Self-Supervised Anomaly Segmentation via Diffusion Models with Dynamic Transformer UNet

Komal Kumar[1], Snehashis Chakraborty[1], Dwarikanath Mahapatra[2],
Behzad Bozorgtabar[3], Sudipta Roy[1]
[1]Artificial Intelligence & Data Science, Jio Institute, Navi Mumbai, India,
[2]Faculty of IT, Monash University, Australia,
[3]Swiss Federal Institute of Technology Lausanne (EPFL) University Hospital Center (CHUV),
Lausanne, Switzerland
suryavansi8650@gmail.com, snehashis.chakraborty@jioinstitute.edu.in, dmahapatra@gmail.com,
behzad.bozorgtabar@epfl.ch, sudipta1.roy@jioinstitute.edu.in

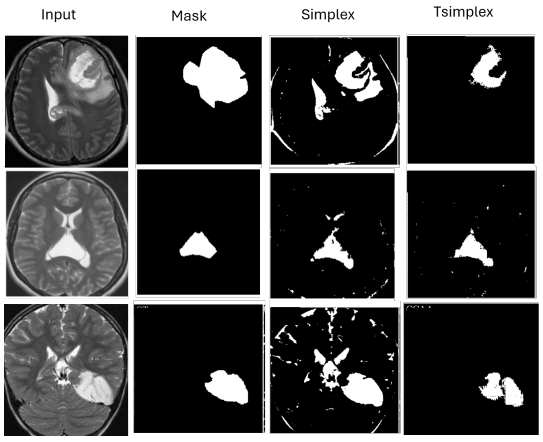
Abstract

A robust anomaly detection mechanism should possess the capability to effectively remediate anomalies, restoring them to a healthy state, while preserving essential healthy information. Despite the efficacy of existing generative models in learning the underlying distribution of healthy reference data, they face primary challenges when it comes to efficiently repair larger anomalies or anomalies situated near high pixel-density regions. In this paper, we introduce a self-supervised anomaly detection method based on a diffusion model that samples from multi-frequency, four-dimensional simplex noise and makes predictions using our proposed Dynamic Transformer UNet (DTUNet). This simplex-based noise function helps address primary problems to some extent and is scalable for three-dimensional and colored images. In the evolution of ViT, our developed architecture serving as the backbone for the diffusion model, is tailored to treat time and noise image patches as tokens. We incorporate long skip connections bridging the shallow and deep layers, along with smaller skip connections within these layers. Furthermore, we integrate a partial diffusion Markov process, which reduces sampling time, thus enhancing scalability. Our method surpasses existing generative-based anomaly detection methods across three diverse datasets, which include BrainMRI, Brats2021, and the MVtec dataset. It achieves an average improvement of $+10.1\%$ *in Dice coefficient,* $+10.4\%$ *in IOU, and* $+9.6\%$ *in AUC. Our source code is made publicly available on Github.*

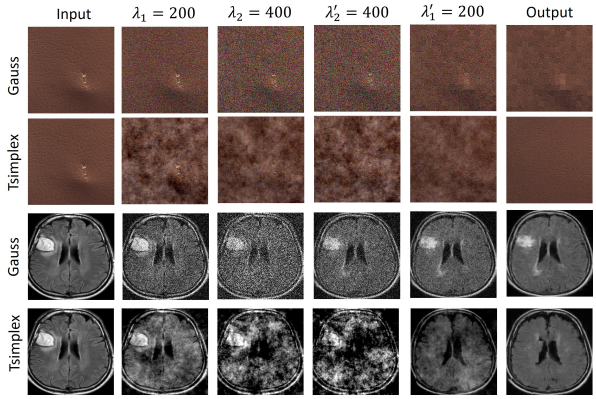
Keywords: *Anomaly Detection, Self-Supervised Learning, Diffusion models, UNet, Noise function.*

1. Introduction

The scarcity of experts with the ability to diagnose and treat specific medical conditions is a pressing concern in developing countries [14]. To illustrate, consider the ratio of dermatologists to the general population, which can plummet to as low as 1 per 216,000 people [10]. This motivation fuels the development of a deep learning system with the ability to localize diseases and thereby prevent misdiagnosis or underdiagnosis [7, 20]. Nonetheless, employing supervised learning models poses notable challenges due to the substantial amount of annotated data they necessitate, making the acquisition process expensive and time-consuming. Self-supervised anomaly detection is a powerful deep learning algorithm that trains on healthy or normal inference data which is used as a threshold for anomalies. The primary aspect of these algorithms is to address unhealthy or abnormal regions, followed by the calculation of the target anomaly using the difference of squares. When it comes to diverse anatomy, anomaly patterns, and distribution shifts [15], image data, especially medical image data, can be quite complex. Generative models have demonstrated their potency in self-supervised representation learning of the underlying distribution, particularly in the context of healthy inference data [14, 31, 46, 49]. Denoising diffusion probabilistic models (DDPMs) [17] have demonstrated remarkable effectiveness in self-supervised representation learning and are capable of generating samples even from complex data distributions with superior convergence, as compared to generative adversarial networks (GANs) and variational autoencoders (VAEs) [5, 46]. The DDPM consists of two steps: a forward noise injection step and a backward denoising step. In the forward step, noise is injected from a $N(0, I)$ distribution, while the denoising backward step stochastically transforms



(a) Detection of complex distribution anomalies without relying on annotations. It detects both big and small anomalies, even those situated near high-intensity regions, challenging existing methods.



(b) Anomaly repair partial diffusion processes with Tsimplex (Our) and Gauss noise, where λ_1 and λ_2 represent intermediate steps in the forward process, while $\lambda_1^{'}$ and $\lambda_2^{'}$ indicate diffusion steps in the backward process. Tsimplex noise recovers the anomaly over Gauss.

