.ca
[2] University of Michigan, Ann Arbor, MI 48109, USA
shiqihe@umich.edu
[3] Oakland University, Rochester, MI 48309, USA
hso@oakland.edu

Abstract. Accurate abnormal region detection in medical images is critical for early diagnosis. Unlike supervised and self-supervised methods, unsupervised methods require no annotated training data and generalize well to unseen abnormalities. Such advantages are achieved by detecting abnormal regions from the differences between an input image and a generated pseudo-normal image, which is similar to the input image but excludes abnormal regions. However, existing unsupervised methods often suffer from high false positive rate at test time due to poor pixel-level matching between the normal regions of the input image and the pseudo-normal image. To address this challenge, we propose MatchGen, a novel plug-and-play framework to enhance the detection performance of existing unsupervised methods by optimizing the pseudo-normal image at test time. This generates an *optimized pseudo-normal image* that accurately matches the normal regions of the input while maintaining a clear distinction from the abnormal regions, which significantly improves the detection performance. Extensive experiments on four real-world datasets demonstrate the outstanding effectiveness of MatchGen.

Keywords: Abnormal region detection · Test time optimization · Medical image analysis

1 Introduction

Detecting abnormal regions in medical images is crucial for early diagnosis and effective treatment. For example, detecting brain tumors in MRI scans enables timely interventions that significantly improve patient outcomes [24]. Identifying abnormalities in retinal images also plays a key role in diagnosing diabetes and diabetic retinopathy [19]. Given the importance of these real-world applications, many effective methods have been developed to detect abnormal regions

* Corresponding author: Lingyang Chu (chul9@mcmaster.ca)

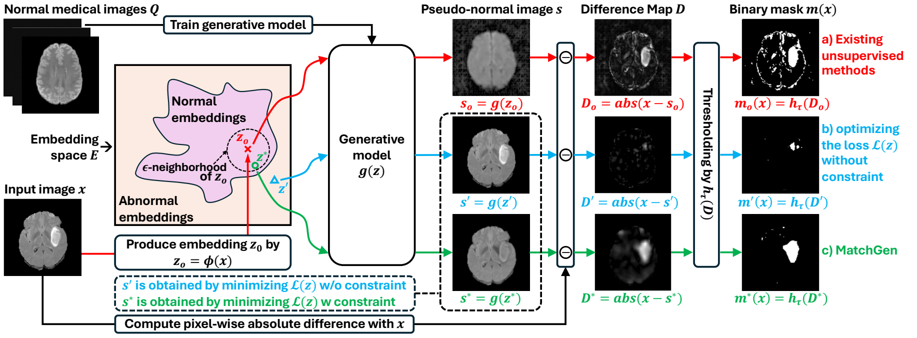
Normal medical images Q
Train generative model
Embedding space E
Normal embeddings
z_o
z^*
z'
ϵ-neighborhood of z_o
Abnormal embeddings
Input image x
Produce embedding z_0 by $z_o = \phi(x)$
Generative model $g(z)$
s' is obtained by minimizing $\mathcal{L}(z)$ w/o constraint
s^* is obtained by minimizing $\mathcal{L}(z)$ w constraint
Compute pixel-wise absolute difference with x
Pseudo-normal image s
$s_o = g(z_o)$
$s' = g(z')$
$s^* = g(z^*)$
Difference Map D
$D_o = abs(x - s_o)$
$D' = abs(x - s')$
$D^* = abs(x - s^*)$
Thresholding by $h_\tau(D)$
Binary mask $m(x)$
$m_o(x) = h_\tau(D_o)$
$m'(x) = h_\tau(D')$
$m^*(x) = h_\tau(D^*)$
a) Existing unsupervised methods
b) optimizing the loss $\mathcal{L}(z)$ without constraint
c) MatchGen

Fig. 1: An example showing: a) the process of the existing unsupervised methods (see red arrows); b) the result when overly optimizing the loss $\mathcal{L}(z)$ in Equation (1) (see blue arrows); and c) the process of MatchGen (see green arrows).

in medical images. Existing methods are broadly categorized into three groups: supervised, self-supervised, and unsupervised methods.

