Robust Self-Supervised Anomaly Detection with Diffusion Models

Anonymous CVPR submission

Paper ID *****

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 dynamical hybrid UNet vision transformer (UDHVT). 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 vision transformers, 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 Leather dataset. It achieves an average improvement of $+10.1\%$ *in Dice coefficient,* $+10.4\%$ *in IOU, and* $+9.6\%$ *in AUC.*

1. Introduction

The scarcity of experts with the ability to diagnose and treat specific medical conditions is a pressing concern in developing countries [19]. To illustrate, consider the ratio of dermatologists to the general population, which can plummet to as low as 1 per 216,000 people [14]. This motivation fuels the development of a deep learning system with the ability to localize diseases and thereby prevent misdiagnosis or underdiagnosis [10, 27]. 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 powerfull deep learning algorithms that train on healthy or normal infrance 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, image data, especially medical image data, can be quite complex. Unsupervised models struggle with sampling from complex data distributions due to their limited assumptions, inadequate representation capac-

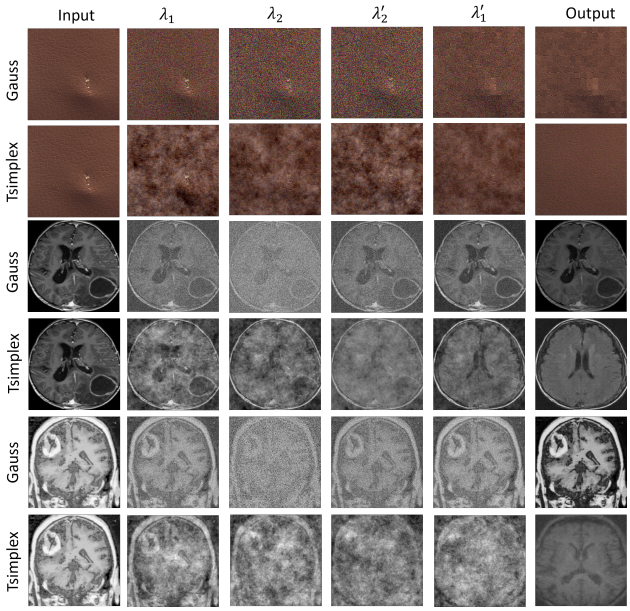


Figure 1. Forward and backward diffusion processes for anomaly detection in Leather (rows 1-2) and BrainMRI (rows 3-6) test data with a partial diffusion step (300), where $\lambda_1(=100)$ and $\lambda_2(=200)$ represent intermediate steps in the forward process, while $\lambda_1^{'}(=200)$ and $\lambda_2^{'}(=100)$ indicate diffusion steps in the backward process.

ity, risk of mode collapse, issues with high-dimensional data, sample efficiency challenges, difficulties in evaluation, and sensitivity to initialization. For tasks requiring precise sampling from highly complex data distributions, alternative, more advanced generative models needed. Generative models have demonstrated their potency in self-supervised representation learning of the underlying distribution, particularly in the context of healthy inference data [19, 35, 50, 52]. Denoising diffusion probabilistic models (DDPMs) [22] 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) [8, 50]. 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 the samples from a gaussion 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 gaussion noise at each diffusion step does not able to recover the anomaly and that results into unrepaired anomaly [52]. Gaussian noise has a constant power spectral density, meaning it has equal power across all frequencies which makes it "white" noise. The author of the paper [52] explored the use of simplex noise [34], a common choice for tasks such as procedural terrain generation, texture synthesis, and creating natural-looking patterns like clouds or marble textures. However, it was discovered that 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, 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 3D and colored images. Moreover, the model was trained using a batch size of one due to the time-complexity, which scales as $\mathcal{O}(batch\_size \times t)$, 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 three dimention and colored images while maintaining the same processing time for batched images.

