Adversarial Denoising Diffusion Model for Unsupervised Anomaly Detection

Jongmin Yu[1], **Hyeontaek Oh**[2], **Jinhong Yang**[3]
[1]Department of Applied Mathematics and Theoretical Physics, University of Cambridge, Centre for Mathematical Sciences, Wilberforce Rd, Cambridge CB3 0WA, United Kingdom
[2]Institute for IT Convergence, Korea Advanced Institute of Science and Technology (KAIST), Daejeon, 34141, Rep. of Korea
[3]Department of Medical Information Technology, Inje University, Kimhae, Republic of Korea, 50834*
jy522@cam.ac.uk, hyeontaek@kaist.ac.kr, jinhong@inje.ac.kr

Abstract

In this paper, we propose the Adversarial Denoising Diffusion Model (ADDM). The ADDM is based on the Denoising Diffusion Probabilistic Model (DDPM) but complementarily trained by adversarial learning. The proposed adversarial learning is achieved by classifying model-based denoised samples and samples to which random Gaussian noise is added to a specific sampling step. With the addition of explicit adversarial learning on data samples, ADDM can learn the semantic characteristics of the data more robustly during training, which achieves a similar data sampling performance with much fewer sampling steps than DDPM. We apply ADDM to anomaly detection in unsupervised MRI images. Experimental results show that the proposed ADDM outperformed existing generative model-based unsupervised anomaly detection methods. In particular, compared to other DDPM-based anomaly detection methods, the proposed ADDM shows better performance with the same number of sampling steps and similar performance with 50% fewer sampling steps.

1 Introduction

The diffusion model is one of the generative models based on thermal non-equilibrium physics [1]. Compared with other generative models such as Generative Adversarial Network (GAN) [2] and Variational Autoencoder (VAE) [3], the diffusion model generates more high-quality data samples, *i.e.,* the diffusion models can derive more discriminative parametric distribution for a given dataset. Based on their superior capability in sampling data, the diffusion models are establishing remarkable success in various domains, such as image and audio generation [4–6] and natural language processing [7]. Those achievements triggered various unsupervised anomaly detection (AD) studies that apply diffusion models to detect data anomalies [8–11].

Those unsupervised AD studies using diffusion models [8–11] have justified that the diffusion models can be used for unsupervised AD and can achieve competitive performance compared with other generative model-based AD methods. The high-quality data sampling capability of the diffusion models can be a strong advantage in driving unsupervised AD methods. However, applying diffusion models is still challenging due to high computational costs caused by iterative sampling or the unstable quality of generated data depending on the number of samples [12].

*Email to the corresponding author: jinhong@inje.ac.kr

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

This paper introduces the Adversarial Denoising Diffusion Model (ADDM), an end-to-end unsupervised AD method using a diffusion model. The proposed ADDM minimises the noise prediction error and explicitly minimises adversarial loss about the denoised sample. Compared to other diffusion model-based methods, ADDM can dramatically reduce the number of data sampling (reverse process) steps by providing an explicit process to improve the quality of denoised data. The ADDM outperforms other diffusion model-based AD methods [8, 10] with 6.2% better performance in our experiments using MRI images for brain tumour detection.

2 Preliminaries

The goal of diffusion models [1] is to find the parameterised data distribution $p_\theta(\mathbf{x}_0)$ using a given data $\mathbf{x}_0 \sim q(\mathbf{x}_0)$. To do this, when a data $\mathbf{x}_0$ is given, the training of diffusion models conducts a *forward process* (*a.k.a.* diffusion process) $q(\mathbf{x}_t|\mathbf{x}_{t-1})$, which adds Gaussian noise to the data and a *reverse process* (*a.k.a.* denoising process) $p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)$, which denoises the given noised data by subtracting predicted noise.