Supervised methods [3] require large training datasets of annotated medical images, with abnormal regions manually labeled by human experts. Obtaining such annotations is expensive and the trained models often cannot generalize well to unseen abnormal regions that differ from the training data [5].

Self-supervised methods [25,26,23,16] aim to eliminate the need for costly human annotations by training models with synthetic abnormal images. However, the discrepancy between synthetic and real abnormal images often causes poor generalization of the models on real-world medical images.

Unsupervised methods [5,32,12,33,17,29,18] address the limitations of supervised and self-supervised methods, because they do not require annotated abnormal images or synthetic abnormal regions for training. Instead, they detect abnormal regions by identifying deviations from normal regions. This is achieved by training a generative model, such as autoencoders [8,5,32], GANs [22,28], and diffusion models [6,18,30,7,29], on normal images without any annotated regions. Denote by $\phi(\cdot)$ and $g(\cdot)$ the encoder (or forward diffusion) and decoder (or reverse diffusion), respectively, of the generative model trained on normal images. The path of red arrows in Figure 1 shows the process of the existing unsupervised methods. We illustrate this in Example 1.

Example 1. Given an *input image* x, the goal is to detect its abnormal regions. First, x is encoded as an embedding $z_o = \phi(x)$, which is then decoded into a *pseudo-normal image* $s_o = g(z_o)$. Since $\phi(x)$ and $g(z)$ are only trained to generate normal images, z_o often belongs to the distribution of normal embeddings, which ensures that s_o retains normal regions of x while omitting abnormal regions [5,32]. Thus, the *difference map* $D_o = \text{abs}(x - s_o)$, which is the absolute pixel-level difference between x and s_o, assigns larger values to the abnormal pixels in x. Finally, a *binary mask* $m_o(x) = h_\tau(D_o)$ is obtained by detecting the pixels in D_o with values above a threshold τ as abnormal pixels.

Most existing unsupervised methods [5,6,32,12,33,17,29] detect abnormal regions by the process in Example 1. Their primary distinction lies in how to generate the pseudo-normal image. The *quality* of the pseudo-normal image s_o, in terms of how well it matches the normal regions and distinguishes the abnormal regions in x, is crucial to the overall detection performance.

Unfortunately, the s_o generated at test time by the existing unsupervised methods [5,6,12,18,32,29] often fails to accurately match the normal regions in x. While some methods [5,12] attempt to minimize the pixel-level difference between the generated image and the input image at training time, that is, when training $\phi(x)$ and $g(z)$ on normal images, the s_o generated at test time still fails to accurately match the normal regions in x since their pixel-level difference is not explicitly minimized at test time. [10] attempts to address this mismatch issue at test time, however it does not explicitly minimize the pixel-level difference either; instead, it uses a normative prior that is specific to models trained with the Evidence Lower BOund (ELBO), which limits its applicability to a broader class of generative models. As a result, existing methods [5,10,18,32,29] often suffer from high false positive rate, because many normal pixels are mistakenly flagged as abnormal in the difference map $D_o = \text{abs}(x - s_o)$.

To the best of our knowledge, accurately matching the normal regions in x when generating s_o at test time is a challenging task because the normal and abnormal regions in x are unknown at test time, which makes it difficult to explicitly minimize the pixel-level difference between the normal regions of s_o and x.

In this paper, we tackle the above task by designing a novel plug-and-play framework named MatchGen to enhance the detection performance of existing unsupervised methods. The key idea is to optimize the pseudo-normal image at test time, such that it retains high pixel-level similarity to the normal regions of the input image while maintaining a clear distinction from abnormalities. This enables MatchGen to effectively reduce false positive rate while ensuring high detection accuracy for abnormal regions. We make the following contributions:

First, we propose a loss function to minimize the pixel-level difference between the pseudo-normal image and the input image at test time. This improves the pixel-level matching between the pseudo-normal image and the input image, which significantly reduces the false positive rate.

Second, to prevent the pseudo-normal image from overfitting the abnormal regions in the input image, we introduce a novel constraint to restrict the minimization of the loss function when generating the pseudo-normal image. This ensures the pseudo-normal image resembles a realistic normal image that remains clearly distinguishable from the abnormal regions in the input image, which improves the detection accuracy of abnormal regions.