Figure 1. Our partial diffusion model $(400/1000)$ trained on healthy reference data for anomaly detection in a self-supervised.

the samples from a Gaussian distribution onto a learned data distribution. We employ this approach to train DDPM on healthy reference data, which maps anomaly data onto the healthy distribution through a diffusion process but Gaussian noise at each diffusion step is not able to recover the anomaly, and that results in unrepaired anomaly (see Figure 1). Gaussian noise has a constant power spectral density, meaning it has equal power across all frequencies which makes it "white" noise. Simplex noise [30] is used for tasks such as procedural terrain generation, texture synthesis, and creating natural-looking patterns like clouds or marble textures. However, models trained based on simplex noise have a few disadvantages, including a decrease in sample quality, particularly when subjected to higher noise levels (a further t value). These models also struggle to repair anomalies situated near other high-frequency information (See Figure 1a), and they exhibit limited exploration capabilities, especially in the context of complex and high-dimensional simplex noise, which affects their focus on tasks like processing colored and higher dimension images. Moreover, the model was trained using a batch size (bs) of one due to the time complexity, which scales as $\mathcal{O}(bs \times t)$, where t represents the time required to sample an image of size (H, W). This motivation led us to address these challenges, resulting in the development of a four-dimensional simplex noise function capable of generating noise for colored and higher-dimension images while maintaining the same processing time for batched images. Additionally, Vision Transformers (ViTs) integrated into UNet architectures offer significant advantages over other models for image generation, including enhanced efficiency, versatility, and robustness [3]. However, ViTs encounter challenges with spatial relevance and weak channel representation, crucial for accurate image recognition and generation [26]. To address these issues, we introduce DTU-Net, a dynamic Transformer UNet architecture inspired by ViTs, which serves as the backbone for diffusion models. DTU-Net incorporates various components such as Patch Embedding, multi-head attention, multi-layer perceptron, and refinement layer, through thorough experimental analyses. Our algorithms offer several advantages over adversarial training, including improved sample quality and stable training, particularly beneficial for smaller datasets.

The contributions of the paper are summarized as follows:

- Our enhancement of the simplex noise function, Tsimplex, reduces processing time for multiple, colored, and higher-dimensional images. By generating averaged sample outputs, we mitigate noise stochasticity and improve sample quality. Tsimplex enables anomaly detection with partial diffusion, significantly reducing backward process time.

- We developed a ViT-based U-Net model as the core for diffusion models, treating noised image and time step as tokens. We improved the multi-head attention mechanism through dynamic interactivity among attention heads.

In addition, we conducted extensive experiments and an in-depth ablation study on three datasets, demonstrating superior performance, especially in the Brain MRI dataset, and providing valuable insights for the research community.

2. Related Work

Self-supervised anomaly detection: In the realm of anomaly detection, Self-Supervised Learning (SSL) plays a pivotal role in training systems to capture intricate relationships within data, with a primary focus on detecting irregular patterns. SSL encompasses two distinct ap-

proaches: Invariance-based methods [1, 5, 6] and generative methods [9, 17, 18]. Generative models, in particular, have made substantial contributions to anomaly detection and have paved the way for addressing more intricate tasks, particularly in self-supervised understanding and the generation of natural images. Authors in [33] introduced DCGAN, showcasing GANs' ability to capture semantic image content, which has led to intriguing applications like vector arithmetic for manipulating visual concepts. Additionally, [50] trained GANs on natural images and employed the trained models for semantic image inpainting, demonstrating the versatility and potential of GANs in various image-related tasks. While VAE models are sometimes criticized for their poor sample quality, GANs [40] come with their set of challenges, including their inability to repair anomalies, training instability, model collapse, and reliance on large datasets. Recent strides in the field of Diffusion Probabilistic Models (DDPM) [9, 17] have showcased their ability to generate higher-quality samples from complex distributions with superior coverage compared to GANs [40] and VAEs. However, these improvements come at the expense of reduced scalability and increased sampling times, primarily due to the necessity of employing long Markov chain sequences [19]. Furthermore, DDPMs also have limitations in capturing larger anomalies caused by Gaussian noise. In their paper [49], the authors introduced a noising scheme for diffusion models based on simplex noise. However, this scheme comes with several significant drawbacks. As the noise level increases, there is a noticeable decrease in sample quality, especially when applying noise to higher values of "t". Moreover, the scheme struggles to effectively repair anomalies located near other high-frequency information, limiting its ability to handle complex data patterns. Additionally, it exhibits limited exploration capabilities, especially concerning 3D and colored images, where its focus is less well-defined. Most crucially, this noising scheme leads to increased sampling times, which can be a significant practical limitation, particularly when dealing with batched images or real-time applications.

Backbone of diffusion models: Indeed, along with the development of diffusion model algorithms [3, 18, 24, 25, 41, 42, 44, 45], the revolution in backbone models plays a crucial and integral role. An illustrative instance is U-Net, constructed upon a convolutional neural network (CNN) and previously utilized in research [17, 43]. The CNN-driven U-Net design features a sequence of down-sampling blocks, a series of up-sampling blocks, and extensive skip connections between these two sets of blocks [9, 35, 39]. This architectural framework has held a prominent position within diffusion models utilized for image generation assignments. Conversely, ViTs [11] have demonstrated promising, and in some cases, superior performance compared to CNNs in a range of tasks. In their paper [3], the authors introduce a straightforward and versatile architecture for image generation using ViTs within diffusion models. Experimental results illustrate that U-ViT performs on par with, if not better than, a CNN-based U-Net of a similar size. However, recent studies [12, 27, 34] that investigate the reasons behind the difference in data efficiency between ViTs and CNNs have led to the conclusion that it attributed to a lack of inductive bias. Based on our experiments we also found out that the ViT-based diffusion backbone [3, 29] fails to generate the image when it comes to small datasets.

3. Methodology

3.1. Background

Diffusion models, specifically Diffusion Probabilistic Models (DDPMs) [17], are a class of generative models that employ a diffusion process resembling a Markov chain. This process comprises sequential steps, where each step involves sampling from a Gaussian distribution. Importantly, the mean of this distribution depends on the current state of the chain. As the number of steps increases, the distribution over the chain converges to a Gaussian distribution. Let's begin with data represented as $x_0 \sim q(x_0)$ and a Markov chain process q progressing from x_1 to x_χ, injecting noise at each step from a normal distribution with a variance schedule β_t:

$$q(x_t \mid x_{t-1}) = N(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta I) \tag{1}$$