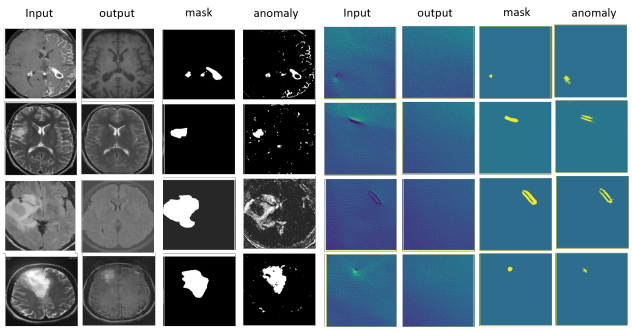


Figure 2. Abnormality reparing under the condition of larger anomaly, anomaly near to the high intensity region and coloured images by our proposed methodology.

Addtionaly, we introduce a simple and versatile Dynamic Hybrid UNet architecture inspired by vision transformers called UDHVT (Figure 3a) designed as the backbone of diffusion models. Our algorithms offer several advantages over adversarial training, including improved sample quality and stable training, particularly beneficial for smaller datasets.

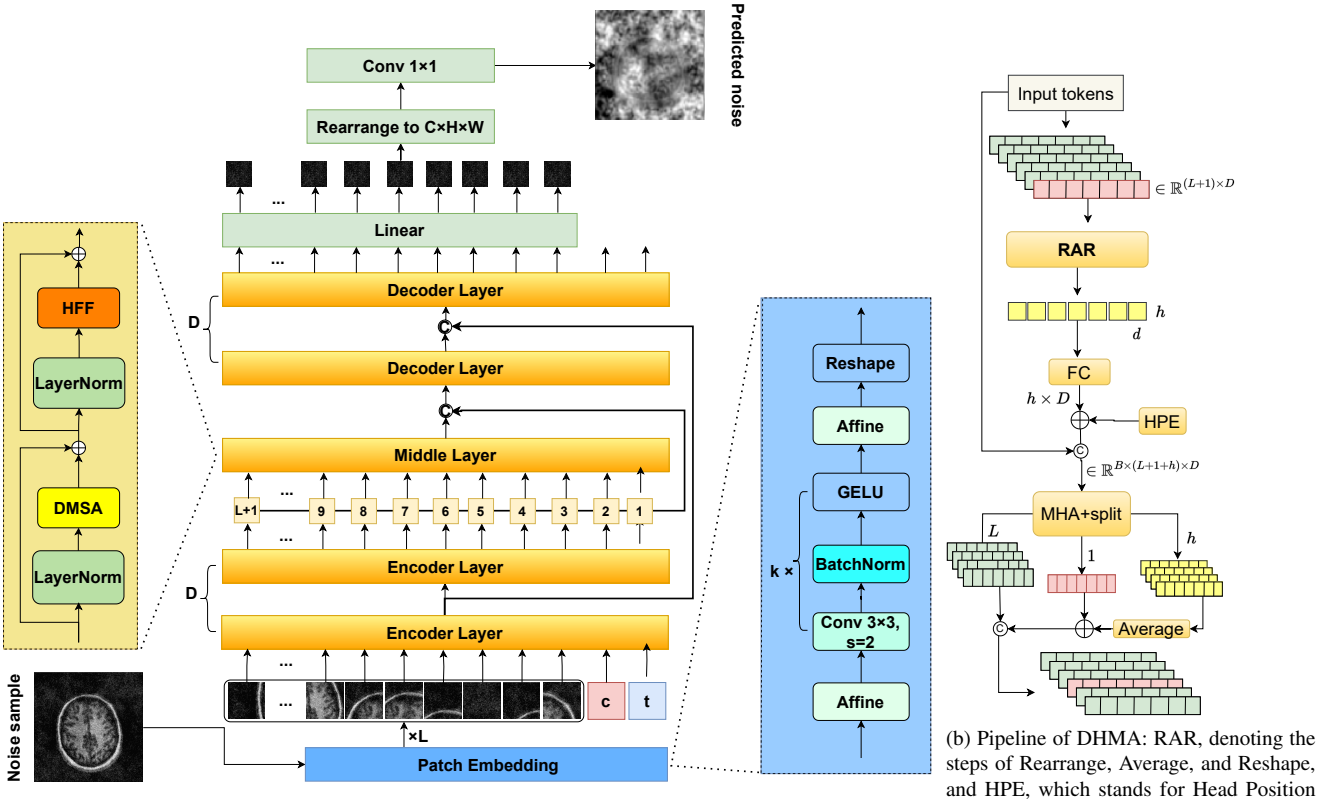
The paper makes several significant contributions, which is be summarized as follows:

- We present Tsimplex, a novel sampling strategy tailored for 3D and colored images. Tsimplex enhances the sampling process, improving the quality of the samples generated.
- We have developed a Vision Transformer-based U-net model as the backbone for diffusion models and improved the multi-head attention mechanism by integrating interactivity between the attention heads.