The forward process $q(\mathbf{x}_t|\mathbf{x}_{t-1})$ is a task to add Gaussian noise to data at a certain time step $t \leq T$. The Gaussian noise is generated by a Gaussian probability $\mathcal{N}(\cdot)$ with scheduled variance: $\beta_1, \ldots, \beta_T$. The entire forward process to generate completely noised sample $\mathbf{x}_T$ is represented by

$$q(\mathbf{x}_{1:T}|\mathbf{x}_0) \coloneqq \prod_{t=1}^{T} q(\mathbf{x}_t|\mathbf{x}_{t-1}), \qquad q(\mathbf{x}_t|\mathbf{x}_{t-1}) \coloneqq \mathcal{N}(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t\mathbf{I}). \tag{1}$$

The reverse process $p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)$ can be considered as a denoising task. For each time step t, a diffusion model predicts a noise and subtracts it from the noised data. This task is represented by Markov Chain so that it can be represented by parametric conditional distribution, as follows:

$$p_\theta(\mathbf{x}_{0:T}) \coloneqq p(\mathbf{x}_T) \prod_{t=1}^{T} p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t), \qquad p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) \coloneqq \mathcal{N}(\mathbf{x}_{t-1}; \boldsymbol{\mu}_\theta(\mathbf{x}_t, t), \boldsymbol{\Sigma}_\theta(\mathbf{x}_t, t)). \tag{2}$$

The learning of the diffusion model is to find the suitable parametric distribution $p_\theta(\mathbf{x}_0)$ representing a given data. However, the negative log-likelihood of $p_\theta(\mathbf{x}_0)$ is not nicely computable, so it is alternatively optimised by variational lower-bound, which is represented by

$$L_{dm} \coloneqq \mathbb{E}\left[-\log p_\theta(\mathbf{x}_0)\right] \leq \mathbb{E}_q\left[-\log \frac{p_\theta(\mathbf{x}_{0:T})}{q(\mathbf{x}_{1:T}|\mathbf{x}_0)}\right] = \mathbb{E}_q\left[-\log p(\mathbf{x}_T) - \sum_{t\geq 1} \log \frac{p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)}{q(\mathbf{x}_t|\mathbf{x}_{t-1})}\right]. \tag{3}$$

In DDPM [13], it found that the forward and reverse processes are represented by reparameterisation tricks; therefore, it can replaced by the minimisation of prediction error for a Gaussian noise and noise prediction obtained by a neural network. As a result, Eq. (3) is more simplified to the $l2$-distance between the Generated Gaussian noise $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and predicted noise $\boldsymbol{\epsilon}_\theta$ at a certain time step. Using the notations, $\alpha_t \coloneqq 1 - \beta_t$ and $\bar{\alpha}_t \coloneqq \prod_{s=1}^{t} \alpha_s$, the simplified loss is represented by

$$\mathbb{E}_{\mathbf{x}_0, \boldsymbol{\epsilon}}\left[\frac{\beta_t^2}{2\sigma_t^2 \alpha_t (1-\bar{\alpha}_t)} \left\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}, t)\right\|^2\right]. \tag{4}$$

3 Adversarial Denoising Diffusion Model

Unlike GAN [2] or VAE [3], which directly generate data from relatively low-dimensional Gaussian noise and classify whether the samples are generated or given (in the case of GAN) or minimise Euclidean distances of generated samples and a given sample (in the case of VAE), to generate data from complete, from a given noise which having the same dimensionality with the data, diffusion models repeatedly performed the reverse process, *i.e.,* predicting the noise present in the given data and remove it.

The diffusion model methodology, that is, generating data through an iterative procedure that gradually restores data from the same dimensional noise, shows outstanding data sampling performance,