In DDPM, instead of repeatedly applying q to sample $x_t \sim q(x_t|x_0)$, it expresses $q(x_t|x_0)$ as a Gaussian distribution using an auxiliary noise variable $\eta \sim N(0, I)$:

$$q(x_t \mid x_0) = N(x_t; \sqrt{\bar{\alpha}_t}x_0, (1-\bar{\alpha}_t)I) \tag{2}$$

$$= \sqrt{\bar{\alpha}_t}x_0 + \eta\sqrt{1-\bar{\alpha}_t} \tag{3}$$

Here, $1-\alpha_t = \beta_t$, and $\bar{\alpha}_t = \prod_{s=0:\chi}\alpha_s$. $1-\bar{\alpha}_t$ serves as a noise scheduler in place of β_t. To sample from the posterior distribution, which is also Gaussian, Bayes' theorem is applied by sampling from each reverse step of the distribution $q(x_{t-1}|x_t)$ for t ranging from χ to 1, eventually reaching $q(x_0)$. The parameters (mean vector and covariance matrix) of $q(x_{t-1}|x_t)$ can be estimated using neural networks, which approximate this distribution. The objective of these neural networks is to minimize the dissimilarity between probability distributions from step t to $t-1$ using Kullback-Leibler divergence (D_{KL}). The loss function for training the parameterized distribution $p_\theta(x_{t-1}|x_t)$ is expressed as the variational lower bound L_{vlb} on the marginal likelihood $p_\theta(x_0)$. It is defined as the sum of terms L_0 through L_T, where $L_0 = -\log p_\theta(x_0|x_1)$, $L_{t-1} = D_{KL}(q(x_{t-1}|x_t, x_0)||p_\theta(x_{t-1}|x_t))$, and $L_T = D_{KL}(q(x_T|x_0)||p(x_T))$. These terms quantify the reconstruction loss, conditional divergence, and final divergence,

respectively, in the context of this variational lower bound formulation.

Conceptually, a neural network can also be seen as a mapping from a simpler Gaussian distribution to a more complex distribution of images. This mapping is thought of as a non-parametric method for defining the mean function of a Gaussian process. A network denoted as $\eta_\theta(x_t,\ t)$ with parameters θ for predicting η trained to simplify the objective by DDPM [17] and enhance the quality of sampling. For a given $x_0 \sim q(x_0)$ and $\eta \sim N(0, I)$ at each step t within the range $[0, \chi]$, the following loss function is defined:

$$L(\theta) = \frac{1}{\chi} \sum_{t=0:\chi} ||\eta - \eta_\theta\left(x_t,\ t\right)||^2 + L_{vlb} \tag{4}$$

3.2. Noise function

The visual world is endlessly captivating due to its consistency across different scales, a phenomenon known as scale invariance, which can be observed in various visual contexts [38]. In natural images, the distribution of frequencies adheres to a power-law distribution, with lower-frequency components playing a more substantial role in defining the image's characteristics [47, 49]. However, there's a notable discrepancy in how DDPMs treat lower-frequency and high-frequency components when Gaussian noise is employed, mainly due to the uniform spectral density of this noise source as shown in Figure 1b. Conversely, diffusion models employing simplex noise tend to assume that lower-frequency components are relatively less corrupted, leading to the recovery of larger anomalous regions in the reverse process (Figure 1b for Tsimplex).

Tsimplex: Simplex noise[1] represents an enhancement of Perlin noise [30], characterized by increased computational efficiency and the generation of smooth and structured randomness. Tsimplex noise relies on gradient noise following simplex and is generated through the amalgamation of numerous noise octaves. In this process, each octave represents a higher-frequency variation of the noise from the previous octave. These octaves are weighted with decreasing amplitude and increasing frequency as outlined in Algorithm 1, yielding a more intricate and detailed noise pattern. The outcome is a textured noise that exhibits gradual and continuous variations across spatial dimensions shown in Figure 3a. Additionally, in Tsimplex noise, we incorporate the temporal dimension (B) and diffusion time steps as a grid along with spatial dimension. Consequently, the total time required to sample noise from Tsimplex for a batch of 1000 steps (approximately 5.97 seconds) become less than simplex noise and close to Gaussian noise, as demonstrated in Figure 2. This improvement is also attributed to the Honeycomb pattern introduced by the additional time step parameter (t), $t \in \mathbb{R}^{bs \times \chi}$, thereby incorporating a hexagonal pattern (honeycomb) due to the inherent characteristics of simplex noise. This incorporation results in a tetrahedral honeycomb pattern. In 2D simplex noise, these shapes manifest as equilateral triangles forming a hexagonal grid, while in 3D simplex noise, the lattice is composed of tetrahedra (See [30] for details). Our experiments show that these added honeycomb patterns contribute to increased symmetry, especially at a minimal octave value (Figure 3), ultimately enhancing sample quality as demonstrated. Furthermore, we have observed that the sample quality of simplex noise decreases as t increases. One of the reasons for the lower sample quality is the asymmetry of simplex noise, as shown in Figure 3c, in contrast to Gaussian noise. In contrast, Tsimplex[2] provides similar sample quality to Gaussian noise, as illustrated in Figure 3b.

[1] https://github.com/lmas/opensimplex

Algorithm 1 Noise(S, t, O, p, μ)

Initialize, $A \leftarrow 1$, $N \leftarrow 0$
$x, y \leftarrow$ spatial grids,
$C \leftarrow$ channel grids $B \leftarrow$ batch grids $\vee$ 3D grids
for i from 1 to O **do**
  $N \leftarrow \text{simplex}(\frac{x}{\mu}, \frac{y}{\mu}, \frac{B}{\mu}, \frac{C}{\mu}, \frac{t}{\mu})$
  $N \leftarrow N + A \cdot N$
  $\mu \leftarrow \frac{\mu}{2}, A \leftarrow A \cdot p$
end for
return N