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 approaches [2]: Invariance-based methods [1, 2, 5, 8, 9] and generative methods [3, 12, 13, 21–23, 33, 52]. 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.

Variational Autoencoders (VAEs) [25], for instance, have garnered acclaim for their ability to generate data while also providing interpretable latent representations. However, VAEs do encounter challenges when tasked with generating high-quality, sharp images. Conversely, Generative Adversarial Networks (GANs) [20] have excelled in producing visually appealing samples, yet they can be challenging to train and may suffer from mode collapse. More recently, Flow-based models like Glow [24] have emerged, offering exact likelihood estimation and high-quality samples but tending to be computationally intensive. When selecting a generative model, it's crucial to consider the specific task requirements and the trade-offs between interpretability, image quality, and computational resources.



(a) Full pipeline of UDHVT as a backbone for diffusion models.

(b) Pipeline of DHMA: RAR, denoting the steps of Rearrange, Average, and Reshape, and HPE, which stands for Head Position Embedding for interaction with the heads.

Figure 3. The complete UDHVT architecture designed for partial diffusion models, where it processes noisy image inputs, including diffusion steps as tokens, and predicts the noise.

GANs have proven to be instrumental in learning highly detailed and realistic images [11, 16, 20]. Authors in [37] introduced Deep Convolutional Generative Adversarial Networks (DCGANs), showcasing GANs' ability to capture semantic image content, which has led to intriguing applications like vector arithmetic for manipulating visual concepts. Additionally, [53] 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 [44] come with their set of challenges, including their inability to repair anomalies, training instability, model collapse, and a reliance on large datasets.

Recent strides in the field of Diffusion Probabilistic Models (DDPM) [13, 22] have showcased their ability to generate higher-quality samples from complex distributions with superior coverage compared to GANs [44] 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 [26]. Furthermore, DDPMs also have limitations in capturing larger anomalies caused by Gaussian noise.

In their paper [52], 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 [5, 6, 15, 23, 29, 30, 45, 47–49], 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 [22, 46]. The CNN-driven U-Net design features a sequence of downsampling blocks, a series of up-sampling blocks, and extensive skip connections between these two sets of blocks

[13, 39, 40, 43]. This architectural framework has held a prominent position within diffusion models utilized for image generation assignments. Conversely, vision transformers (ViT) [17] have demonstrated promising, and in some cases, superior performance compared to CNNs in a range of tasks. In their paper [6], the authors introduce a straightforward and versatile architecture for image generation using Vision Transformers 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 [18, 32, 38] 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. The paper [31] attempts to bridge the gap between Vision Transformers and Convolutional Neural Networks as a potential solution to enhance their respective biases.

3. Methodology

3.1. Overview of Diffusion Models

Diffusion models, specifically Diffusion Probabilistic Models (DDPMs) [22], 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 normal ditribution 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 is also be seen as a mapping from a simpler Gaussian distribution to a more complex distribution of images. This mapping 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 [22] 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 [42]. 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 [51, 52]. 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 the Figure 1. 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 (1 for Tsimplex).