Last, we conducted extensive experiments on seven unsupervised methods and four real-world medical image datasets to demonstrate the effectiveness of MatchGen in enhancing the detection performance of existing unsupervised methods. We also performed a case study to visually analyze MatchGen's outstanding effectiveness.

2 Task and Methodology

2.1 Task Definition

Given the trained encoder $\phi(x)$ and decoder $g(z)$ of an existing unsupervised abnormal region detection method, the goal of our task is to enhance the detection performance by optimizing the pseudo-normal image at test time. We adopt the same unsupervised setting as the existing unsupervised methods in the literature [5,32,30], that is: 1) no training images with annotated abnormal regions are used to train the abnormal region detection model; 2) a set of normal medical images, denoted by Q, is used to train a generative model that captures the distribution of normal medical images; and 3) a small set of validation images with annotated abnormal regions is used to select hyper-parameter values.

2.2 Problem Formulation and Solution

We tackle the above task by proposing a general framework named MatchGen to boost the detection performance of existing unsupervised approaches [5,12,30,2]. The **key idea** of MatchGen is to generate an *optimized pseudo-normal image*, denoted by s^*, which is optimized at test time to accurately match the normal regions in x while maintaining a clear distinction from the abnormal regions in x. Following this idea, we formulate the problem to generate s^* as a *constrained optimization problem*:

$$\min_{z} \|x - g(z)\|_1 \quad \text{s.t. } \|z - z_o\|_2 \leq \epsilon, \tag{1}$$

where $\|\cdot\|_1$ and $\|\cdot\|_2$ are L1-norm and L2-norm, respectively, x is the input image, $g(z)$ is the decoder trained together with the encoder $\phi(x)$ on Q, z is an embedding used as the input for $g(z)$ to generate the pseudo-normal image $s = g(z)$, and $z_o = \phi(x)$ is the embedding derived from the input image x. Please refer to Figure 1 and Example 1 for the meaning of the notations.

The loss function in Equation (1), denoted by $\mathcal{L}(z) = \|x - g(z)\|_1$, is the L1-norm of the difference map $D = \text{abs}(x - s)$, where $s = g(z)$ is the pseudo-normal image.

Due to the intrinsic property of L1-norm to promote sparsity by encouraging shrinkage of small values to zero [27], minimizing $\mathcal{L}(z)$ produces a sparse difference map D, which reduces the pixel-wise difference between the normal regions of x and s. This enables s to accurately match the normal regions, which reduces the false positive rate of abnormal region detection.

However, excessively minimizing $\mathcal{L}(z)$ without any constraint will lead to poor detection performance. As shown by the blue arrow path in Figure 1, excessively minimizing $\mathcal{L}(z)$ reduces the pixel-wise difference in both the normal regions and abnormal regions, thus the pseudo-normal image s' closely resembles both the normal and abnormal regions in x. This produces a poor difference map D' that fails to effectively detect the abnormal regions in x.

The cause of the poor detection performance when excessively minimizing $\mathcal{L}(z)$ lies in the embedding space E. By empirically analyzing the embedding

Algorithm 1 MatchGen

```
Require: The $\phi(x)$ and $g(z)$ trained on $Q$, input image $x$, and thresholds $\epsilon$ and $\tau$.
Ensure: A binary mask $m(x)$ on $x$.
Compute $z_o = \phi(x)$.
Obtain $z^*$ by solving the problem in Equation (1).
Compute optimized pseudo-normal image $s^* = g(z^*)$.
Compute difference map $D^* = \text{abs}(x - s^*)$.
Compute binary mask $m^*(x) = h_\tau(D^*)$ by thresholding $D^*$ with $\tau$.
Return $m(x) \leftarrow m^*(x)$.
```

z' that excessively minimizes $\mathcal{L}(z)$, we discovered that z' is often distant from z_o. This means z' is likely outside of the distribution of the normal embeddings generated by the encoder $\phi(x)$. Thus, even though $\phi(x)$ and $g(z)$ are only trained on normal images, $g(z')$ still generates an abnormal image from z'.