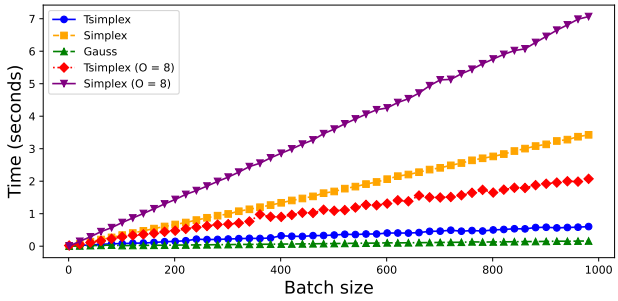
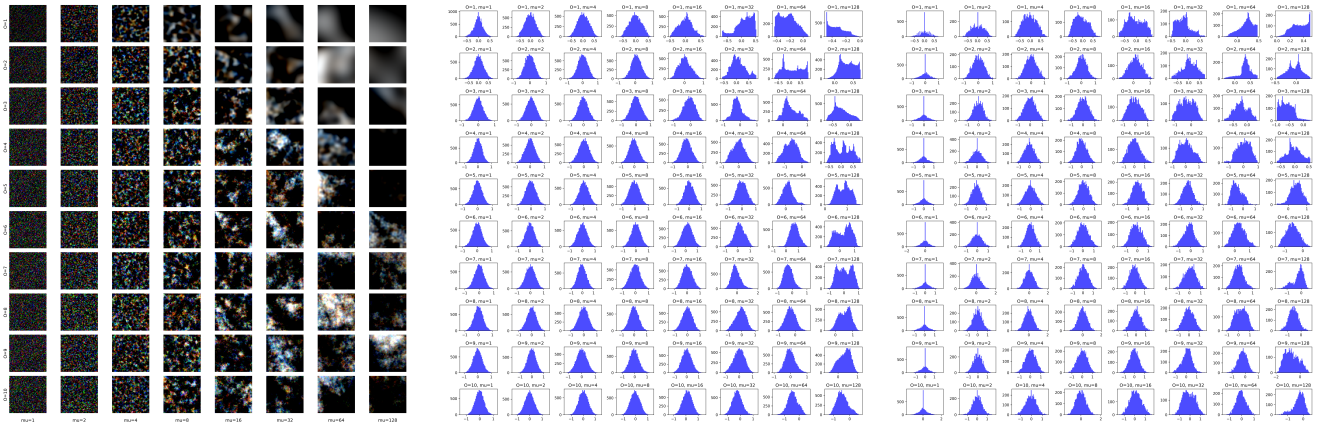


Figure 2. Time Required for sampling up to 1000 steps for grid of 64×64 and for O octaves. Tsimplex requires less time than simplex as batch size goes further.

3.3. Dynamic Transformer UNet

DTU-Net serves as a fundamental component in diffusion models (see Figure 4a) for anomaly detection and can be applied across a variety of tasks within diffusion modeling. The primary goal of DTU-Net is to minimize the loss as defined in equation 4 which combines $L2-norm$ with L_{vlb} for the robust self-supervised training and generate predic-

[2] Please see more analysis and generation method in Supplementary Material



(a) Tsimplex noise structure (b) Tsimplex noise distribution (c) Simplex noise distribution

Figure 3. Comparing Tsimplex and Simplex noise, we analyze the impact of two variables: octave (O) on the y-axis and frequency (μ) on the x-axis, examining their influence on both distribution and structure. We selected a minimum value of O with the highest symmetry.

tions by removing the noise to reconstruct the image. It takes the noisy input x_t, the time step t, and predicts the noise added to x_t.

Sliding window patch embedding: Following the architectural principles of ViT [11], DTU-Net divides the input images into patches and treats all patches, along with time, as tokens. The whole process of sliding window patch embedding (SPE) is shown in Figure 4a and the function is formulated as follows.

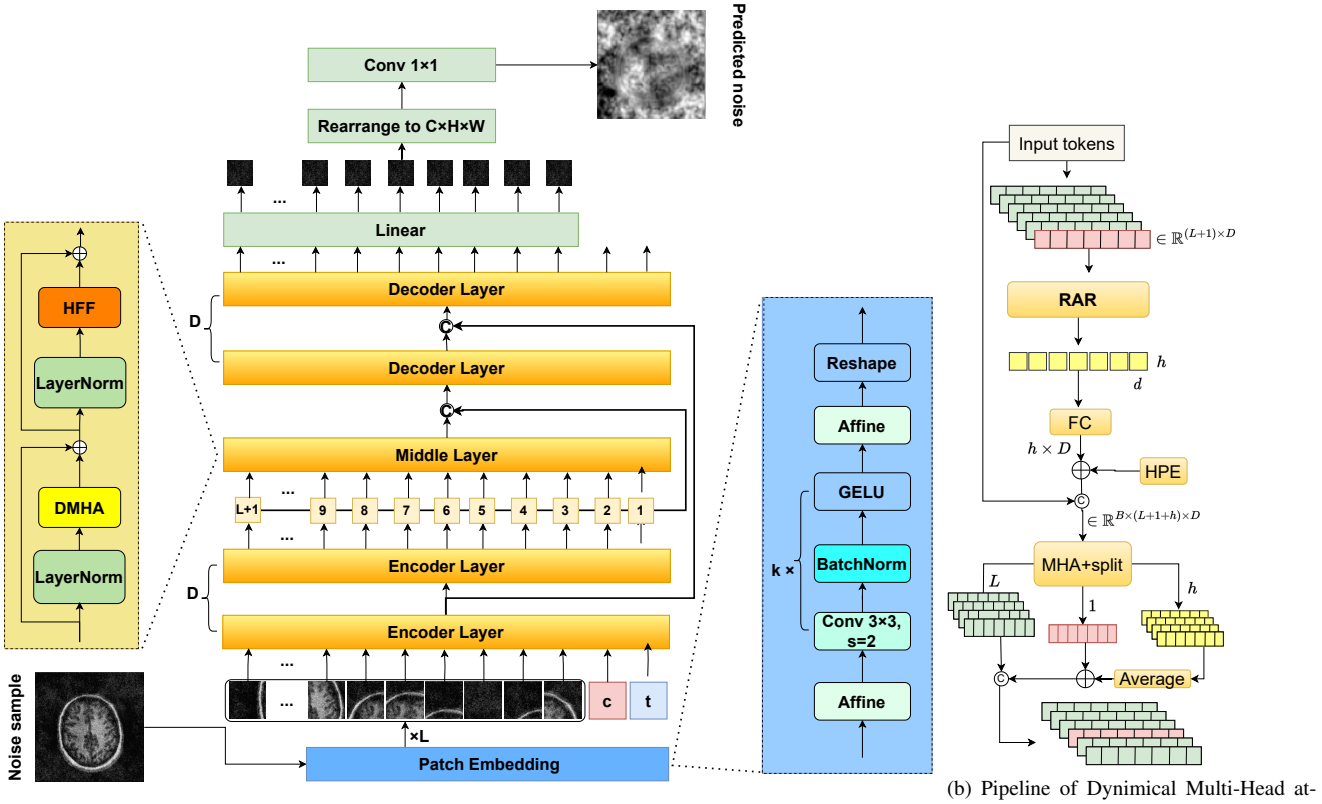
$$\text{Affine}(x) = \text{Diag}(\nu)x + \phi \tag{5}$$

Where ν, and ϕ are learnable parameters initilized with 1 and 0 resepctively. The output from the affine function undergoes a series of operations, including a Conv(3,3) operation, followed by batch normalization and activation functions, and this sequence is repeated up to k times. Finally, the result is post-processed once again through an affine function to get the sliding window patch embedding ($L = \frac{H \times W}{P^2}$).