Tsimplex: Simplex noise represents an enhancement of Perlin noise [34], characterized by increased computational efficiency and the generation of smooth and structured randomness. It relies on gradient noise 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, yielding a more intricate and detailed noise pattern. The ultimate outcome is a textured noise that exhibits gradual and continuous variations across spatial dimensions shown in the Figure 5a. The generation of Tsimplex noise is outlined in Algorithm 1, while the pseudocode for the simplified simplex function can be found in Supplementary Material Algorithm 1. In the case of Tsimplex noise, we

follow similar steps as detailed in [52], but with additional exploration for three-dimensional and colored images. Notably, the total time required to sample noise from Tsimplex for a batch of 1000 steps is approximately 5.97 seconds, which is less than simplex noise and close to Gaussian noise, as demonstrated in Figure 4. This improvement is attributed to the Honeycomb pattern introduced by the additional time step parameter, $t \in \mathbb{R}^{batch\_size \times num\_step}$. 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 5c, in contrast to Gaussian noise. In contrast, Tsimplex provides similar sample quality to Gaussian noise, as illustrated in Figure 5b. For a more in-depth examination of the symmetry of simplex noise, please refer to SUPPLEMENTARY MATERIAL.

Algorithm 1 GenNoise(S, t, $O = 8$, $p = 0.9$, $\mu = 128$)

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

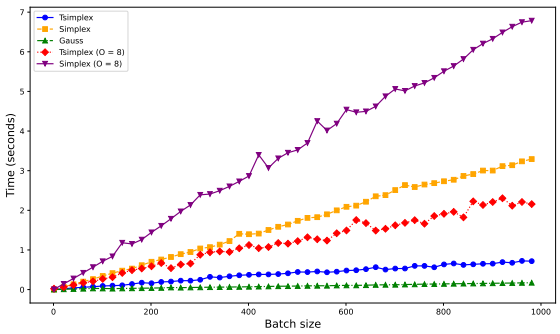


Figure 4. Time Required for sampling upto 1000 steps for grid of 64×64 and O represents number of octave

3.3. Dynimic Hybrid UNet

UDHVT serves as a fundamental component in diffusion models (see Figure 3a) for anomaly detection and can be applied across a variety of tasks within diffusion modeling. The primary goal of UDHVT is to minimize the loss as defined in equation 4 which combine $L2-norm$ with L_{vlb} for the robust self-supervised traning and generate predictions 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 [17], UDHVT 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 the Figure 3a and some function can formulated as follows.

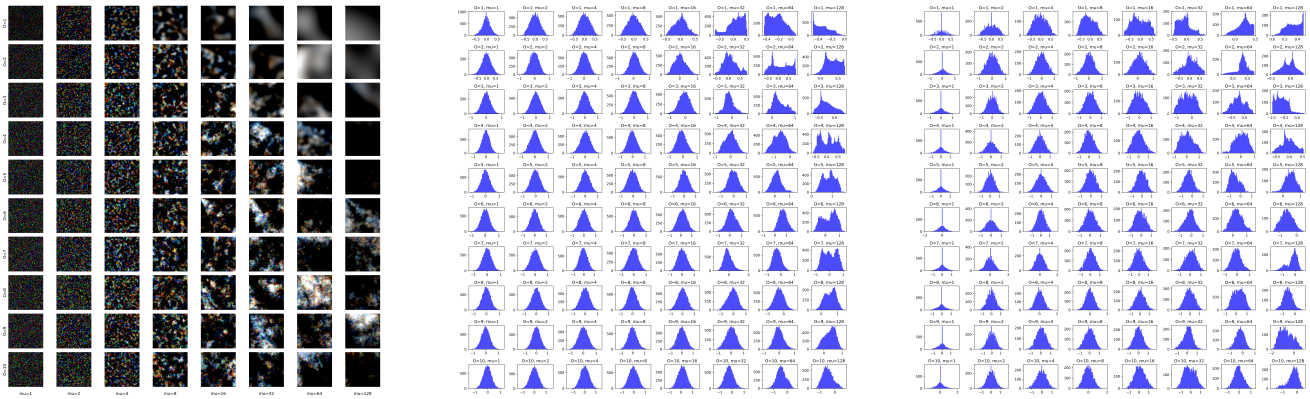
$$\text{Affine}(x) = \text{Diag}(\nu)x + \phi \tag{5}$$

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 ($L = \frac{H \times W}{P^2}$).

UDHVT Layer: Following the architecture of UNet, UDHVT 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 3a. These blocks primarily include dynimical multi-head attention (DMHA) (refer to Figure 3b) and a hybrid feed forward (HFF) block (see Figure 6). Inspired by [31], 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 vision transformer with window attention [28].