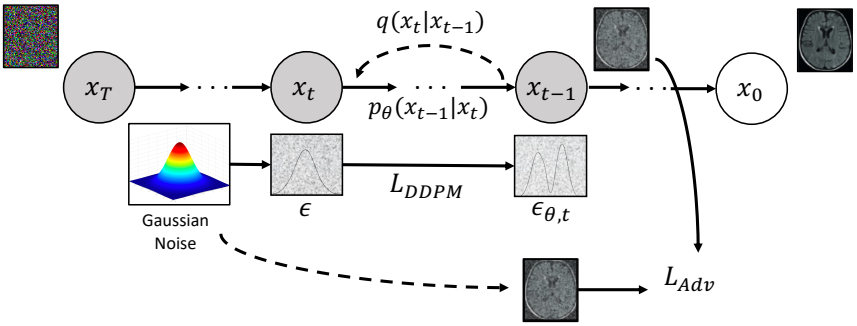

Figure 1: Illustration for Learning of Adversarial Denoising Diffusion Model (ADDM). L_{DDPM} denotes the loss function to minimise the $l2$-distance between the generated error from Gaussian distribution and the predicted error by a model. L_{Adv} indicates the adversarial loss between the model-based denoised sample x_{t-1} and virtually generated noised samples, for the certain time step $t-1$.

but compared to GAN and VAE, the sampling process of the diffusion model is cost-efficient. Additionally, the quality of generated data is varied depending on the number of sampling steps [12]. This issue is because diffusion model learning involves noise prediction, which is not closely related to the semantic properties of data. When generating noise, $\mathbf{x}_0$ is used conditionally (See Eq. (2)), but the model ultimately learns the minimum prediction error between Gaussian noise and the noise present in the image (See Eq. (4)). In this study, adversarial learning is proposed to minimise changes in the diffusion model's structure and preserve the semantic characteristics of data.

Figure 1 shows the architectural details of the proposed ADDM. The objective function of the adversarial denoising diffusion model (ADDM) is defined by the summation of the DDPM loss (Same with Eq. (4)) and the additional adversarial loss using a balancing weight λ, represented by

$$\begin{aligned} L_{\mathrm{ADDM}} \coloneqq & \underbrace{\mathbb{E}_{t,\mathbf{x}_0,\boldsymbol{\epsilon}}\left[\left\|\boldsymbol{\epsilon}-\boldsymbol{\epsilon}_\theta(\sqrt{\bar{\alpha}_t}\mathbf{x}_0+\sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon},t)\right\|^2\right]}_{L_{\mathrm{DDPM}}} \\ & + \underbrace{\lambda\mathbb{E}_{t,\mathbf{x}_0,\boldsymbol{\epsilon}}[\log D(q(\mathbf{x}_{t-1}))]+\mathbb{E}_{t,\mathbf{x}_0,\boldsymbol{\epsilon}}[\log\left(1-D(p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t))\right)]}_{L_{\mathrm{Adv}}}, \end{aligned} \tag{5}$$

where D denotes the discriminator to apply the adversarial learning on ADDM. $q(\mathbf{x}_t|\mathbf{x}_{t-1})$ defines the t-step noised data obtained by the forward process (Eq. (1)), and $p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)$ indicates the denoised data produced by the reverse process (Eq. (2)).

L_{DDPM} is a simplified version of Eq. (4). DDPM [13] supposes that the simplified loss function is more beneficial to the quality of the sample and implementation efficiency; therefore, we keep this loss format in our study. L_{Adv} is the loss term for the adversarial learning about the denoised data. L_{Adv} tries to distinguish between the real noise data $q(\mathbf{x}_t|\mathbf{x}_{t-1})$ and the denoised data $p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)$.

In predicting a noise for the noise removal process for data generation through DDPM, the ADDM model can learn semantic features of the data and information about the noise to be predicted. Through this direct learning of semantic features of data, we expect to reduce the number of sampling steps required to generate sufficient quality data. Besides, the adversarial loss helps to reduce blurriness, which means it helps to learn high-frequency features to add more local details in denoising the data. This property will help prove the accuracy of AD methods by reducing the false-positive results. Through ablation studies, we will demonstrate that the proposed adversarial learning helps AD.

4 Experiment

4.1 Experimental settings

Dataset and Protocol: We utilise the two datasets. The first dataset is the Neurofeedback Skull-Stripped (NFBS) repository [14]. The repository provides 125 MRI images captured from normal people, so there are no anomalies in the MRI images. The second dataset contains 22 T1-weighted MRI scans provided by the Centre for Clinical Brain Sciences from the University of Edinburgh [15].