To avoid excessively minimizing $\mathcal{L}(z)$, we introduce the constraint $\|z - z_o\|_2 \leq \epsilon$ in Equation (1), which restricts the feasible search space of z^* to the ϵ-neighborhood of $z_o = \phi(x)$. As shown by the path of green arrows in Figure 1, keeping ϵ small heuristically ensures that the ϵ-neighborhood centered at z_o is in the distribution of the normal embeddings generated by $\phi(x)$.

This prevents $s^* = g(z^*)$ from drifting far from normal images, which helps ensure s^* to clearly distinguish itself from the abnormal regions in x. Consequently, the difference map $D^* = \text{abs}(x - s^*)$ can effectively detect the abnormal regions in x.

In summary, the optimized pseudo-normal image s^* generated by MatchGen at test time accurately matches the normal regions in x while maintaining a clear distinction from the abnormal regions. This explains why MatchGen can achieve a low false positive rate for normal regions while maintaining a high detection accuracy for abnormal regions.

We solve the problem in Equation (1) by the penalty method [20], which converts Equation (1) to the following *unconstrained optimization problem*:

$$\min_z \|x - g(z)\|_1 + \lambda \max\left(\|z - z_o\|_2 - \epsilon, 0\right), \tag{2}$$

where the penalty parameter λ is initialized to be $\lambda = 1$, and it is gradually increased in the optimization process until $\max\left(\|z - z_o\|_2 - \epsilon, 0\right) = 0$. This ensures the final solution z^* is a feasible solution to Equation (1). Algorithm 1 concludes the whole process of MatchGen.

3 Experiments

Datasets. We use four public medical images datasets named BraTS2021 [4,1], BTCV [15,1], RESC [11] and IDRiD [19,1]. For each dataset, we construct a training dataset $\mathcal{D}_{train}$ of normal images, a validation dataset $\mathcal{D}_{val}$ of abnormal images, and a testing dataset $\mathcal{D}_{test}$ of abnormal images. The abnormal regions

Table 1: The essential information of each dataset.

Datasets	**Modality**	$\mathcal{D}_{train}$	$\mathcal{D}_{val}$	$\mathcal{D}_{test}$	**Image Type**	**Resolution**
		# normal	# abnormal	# abnormal		
BraTS2021	Brain MRI	4,500	100	400	grayscale	128×128×1
BTCV	Liver CT	3,200	100	400	grayscale	512×512×1
RESC	Retinal OCT	6,200	100	400	grayscale	256×256×1
IDRiD	Fundus Images	7,000	100	400	color	128×128×3

of the abnormal images in $\mathcal{D}_{val}$ and $\mathcal{D}_{test}$ are annotated at the pixel level. The essential information of each dataset is reported in Table 1. We thank the original authors for the excellent datasets.

Evaluation Metrics. We evaluate the detection performance by two classic metrics, such as *Dice score* [5] denoted by `Dice`, and *pixel-level average precision* [18,21] denoted by $\texttt{AP}_{\texttt{pix}}$. Both metrics measure how well the detected abnormal region aligns with the ground truth abnormal region. Larger values of `Dice` and $\texttt{AP}_{\texttt{pix}}$ indicate better performance. We compute the metric value on each image in $\mathcal{D}_{test}$, and report the mean value across all the images in $\mathcal{D}_{test}$.

Baseline Methods and Base Models. Our baseline methods consist of four *self-supervised methods* [16,25,26,23] and nine unsupervised methods categorized as: the *DDPM-based methods* [30,7], the *AE-based methods* [5,14,32,33,12], and the *GAN-based methods* [2,31]. We use each of the AE-based and GAN-based methods as the *base model* to implement MatchGen. The DDPM-based methods are not used as base models because their "decoder" $g(z)$ operates as a stochastic reverse-diffusion process, which prevents Algorithm 1 from effectively passing gradients through $g(z)$.