DTU-Net Layer: Following the architecture of UNet, DTU-Net also comprises three types of layers: Encoder, Middle, and Decoder. These layers consist of the same types of blocks, as indicated by the colors in Figure 4a. These blocks primarily include dynamic multi-head attention (DMHA) (refer to Figure 4b) and a hybrid feedforward (HFF) block (see Figure 5). Inspired by [26], we incorporate the Head token into our DMHA. This addresses the issue of inductive bias in ViTs by allowing interactions between multiple heads, in contrast to the hierarchical structure of the ViT with window attention [22].

DMHA: The mechanism is formulated as follows: For $x \in \mathbb{R}^{bs \times L \times D}$, we first apply Rearrange Average and Reshape (RAR), i.e., we rearrange D into $h \times d$, average with respect to $L + 1$ tokens, and reshape into $h \times d$. The output is then projected into $h \times D$ through a fully connected layer (FC), followed by layer normalization and activation. The head tokens are added with the head position embedding (HPE) so the position embedding of the head will not be forgotten. We concatenate the generated head token with shortcut input and feed it to the multi-head attention. Finally, output is transformed into the original shape by splitting into $L \times D$, $1 \times D$, and $h \times D$ as shown in Figure 4b.

HFF: We employ a channel attention mechanism to consolidate the features of patch tokens into the class token as visualized in Figure 5. Before reaching the projection layers, we split the class token. Subsequently, the patch tokens undergo processing within a depth-wise convolutional (DW-Conv) integrated feedforward network, which includes a shortcut. The resulting output patch tokens are then subjected to averaging, producing a weight vector referred to as W. Following the squeeze-excitation operation, the output weight vector is channel-wise multiplied with the class token. This recalibrated class token is then joined with the output patch tokens to reconstruct the token. In DTU-Net, we integrate skip connections, much like those employed in the UNet architecture, into the diffusion models, establishing connections between shallow and deep layers. The primary goal is to furnish pixel-level information, which is particularly sensitive to fine-grained features. Consequently, the incorporation of extensive skip connection shortcuts amplifies feature communication and preserves the fidelity of pixel-level details. Additionally, DTU-Net employs a Conv(3, 3) block before predicting the noise. This step is intended to mitigate artifacts that may arise in images due to the attention mechanism.



(a) Full pipeline of DTU-Net as a backbone for diffusion models.

(b) Pipeline of Dynimical Multi-Head attention

Figure 4. The complete DTU-Net architecture designed for partial diffusion models, where it processes noisy image inputs, including diffusion steps as tokens, and predicts the noise.

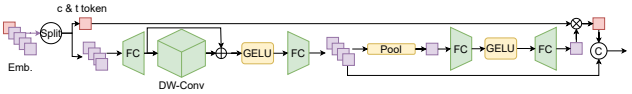


Figure 5. Hybrid Feed Forward network integrated as a Multi-Layer Perceptron for DTU-Net pipeline.

4. Experiments

All experiments in this study are conducted using the DDPM algorithm as the foundation. For DTU-Net, the hyperparameters used to approximate η_θ closely resemble those in the ViT model outlined in [27]. The model is implemented using PyTorch and trained on a single GPU, specifically the NVIDIA RTX A4000. For the image settings, we used images resized to 224×224 pixels and a batch size of 32. The training process ran for 3000 epochs with a time step of 1000. We used 1 or 3 channels, a cosine schedule for the beta parameter, and an $l2$-norm loss type. The learning rate was set to 1×10^{-4}. The patch size was 16, the embedding dimension was 384, the model depth was 6, the number of attention heads was 6, and the MLP ratio was 4. The number of classes was either null or 2, and the exponential moving average (EMA) rate was 0.9999. For the Tsimplex parameters, we used an octave value of 6, a frequency of 64, and a persistence of 0.9. We train only in healthy images with the goal of repairing the anomaly. We compute the $(\{anomaly - repaired\}\text{image})^2$ followed by binarization for testing the method on segmentation tasks using a variety of segmentation measures [3].

4.1. Datasets

Brain MRI: We utilize the healthy brain dataset sourced from the NFBS repository [32]. This dataset comprises T1-weighted MRI scans with dimensions of $256 \times 256 \times 192$. For our experiments, we focus on $2D$ slices of size 256×192 in the axial plane. Specifically, we allocate 100 of these slices for training purposes and reserve 25 for testing the algorithms. For anomaly detection, we curate a set of 154 tumor images from Kaggle, deliberately choosing a diverse range to pose a challenging task in tumor detection. In Figures 1a and 1b, we exclusively showcase images from this tumor dataset, highlighting the substantial variations compared to the healthy brain dataset.

MVTec: To train our model on typical inference data, we employ a MVTec dataset [4] comprisingcomprises 15 categories with 3629 images for normal images. For the testing

[3] More details are provided in supplementary

phase, we evaluate the model's performance on abnormal inferences consists of 1725 images, which involve various anomalies such as color variations, cuts, folds, glue marks, and punctures.

AnnoBrats: From the BRATS 2021 (Brain Tumor Segmentation) dataset [2, 19], we initially preprocess it into healthy and anomaly datasets using segmentation masks for model training and testing, respectively. We select the top 1306 (40%) 2D slices with dimensions of $4 \times 240 \times 155$, utilizing all four modalities, as anomalies are more discernible in this perspective. For testing, we employ the top 1935 (60%) 2D slices with dimensions of $4 \times 240 \times 155$, along with segmentation masks.

4.2. Results

To evaluate our method, we segment unhealthy/anomalous regions in the test dataset and employ segmentation metrics, including the Sørensen–Dice coefficient (Dice), Intersection over Union (IOU), Precision, and Recall. The results for comparison are presented in Table 1. Additionally, we conduct an area under the curve (AUC) for state-of-the-art comparison, as depicted in Table 2.In autoencoders (AE) and VAE, we utilize architectures similar to those in [36]. For the diffusion-based models, we first identify the optimal diffusion step range for anomaly detection and subsequently compute the results. As illustrated in 1, DTU-Net outperforms other methods and exhibits lower deviation compared to AnnoDDPM, though it does exhibit slightly higher deviation compared to DDPM. DTU-Net leverages sampling from Tsimplex, which exhibits fewer stochastic patterns than simplex, owing to the inclusion of batch sampling. In contrast, DDPM samples from Gaussian noise, which has fewer stochastic patterns[4].