Method	Dice	AUC	IoU	Precision	Recall
AE (Spatial) [17]	0.252	0.707	0.162	0.258	0.279
VAE (Dense) [17]	0.317	0.734	0.203	0.297	0.313
f-AnoGAN [18]	0.128	0.789	0.093	0.362	0.080
Transformer [19]	0.241	0.695	0.193	0.275	0.120
Pinaya *et al.* [10] ($T = 1000$)	0.375	0.815	0.238	0.367	0.452
AnoDDPM [8] ($T = 1000$)	0.383	0.863	0.269	0.373	0.468
ADDM ($T = 300$)	0.301	0.811	0.213	0.293	0.369
ADDM ($T = 500$)	0.379	0.861	0.271	0.361	0.491
ADDM ($T = 1000$)	**0.403**	**0.917**	**0.289**	**0.392**	**0.508**

Table 1: Quantitative performance comparison of unsupervised AD methods based on generative models. T denotes the number of sampling steps of a diffusion model.

The second dataset provides MRI images containing brain tumours. To follow an experimental protocol for the unsupervised AD, the training dataset has to be composed of normal samples only. We train the ADDM using the first dataset and conduct AD experiments using the second dataset. We refer to the experiment protocol of AnoDDPM [8].

Implementations: The resolution of images is resized to 256×256. Adam optimiser [16] is used for the optimisation algorithm. The balancing weight λ is set by 0.05. The number of epochs and the batch size are set by 3000 and 4, respectively. To demonstrate the effectiveness of adversarial learning, we evaluate the AD performance with 300, 500, and 1000 sampling steps (T). The learning rate is initialised by 0.0001, and it is decayed by multiplying 0.999 for every 200 epoch.

4.2 Experimental results

Effectivenss of L_{Adv}: We train the ADDM with 300, 500, and 1000 sampling steps. Table 1 contains the quantitative performances of the ADDM depending on T. The ADDM obtains the best performance with 1000 sampling steps. It produces 0.403 Dice and 0.917 AUC. The overall performance of the ADDM with 1000 sampling steps is better than the other two ADDMs trained with 300 and 500 sampling steps. However, the ADDMs trained with 300 and 500 sampling steps produce competitive performance compared with other DDPM-based models [8, 10]. In particular, the ADDM trained with 500 sampling steps outperforms the Pinaya *et al.* [10], which is structurally equivalent to the DDPM with 1000 sampling steps. The experimental results justify that the adversarial learning on the ADDM improves the robustness of the diffusion models with respect to the sampling step.

Comparison with SOTA methods: We compare the proposed method with various generative model-based AD methods [8, 17–19]. Table 1 shows the quantitative results on the dataset. Listed methods have been chosen for performance comparison: Autoencoder (AE), Variational AE, [17], f-AnoGAN [18], Transformer [19], Pinaya *et al.* [10], and AnoDDPM [8]. In particular, Pinaya *et al.* [10] and AnoDDPM [8] are built based on DDPM [13], so their baseline methods are similar to the ADDM. Both approaches compile reconstruction-based AD methods for MRI images using diffusion models. Interestingly, the architectural details of Pinaya *et al.* [10] are almost identical to the DDPM [13]. AnoDDPM [8] is built based on the DDPM and uses a new noising approach called *Simplex Noise*.

The quantitative results in Table 1 show that the ADDM outperforms other methods. The ADDM (T=1000) produces 0.917 AUC, which is 6.2% higher performance than the second-ranked method (AnoDDPM). Moreover, the proposed ADDM trained with 500 sampling steps also achieves competitive performance with the AnoDDPM that requires 1000 sampling steps. This result shows that the proposed ADDM is more cost-efficient than AnoDDPM. Overall, the experimental results show that the proposed adversarial loss improves AD performances and outperforms existing SOTA AD detection methods on MRI images.