Implementation Details. For MatchGen built on different base models, we used a learning rate of 10^{-3} with the Adam optimizer [13] to run Algorithm 1. Our code is written in Pytorch 2.0.1 with CUDA 11.8. We use the original code for the baseline methods released by their authors. Following the routine of existing unsupervised methods [5,17,32,9,12,30], for each compared method, we train the generative models $\phi(x)$ and $g(z)$ using $\mathcal{D}_{train}$. For each compared method, we tune the hyperparameters by a grid search on the annotated validation dataset $\mathcal{D}_{val}$. The hyperparameters that yield the highest `Dice` on $\mathcal{D}_{val}$ are selected for evaluating the final detection performance on $\mathcal{D}_{test}$. The search scope for MatchGen is $\epsilon \in [0, 0.6]$ and $\tau \in [0, 1]$. All experiments were conducted on an NVIDIA 4090 GPU. Our code is available at `https://github.com/lele0007/MatchGen`.

3.1 How Are the Detection Results?

The detection results are reported in Table 2, where the best results among all methods are marked in red boxes. For each row of a base model, the immediately following row marked by "+MG" shows the results of MatchGen (MG) implemented on the base model. We can draw the following conclusions from the results in Table 2.

Table 2: The detection performance on $\mathcal{D}_{test}$. "+MG" means applying MatchGen.

Methods		**BraTS2021**		**BTCV**		**RESC**		**IDRiD**	
		↑ Dice	↑ $\texttt{AP}_{\texttt{pix}}$	↑ Dice	↑ $\texttt{AP}_{\texttt{pix}}$	↑ Dice	↑ $\texttt{AP}_{\texttt{pix}}$	↑ Dice	↑ $\texttt{AP}_{\texttt{pix}}$
Self-supervised methods	CutPaste [16]	20.20	12.97	13.29	11.05	12.09	8.51	1.69	1.66
	FPI [25]	16.03	10.82	14.40	12.79	10.25	6.64	6.57	5.71
	PII [26]	23.41	14.47	15.24	13.76	10.98	7.71	15.29	5.99
	NSA [23]	33.80	23.35	15.86	14.20	10.66	7.67	15.05	10.08
DDPM-based methods	AnoDDPM [30]	52.24	47.93	26.84	20.18	25.01	16.06	20.13	12.62
	THOR [7]	53.42	51.44	22.71	18.31	27.25	17.18	15.52	8.72
AE-based methods	AE-ℓ_1 [5]	34.97	25.82	21.78	16.17	26.68	16.55	21.87	13.97
	AE-ℓ_1+MG	41.52	35.93	25.73	18.37	30.01	18.76	27.98	28.86
	CeAE [33]	33.15	25.36	21.28	15.41	27.33	16.97	25.75	20.19
	CeAE+MG	38.76	33.68	25.78	18.43	30.08	18.79	31.17	26.62
	VAE [14]	38.75	29.29	22.02	16.74	26.20	15.97	29.46	23.27
	VAE+MG	42.53	37.30	25.79	18.79	29.98	18.68	33.54	30.35
	VAE-Grad [32]	37.10	33.98	22.24	16.20	27.81	16.98	26.53	20.46
	VAE-Grad+MG	40.93	34.75	25.73	18.14	31.62	19.71	30.27	28.33
	DAE [12]	60.12	54.50	26.59	21.69	26.14	16.17	15.24	8.94
	DAE+MG	69.79	72.03	29.56	25.02	29.74	18.70	18.02	11.53
GAN-based methods	GANomaly [2]	36.02	27.16	25.02	18.17	24.30	15.55	15.49	8.86
	GANomaly+MG	42.79	36.15	28.01	20.82	29.72	18.45	19.09	10.63
	DDGAN [31]	26.08	18.63	20.93	15.96	26.65	16.11	23.30	18.82
	DDGAN+MG	39.26	35.00	25.72	18.12	30.26	18.64	27.54	25.01

First, the self-supervised methods is often inferior to the unsupervised methods. Because self-supervised methods train detection models on synthetic abnormal images that differ substantially from real abnormal images, thus their generalization ability in detecting real abnormal regions is limited [21,9].