4.3. Ablation studies

Noise functions: Based on our experiments, we have observed a slight decrease in sample quality as the diffusion step increases. This phenomenon is likely attributable to the asymmetry in Tsimplex noise. We use structural similarity index measure (SSIM) to compare the compare quality of reconstruction which is shown in Figure 6. Tsimplex gives a better SSIM than other noise functions.

[4]Please see the supplementary material for more detailed discussion and further results on objectives

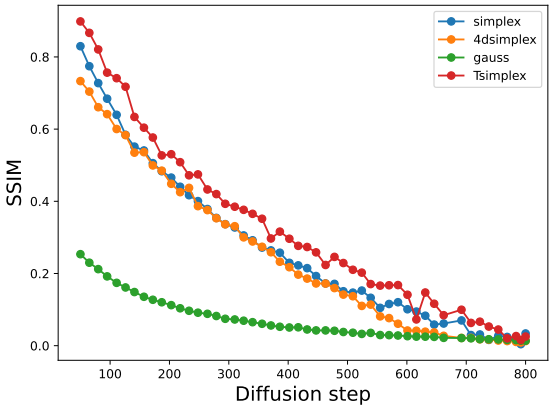


Figure 6. Effect of diffusion steps on SSIM with the backbone DTU-Net and varity of noise functions.

Effect of Diffusion Steps: The choice of diffusion steps (t) stands as a pivotal parameter in the simplex noise diffusion model for anomaly detection. In our devised approach, labeled the partial diffusion model (PDM), we strategically circumvent unnecessary diffusion steps following anomaly repair. To investigate the impact of this parameter, we conducted a series of experiments encompassing various diffusion models, each with distinct time steps ranging from 0 to 800, all grounded in simplex noise, as illustrated in Figure 7.

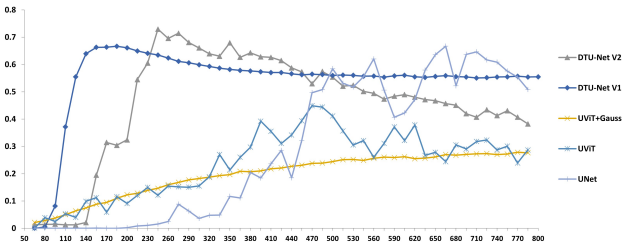


Figure 7. Diffusion steps range selection based on best Dice score for all simplex noise based model.

Additionally, we implemented a strategy aimed at reducing the stochasticity of the noise function by averaging the outputs of n-samples during the training of the model, denoted as DTU-Net V1. All instances of the PDM involve the careful selection of the optimal range of diffusion steps, maximizing the Dice score across all experiments. It is evident from the results that our proposed models, DTU-Net V1 and DTU-Net V2, achieve the highest Dice scores with fewer steps. Furthermore, the stochastic nature is notably reduced in DTU-Net V1 compared to the model without averaging during training (DTU-Net V2). The Dice score also improve robustness in DTU-Net V1, not decline as observed in DTU-Net V2.

Backbone Configuration: To assess the impact of all

Table 1. Performance Comparison of the model's ability to segment abnormal regions. Square error is employed as a predictor of the mask. We use the same architecture of VAE, and AE based on ResNet. Results in the last three rows use Tsimplex noise. The best results are highlighted.

(a) Brain MRI

Model	Dice (↑)	IOU (↑)	Recall (↑)	Precision (↑)
AE [48]	0.098±0.01	0.063±0.04	0.109±0.09	0.130±0.04
VAE [16]	0.111±0.05	0.060±0.03	0.113±0.02	0.132±0.04
CE [28]	0.242±0.20	0.152±0.14	0.25±0.218	0.275±0.22
AnnoGAN [40]	0.140±0.00	0.108±0.00	0.380±0.01	0.018±0.01
DDPM	0.017±0.00	0.006±0.01	0.013±0.02	0.042±0.02
AnnoDDPM [49]	0.334±0.29	0.243 ±0.23	0.607±0.45	0.263±0.25
DTU-Net	**0.466±0.12**	**0.364±0.18**	**0.705±0.13**	**0.383±0.19**

(b) AnnoBrats2021

Model	Dice (↑)	IOU (↑)	Recall (↑)	Precision (↑)
AE [48]	0.015±0.04	0.063±0.02	0.125±0.02	0.112±0.04
VAE [16]	0.016±0.04	0.05±0.02	0.115±0.02	0.115±0.04
CE [28]	0.239±0.19	0.154±0.14	0.245±0.21	0.265±0.22
AnnoGAN [40]	0.135±0.00	0.098±0.00	**0.384±0.00**	0.085±0.00
DDPM [17]	0.010±0.01	0.004±0.00	0.007±0.01	0.036±0.04
AnnoDDPM [49]	0.334±0.25	0.146±0.13	0.083±0.18	0.042±0.14
DTU-Net	**0.428±0.13**	**0.280±0.17**	0.210±0.13	**0.299±0.12**

Table 2. MvTec Dataset result on 15 objects by AUROC for state-of-the-art comparison.

Classes	PatchCore [37]	PaDiM [8]	SimpleNet [23]	DMAD [21]	DRAEM [51]	AnnoGAN [40]	FastRecon [13]	AnnoDDPM [49]	DDPM	Ours
Carlet	98.85	99.10	98.71	99.56	96.27	91.63	92.58	98.63	51.10	99.15
Grid	98.34	97.30	98.33	99.61	99.80	94.09	90.71	96.65	51.46	99.85
Leather	99.66	99.20	98.67	99.76	99.32	97.39	94.60	96.39	51.51	99.90
Tile	97.20	94.10	98.20	98.06	99.41	81.78	79.12	97.94	51.46	99.80
Wood	97.17	94.90	96.66	97.82	97.79	93.69	93.27	97.84	49.91	98.35
Bottle	99.32	98.30	97.10	99.47	99.15	99.54	96.47	96.09	51.15	99.41
Cable	98.97	96.70	99.32	98.62	93.20	93.01	94.25	98.52	50.63	98.87
Capsule	98.44	98.50	98.74	98.61	96.47	96.67	98.33	98.13	50.74	98.61
Huelnut	99.37	98.20	99.32	99.56	99.85	99.26	98.85	99.43	51.46	99.85
Metal nut	99.23	97.20	98.55	98.89	99.09	95.29	94.33	99.02	51.46	99.85
Pill	96.99	95.70	99.42	97.98	98.27	97.79	95.24	95.54	51.25	99.15
Screw	98.73	98.50	99.03	99.81	95.69	97.70	98.57	99.40	51.10	99.21
Toothbrush	99.37	98.80	98.63	99.71	99.08	99.02	96.47	96.23	50.99	99.42
Transistor	98.21	97.50	98.94	97.10	92.04	89.55	91.10	96.96	47.95	96.46
Ziplrr	99.11	98.50	99.61	98.97	99.42	83.65	80.03	96.59	51.20	99.61
Average	98.60	97.39	98.62	98.90	97.66	94.00	92.93	97.56	50.89	**99.17**