5 Conclusion

In this work, we have presented an adversarial denoising diffusion model (ADDM) for AD on MRI images, which can derive a more discriminative AD method based on diffusion models. Similar to DDPM, the proposed ADDM learns a noise prediction model based on the forward and reverse

processes. The adversarial learning is applied to improve the quality of denoised data explicitly. Our ablation studies have shown that adversarial learning is helpful not only in improving AD performance but also in reducing the sampling step to obtain high-quality data. In comparison with existing AD methods based on generative models, ADDM outperforms existing SOTA AD methods.

6 Acknolwdgement

This work was partly supported by the Institute of Information & Communications Technology Planning & Evaluation(IITP) grant funded by the Korea government(Ministry of Science and ICT; MSIT) (2020-0-00048, Development of 5G-IoT Trustworthy AI-Data Commons Framework) and by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (NRF-2022R1C1C2003437)

References

[1] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 2256–2265, Lille, France, 07–09 Jul 2015. PMLR.

[2] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014.

[3] Diederik P Kingma and Max Welling. Auto-encoding variational bayes, 2022.

[4] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, Jonathan Ho, David J Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 36479–36494. Curran Associates, Inc., 2022.

[5] Congyue Deng, Chiyu "Max" Jiang, Charles R. Qi, Xinchen Yan, Yin Zhou, Leonidas Guibas, and Dragomir Anguelov. Nerdi: Single-view nerf synthesis with language-guided diffusion as general image priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20637–20647, June 2023.

[6] Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. Imagic: Text-based real image editing with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6007–6017, June 2023.

[7] Zhenghao Lin, Yeyun Gong, Yelong Shen, Tong Wu, Zhihao Fan, Chen Lin, Nan Duan, and Weizhu Chen. Text generation with diffusion language models: A pre-training approach with continuous paragraph denoise. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 21051–21064. PMLR, 23–29 Jul 2023.

[8] 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 (CVPR) Workshops*, pages 650–656, June 2022.

[9] Julia Wolleb, Florentin Bieder, Robin Sandkühler, and Philippe C. Cattin. Diffusion models for medical anomaly detection. In Linwei Wang, Qi Dou, P. Thomas Fletcher, Stefanie Speidel, and Shuo Li, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022*, pages 35–45, Cham, 2022. Springer Nature Switzerland.

[10] Walter H. L. 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, David Werring, Geraint Rees, Parashkev Nachev, Sebastien Ourselin, and M. Jorge Cardoso. Fast unsupervised brain anomaly detection and segmentation with diffusion models. In Linwei Wang, Qi Dou, P. Thomas Fletcher, Stefanie Speidel, and Shuo Li, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022*, pages 705–714, Cham, 2022. Springer Nature Switzerland.

[11] Mark S. Graham, Walter H.L. Pinaya, Petru-Daniel Tudosiu, Parashkev Nachev, Sebastien Ourselin, and Jorge Cardoso. Denoising diffusion models for out-of-distribution detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 2947–2956, June 2023.

[12] Florinel-Alin Croitoru, Vlad Hondru, Radu Tudor Ionescu, and Mubarak Shah. Diffusion models in vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[13] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[14] Benjamin Puccio, James P Pooley, John S Pellman, Elise C Taverna, and R Cameron Craddock. The preprocessed connectomes project repository of manually corrected skull-stripped t1-weighted anatomical mri data. *Gigascience*, 5(1):s13742–016, 2016.

[15] Cyril R Pernet, Krzysztof J Gorgolewski, Dominic Job, David Rodriguez, Ian Whittle, and Joanna Wardlaw. A structural and functional magnetic resonance imaging dataset of brain tumour patients. *Scientific data*, 3(1):1–6, 2016.

[16] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[17] 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*, 69:101952, 2021.

[18] 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.

[19] Walter Hugo Lopez Pinaya, Petru-Daniel Tudosiu, Robert Gray, Geraint Rees, Parashkev Nachev, Sébastien Ourselin, and M Jorge Cardoso. Unsupervised brain anomaly detection and segmentation with transformers. *arXiv preprint arXiv:2102.11650*, 2021.