Second, MatchGen (MG) consistently outperforms its corresponding base model, which demonstrates the effectiveness of MG in improving the detection performance. In particular, MG achieves the largest performance gain over the base model on BraTS2021 because its images contain less noise than the images in the other datasets. This allows MG to generate higher-quality pseudo-normal images that more accurately match the normal regions of the input images.

In summary, we can conclude from Table 2 that MatchGen effectively improves the detection performance of its base models and the best detection results in Table 2 are always achieved by MatchGen.

3.2 A Case Study to Visually Analyze Detection Results

In this section, we conduct a case study to visually analyze the detection results of each base model and its upgraded version enhanced by MatchGen (MG). We can draw the following conclusions from the results in Figure 2.

First, the base models exhibit large false positive areas marked blue in their detection maps, indicating high false positive rates (FPR). This stems from their test-time pseudo-normal images failing to accurately match normal regions in the input, which leads to poor difference maps that misclassify normal pixels as abnormal. Consequently, the high FPR limits the detection performance of the base models.

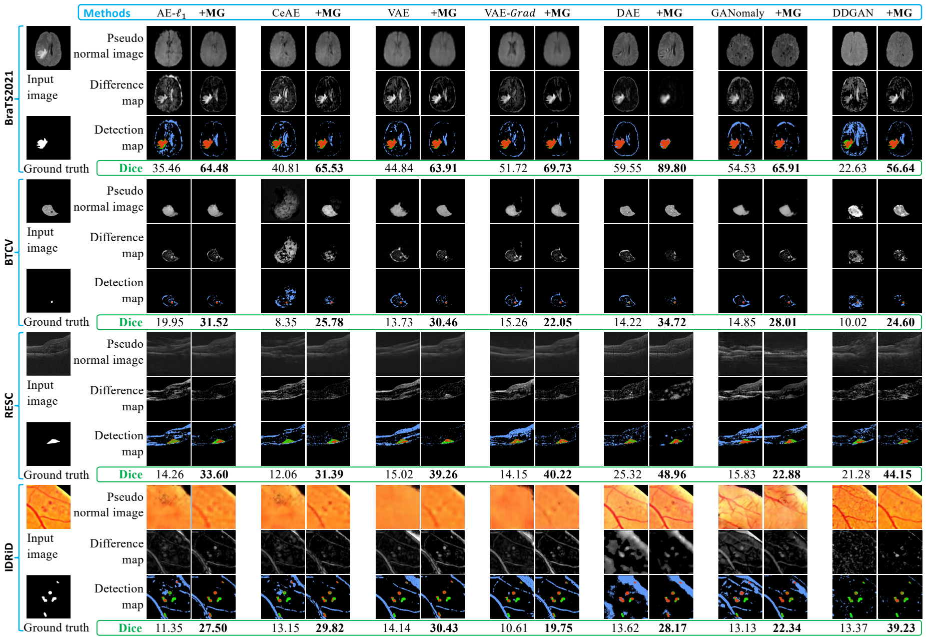

Fig. 2: The case study of detection results. The first column shows the input image and its ground truth abnormal regions. The rest of the columns show two columns per group, where the left column shows the results of a base model and the right column (i.e., +MG) shows the results when applying MG to the base model. In the detection map, the areas in **red**, **blue**, **green**, and **black** are true positives, false positives, false negatives, and true negatives, respectively. The green box marks the `Dice` scores of the detection maps.

Second, for each base model enhanced by MatchGen (denoted by "+MG"), the false positive area marked blue in the detection map is reduced, indicating that MatchGen effectively lowers the FPR by minimizing $\mathcal{L}(z)$. Meanwhile, the constraint $\|z - z_o\|_2 \leq \epsilon$ prevents the pseudo-normal image from overfitting the abnormal regions, enabling effective abnormality detection. As a result, the "+MG" models achieve higher `Dice` scores than their base versions, demonstrating MatchGen's effectiveness in improving detection performance.

In summary, the case study in Figure 2 shows that MatchGen outperforms the base models by reducing false positives through more accurate pixel-level matching between the pseudo-normal and input images.