the modifications made to DTU-Net, in addition to the ViT model, we conducted a series of experiments involving various backbone variations. We evaluated these variations based on Dice and AUC scores, which are summarized in Table 3. The Leather dataset was chosen for this comparison due to its inherent difficulty in segmenting anomalies.As depicted in Figure 4a, we examined four types of variations, including changes in Patch Embedding (PE and SPE), Multi-Head Attention (MHA and DMHA), Multi-Layer Perceptron (MLP and HFF), and Refinement Layer (Conv(3,3) and an additional Conv(3,3) layer for output refinement). Among these configurations, DTU-Net with the setup of PE+DMHA+MLP+Conv outperformed others in terms of AUC scores. However, the PE+DMHA +MLP+Conv configuration achieved superior results in Dice scores.

Weakly Supervised Segmentation: We explore three different backbone models for the backward diffusion process: UNet, UViT, DiT [9], and our proposed DTU-Net, in conjunction with the Tsimplex noise function. Guided diffusion models [9, 29] have demonstrated superior performance when trained in a weakly supervised (WS) manner, such as by using image-level labels. We trained our model using image-level labels to evaluate its effectiveness, and the results are presented in Table 4, specifically in the

Table 3. Comparison of DTU-Net variations with DICE and AUC scores.

DTU-Net Variations	Dice	AUC
PE+MHA+MLP+Conv	0.293±0.230	0.640±0.092
SPE+MHA+MLP+Conv	0.140±0.208	0.621±0.099
PE+DMHA+MLP+Conv	0.268±0.232	**0.757±0.131**
SPE+DMHA+MLP+Conv	0.147±0.204	0.711±0.070
PE+DMHA+HFF+Conv	0.163±0.177	0.706±0.110
SPE+DMHA+HFF+Conv	0.181±0.128	0.667±0.092
SPE+DMHA+HFF+RConv	**0.306±0.234**	0.636±0.137

shaded columns (G) for the BrainMRI dataset. Furthermore, To illustrate the overall impact of our modifications on baseline methods and to compare with WS tasks, we also report results for the same dataset, which includes highly challenging abnormal images (see Figure 1a). Our experiments indicate that Tsimplex generates better samples for recovering or repairing anomalies, even when used with the UNet and UViT models. The DTU-Net model combined with Tsimplex surpasses other combinations, primarily due to configuration modifications (refer to Table 3). Additionally, we observe superior performance compared to DiT [29] in both DICE and AUROC measures for image-level labels.

Table 4. Impact of modifications on anomaly detection performance in the BrainMRI dataset, measured by DICE/AUROC scores. Shaded columns (G) represents the training with image-lavel.

Noise(↓)/Model(→)	UNet	UViT	DTU-Net	DiT(G)	DTU-Net(G)
Gauss	0.02/0.64	0.10/0.69	0.08/0.61	0.13/0.66	0.18/0.65
Simplex	0.33/0.68	0.34/0.70	0.40/0.74	0.41/0.75	0.41/0.77
Tsimplex	0.35/0.68	0.39/0.73	0.46/0.75	0.42/0.76	0.50/0.78

5. Conclusion

Our self-supervised anomaly detection model, which leverages partially observed diffusion steps to significantly reduce sampling time, has shown remarkable effectiveness. The combination Tsimplex diffusion and DTU-Net, a ViT-based backbone, has not only enabled our method to generate high-quality anomaly maps but has also led to achieving good scores for anomaly segmentation across three different image datasets without the need for annotations.

References

[1] Yuki Markus Asano, Christian Rupprecht, and Andrea Vedaldi. Self-labelling via simultaneous clustering and representation learning. *arXiv preprint arXiv:1911.05371*, 2019. 3

[2] Ujjwal Baid, Satyam Ghodasara, Suyash Mohan, Michel Bilello, Evan Calabrese, Errol Colak, Keyvan Farahani, Jayashree Kalpathy-Cramer, Felipe C Kitamura, Sarthak Pati, et al. The rsna-asnr-miccai brats 2021 benchmark on brain tumor segmentation and radiogenomic classification. *arXiv preprint arXiv:2107.02314*, 2021. 7

[3] Fan Bao, Shen Nie, Kaiwen Xue, Yue Cao, Chongxuan Li, Hang Su, and Jun Zhu. All are worth words: A vit backbone for diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22669–22679, 2023. 2, 3

[4] Paul Bergmann, Kilian Batzner, Michael Fauser, David Sattlegger, and Carsten Steger. The mvtec anomaly detection dataset: a comprehensive real-world dataset for unsupervised anomaly detection. *International Journal of Computer Vision*, 129(4):1038–1059, 2021. 6

[5] Sam Bond-Taylor, Adam Leach, Yang Long, and Chris G Willcocks. Deep generative modelling: A comparative review of vaes, gans, normalizing flows, energy-based and autoregressive models. *IEEE transactions on pattern analysis and machine intelligence*, 2021. 1, 3

[6] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in neural information processing systems*, 33:9912–9924, 2020. 3

[7] Snehashis Chakraborty, Komal Kumar, Balakrishna Pailla Reddy, Tanushree Meena, and Sudipta Roy. An explainable ai based clinical assistance model for identifying patients with the onset of sepsis. In *2023 IEEE 24th International Conference on Information Reuse and Integration for Data Science (IRI)*, pages 297–302, 2023. 1

[8] Thomas Defard, Aleksandr Setkov, Angelique Loesch, and Romaric Audigier. Padim: a patch distribution modeling framework for anomaly detection and localization. In *International Conference on Pattern Recognition*, pages 475–489. Springer, 2021. 8

[9] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021. 3, 8

[10] NC Dlova, A Chateau, N Khoza, A Skenjane, Z Mkhize, OS Katibi, A Grobler, JT Gwegweni, and A Mosam. Prevalence of skin diseases treated at public referral hospitals in kwazulu-natal, south africa. *British journal of dermatology*, 178(1):e1–e2, 2018. 1

[11] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 3, 5

[12] Stéphane d'Ascoli, Hugo Touvron, Matthew L Leavitt, Ari S Morcos, Giulio Biroli, and Levent Sagun. Convit: Improving vision transformers with soft convolutional inductive biases. In *International Conference on Machine Learning*, pages 2286–2296. PMLR, 2021. 3

[13] Zheng Fang, Xiaoyang Wang, Haocheng Li, Jiejie Liu, Qiugui Hu, and Jimin Xiao. Fastrecon: Few-shot industrial anomaly detection via fast feature reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 17481–17490, 2023. 8

[14] Alvaro Gonzalez-Jimenez, Simone Lionetti, Marc Pouly, and Alexander A Navarini. Sano: Score-based diffusion model for anomaly localization in dermatology. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2987–2993, 2023. 1

[15] Hao Guan and Mingxia Liu. Domain adaptation for medical image analysis: A survey. *IEEE Transactions on Biomedical Engineering*, 69:1173–1185, 2021. 1

[16] Kuan Han, Haiguang Wen, Junxing Shi, Kun-Han Lu, Yizhen Zhang, Di Fu, and Zhongming Liu. Variational autoencoder: An unsupervised model for encoding and decoding fmri activity in visual cortex. *NeuroImage*, 198:125–136, 2019. 8

[17] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 1, 3, 4, 8

[18] Diederik Kingma, Tim Salimans, Ben Poole, and Jonathan Ho. Variational diffusion models. *Advances in neural information processing systems*, 34:21696–21707, 2021. 3

[19] Komal Kumar, Snehashis Chakraborty, and Sudipta Roy. Self-supervised diffusion model for anomaly segmentation in medical imaging. In Pradipta Maji, Tingwen Huang, Nikhil R. Pal, Santanu Chaudhury, and Rajat K. De, editors, *Pattern Recognition and Machine Intelligence*, pages 359–368, Cham, 2023. Springer Nature Switzerland. 3, 7

[20] Komal Kumar, Balakrishna Pailla, Kalyan Tadepalli, and Sudipta Roy. Robust msfm learning network for classification and weakly supervised localization. In *2023 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, Paris, France, October 2, 2023 2023. IEEE. 1

[21] Wenrui Liu, Hong Chang, Bingpeng Ma, Shiguang Shan, and Xilin Chen. Diversity-measurable anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12147–12156, 2023. 8

[22] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 5

[23] Zhikang Liu, Yiming Zhou, Yuansheng Xu, and Zilei Wang. Simplenet: A simple network for image anomaly detection and localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20402–20411, 2023. 8

[24] Cheng Lu, Kaiwen Zheng, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Maximum likelihood training for score-based diffusion odes by high order denoising score matching. In *International Conference on Machine Learning*, pages 14429–14460. PMLR, 2022. 3

[25] Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. *Advances in Neural Information Processing Systems*, 35:5775–5787, 2022. 3

[26] Zhiying Lu, Hongtao Xie, Chuanbin Liu, and Yongdong Zhang. Bridging the gap between vision transformers and convolutional neural networks on small datasets. *Advances in Neural Information Processing Systems*, 35:14663–14677, 2022. 2, 5

[27] Namuk Park and Songkuk Kim. How do vision transformers work? *arXiv preprint arXiv:2202.06709*, 2022. 3, 6

[28] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2536–2544, 2016. 8

[29] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4195–4205, 2023. 3, 8

[30] Ken Perlin. Improving noise. In *Proceedings of the 29th annual conference on Computer graphics and interactive techniques*, pages 681–682, 2002. 2, 4

[31] 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. 1

[32] 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. 6

[33] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015. 3

[34] Maithra Raghu, Thomas Unterthiner, Simon Kornblith, Chiyuan Zhang, and Alexey Dosovitskiy. Do vision transformers see like convolutional neural networks? *Advances in Neural Information Processing Systems*, 34:12116–12128, 2021. 3

[35] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 3

[36] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015. 7

[37] Karsten Roth, Latha Pemula, Joaquin Zepeda, Bernhard Schölkopf, Thomas Brox, and Peter Gehler. Towards total recall in industrial anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14318–14328, 2022. 8

[38] Daniel L. Ruderman. Origins of scaling in natural images. *Vision Research*, 37(23):3385–3398, 1997. 4

[39] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022. 3

[40] Thomas Schlegl, Philipp Seeböck, Sebastian M Waldstein, Ursula Schmidt-Erfurth, and Georg Langs. Unsupervised anomaly detection with generative adversarial networks to guide marker discovery. In *International conference on information processing in medical imaging*, pages 146–157. Springer, 2017. 3, 8

[41] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *In 9th International Conferenceon Learning Representations*, 2021. 3

[42] Yang Song, Conor Durkan, Iain Murray, and Stefano Ermon. Maximum likelihood training of score-based diffusion models. *Advances in Neural Information Processing Systems*, 34:1415–1428, 2021. 3

[43] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32, 2019. 3

[44] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020. 3

[45] Arash Vahdat, Karsten Kreis, and Jan Kautz. Score-based generative modeling in latent space. *Advances in Neural Information Processing Systems*, 34:11287–11302, 2021. 3

[46] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017. 1

[47] van A Van der Schaaf and JH van van Hateren. Modelling the power spectra of natural images: statistics and information. *Vision research*, 36(17):2759–2770, 1996. 4

[48] Chathurika S Wickramasinghe, Daniel L Marino, and Milos Manic. Resnet autoencoders for unsupervised feature learning from high-dimensional data: Deep models resistant to performance degradation. *IEEE Access*, 9:40511–40520, 2021. 8

[49] Julian Wyatt, Adam Leach, Sebastian M Schmon, and Chris G Willcocks. Anoddpm: Anomaly detection with denoising diffusion probabilistic models using simplex noise. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 650–656, 2022. 1, 3, 4, 8

[50] Raymond Yeh, Chen Chen, Teck Yian Lim, Mark Hasegawa-Johnson, and Minh N. Do. Semantic image inpainting with perceptual and contextual losses. *arXiv preprint arXiv:1607.07539*, 2016. 3

[51] Vitjan Zavrtanik, Matej Kristan, and Danijel Skočaj. Draema discriminatively trained reconstruction embedding for surface anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8330–8339, 2021. 8