4 Conclusion and Future Work

In this work, we introduced MatchGen, a novel plug-and-play framework that significantly enhances the performance of unsupervised abnormal region detection methods while remaining practical for clinical deployment. MatchGen mitigates high false positive rates by optimizing a pseudo-normal image at test time to closely match normal regions, with a constraint that prevents overfitting to abnormalities and preserves accurate detection. It is compatible with a broad range of unsupervised methods based on differentiable encoder–decoder architectures. The lightweight optimization of MatchGen processes each image in under 10 seconds on an RTX 4090 GPU, which is practical for clinical diagnostic workflows. Its per-image processing ensures low memory cost and enables parallelization in batch-processing cases such as radiology departments and multi-clinic server-client systems. Extensive experiments on four real-world medical image datasets demonstrate its strong performance. In future work, we will extend MatchGen to improve DDPM-based methods by guiding their reverse diffusion process.

Acknowledgments. This work is supported in part by the NSERC Discovery Grant program (RGPIN-2022-04977), the Canadian Institutes of Health Research (CIHR) (Funding Reference Number: ACD 187254), and the NSERC sMAP CREATE grant #542989-2020.

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

References

1. Creative Commons Attribution 4.0 International (CC BY 4.0) License: https://creativecommons.org/licenses/by/4.0
2. Akcay, S., Atapour-Abarghouei, A., Breckon, T.P.: GANomaly: Semi-Supervised Anomaly Detection via Adversarial Training. In: Asian Conference on Computer Vision. pp. 622–637 (2019)
3. Asgari Taghanaki, S., Abhishek, K., Cohen, J.P., Cohen-Adad, J., Hamarneh, G.: Deep Semantic Segmentation of Natural and Medical Images: A Review. Artificial Intelligence Review **54**, 137–178 (2021)
4. Baid, U., Ghodasara, S., Mohan, S., Bilello, M., Calabrese, E., Colak, E., Farahani, K., Kalpathy-Cramer, J., Kitamura, F.C., Pati, S., et al.: The RSNA-ASNR-MICCAI BraTS 2021 Benchmark on Brain Tumor Segmentation and Radiogenomic Classification. arXiv preprint arXiv:2107.02314 (2021)
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. 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. pp. 1019–1032. PMLR (2024)
7. Bercea, C.I., Wiestler, B., Rueckert, D., Schnabel, J.A.: Diffusion Models with Implicit Guidance for Medical Anomaly Detection. In: International Conference on Medical Image Computing and Computer Assisted Intervention. pp. 211–220 (2024)
8. Cai, Y., Chen, H., Cheng, K.T.: Rethinking Autoencoders for Medical Anomaly Detection from A Theoretical Perspective. In: International Conference on Medical Image Computing and Computer Assisted Intervention. pp. 544–554 (2024)

9. Cai, Y., Zhang, W., Chen, H., Cheng, K.T.: MedIAnomaly: A Comparative Study of Anomaly Detection in Medical Images. Medical Image Analysis p. 103500 (2025)
10. 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)
11. Hu, J., Chen, Y., Yi, Z.: Automated Segmentation of Macular Edema in OCT Using Deep Neural Networks. Medical image analysis **55**, 216–227 (2019)
12. Kascenas, A., Pugeault, N., O'Neil, A.Q.: Denoising Autoencoders for Unsupervised Anomaly Detection in Brain MRI. In: International Conference on Medical Imaging with Deep Learning. pp. 653–664 (2022)
13. Kingma, D.P., Ba, J.: Adam: A Method for Stochastic Optimization. arXiv preprint arXiv:1412.6980 (2014)
14. Kingma, D.P., Welling, M.: Auto-Encoding Variational Bayes. In: International Conference on Learning Representations (2014)
15. Landman, B., Xu, Z., Igelsias, J., Styner, M., Langerak, T., Klein, A.: MICCAI Multi-Atlas Labeling Beyond the Cranial Vault–Workshop and Challenge. In: International Conference on Medical Image Computing and Computer Assisted Intervention Workshop. p. 12 (2015)
16. Li, C.L., Sohn, K., Yoon, J., Pfister, T.: Cutpaste: Self-Supervised Learning for Anomaly Detection and Localization. In: The IEEE/CVF conference on computer vision and pattern recognition. pp. 9664–9674 (2021)
17. Mao, Y., Xue, F.F., Wang, R., Zhang, J., Zheng, W.S., Liu, H.: Abnormality Detection in Chest X-ray Images Using Uncertainty Prediction Autoencoders. In: International Conference on Medical Image Computing and Computer Assisted Intervention. pp. 529–538 (2020)
18. Naval Marimont, S., Siomos, V., Baugh, M., Tzelepis, C., Kainz, B., Tarroni, G.: Ensembled cold-diffusion restorations for unsupervised anomaly detection. In: International Conference on Medical Image Computing and Computer Assisted Intervention. pp. 243–253 (2024)
19. Porwal, P., Pachade, S., Kokare, M., Deshmukh, G., Son, J., Bae, W., Liu, L., Wang, J., Liu, X., Gao, L., et al.: IDRiD: Diabetic Retinopathy–Segmentation and Grading Challenge. Medical Image Analysis **59**, 101561 (2020)
20. Powell, M.J.: A Method for Nonlinear Constraints in Minimization Problems. Optimization, pp. 283–298 (1969)
21. Sato, J., Suzuki, Y., Wataya, T., Nishigaki, D., Kita, K., Yamagata, K., Tomiyama, N., Kido, S.: Anatomy-Aware Self-Supervised Learning for Anomaly Detection in Chest Radiographs. Iscience **26**, 7 (2023)
22. 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)
23. Schlüter, H.M., Tan, J., Hou, B., Kainz, B.: Natural Synthetic Anomalies for Self-Supervised Anomaly Detection and Localization. In: European Conference on Computer Vision. pp. 474–489 (2022)
24. She, D., Zhang, Y., Zhang, Z., Li, H., Yan, Z., Sun, X.: Eoformer: Edge-oriented Transformer for Brain Tumor Segmentation. In: International Conference on Medical Image Computing and Computer Assisted Intervention. pp. 333–343 (2023)
25. Tan, J., Hou, B., Batten, J., Qiu, H., Kainz, B.: Detecting Outliers with Foreign Patch Interpolation. arXiv preprint arXiv:2011.04197 (2020)
26. Tan, J., Hou, B., Day, T., Simpson, J., Rueckert, D., Kainz, B.: Detecting Outliers with Poisson Image Interpolation. In: International Conference on Medical Image Computing and Computer Assisted Intervention. pp. 581–591 (2021)

27. Tibshirani, R.: Regression Shrinkage and Selection via the Lasso. Journal of the Royal Statistical Society Series B: Statistical Methodology **58**, 267–288 (1996)
28. Vyas, B., Rajendran, R.M.: Generative Adversarial Networks for Anomaly Detection in Medical Images. International Journal of Multidisciplinary Innovation and Research Methodology **2**, 52–58 (2023)
29. Wolleb, J., Bieder, F., Friedrich, P., Zhang, P., Durrer, A., Cattin, P.C.: Binary noise for binary tasks: Masked bernoulli diffusion for unsupervised anomaly detection. In: International Conference on Medical Image Computing and Computer Assisted Intervention. pp. 135–145 (2024)
30. Wyatt, J., Leach, A., Schmon, S.M., Willcocks, C.G.: AnoDDPM: Anomaly Detection with Denoising Diffusion Probabilistic Models using Simplex Noise. In: The IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 650–656 (2022)
31. Xiao, Z., Kreis, K., Vahdat, A.: Tackling the Generative Learning Trilemma with Denoising Diffusion GANs. arXiv preprint arXiv:2112.07804 (2021)
32. Zimmerer, D., Isensee, F., Petersen, J., Kohl, S., Maier-Hein, K.: Unsupervised Anomaly Localization Using Variational Auto-Encoders. In: International Conference on Medical Image Computing and Computer Assisted Intervention. pp. 289–297 (2019)
33. Zimmerer, D., Kohl, S.A., Petersen, J., Isensee, F., Maier-Hein, K.H.: Context-encoding Variational Autoencoder for Unsupervised Anomaly Detection. arXiv preprint arXiv:1812.05941 (2018)