DMHA: The mechanism is formulated as follows: For $x \in \boldsymbol{R}^{b \times L \times D}$, 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 head will not be forgeted. We concatinate the generated head token with shortcut input and feed to the multi-head attetion. Finlly output is transformed into original shape by splitting into $L \times D$, $1 \times D$, and $h \times D$ as shown in the Figure 3b.

HFF: We employ a channel attention mechanism to consolidate the features of patch tokens into the class token as visualized in Figure 6. Before reaching the projection layers, we split the class token. Subsequently, the patch tokens undergo processing within a depth-wise convolutional (DW-Conv) integrated feed forword 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 UDHVT, we integrate skip connections, much like those employed in the Unet architecture, into the diffusion



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

Figure 5. Comparing Tsimplex and Simplex noise, we analyze the impact of two variables: octave (N) on the y-axis and frequency (μ) on the x-axis, examining their influence on both distribution and structure.

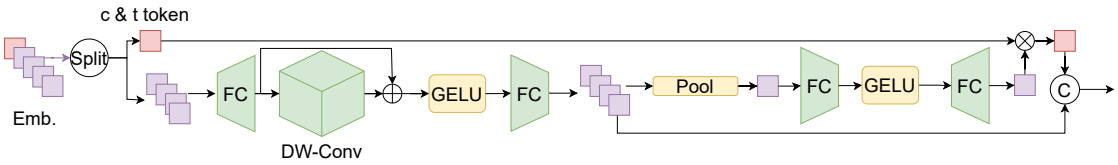


Figure 6. Pipeline of Hybrid Feed Forward network.

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, UDHVT 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.

4. Experiments

4.1. Implementation details

All experiments in this study are conducted using the DDPM algorithm as the foundation. For DHUVT, the hyperparameters used to approximate η_θ closely resemble those in the ViT model outlined in [32]. We employ the hyperparameters detailed in SUPPLEMENTARY Table 3. The model is implemented using PyTorch and trained on a single GPU, specifically the NVIDIA RTX A4000.

4.2. Datasets

Brain MRI: We utilize the healthy brain dataset sourced from the NFBS repository [36]. 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 1 and 2, we exclusively showcase images from this tumor dataset, highlighting the substantial variations compared to the healthy brain dataset.

Lether: To train our model on typical inference data, we employ a dataset comprising 245 normal images. For the testing phase, we evaluate the model's performance on abnormal inferences, which involve various anomalies in leather such as color variations, cuts, folds, glue marks, and punctures. These anomalies are sourced from the MVTec dataset [7].

AnnoBrats: From the BRATS 2021 (Brain Tumor Segmentation) dataset [4, 26], 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.

4.3. 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) comparison, as depicted in Figure 7.

We explore three different backbone models for the backward diffusion process: UNet, UViT, and our proposed DHUVT, in conjunction with the Tsimplex noise function. In autoencoders (AE) and VAE, we utilize architectures similar to those in [41]. For the diffusion-based models, we first identify the optimal diffusion step range for anomaly detection and subsequently compute the results.

As illustrated in Figure 7, UDHVT outperforms other methods and exhibits lower deviation compared to AnnoDDPM, though it does exhibit slightly higher deviation compared to DDPM. UDHVT 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. Additionaly,

We also showcase the results of sampling from Tsimplex and Gaussian noise on the BrainMRI and leather (channels = 3) subset of the MVTec AD dataset, demonstrating excellent healthy reconstructions in Figure 1 and 2.

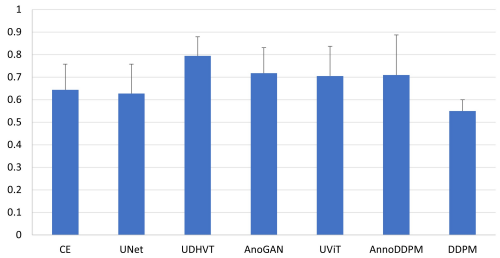


Figure 7. Comparision with the AUC matric on BrainMRI dataset for CE [33], AnnoGAN [44], AnnoDDPM [52], DDPM [22] with gaussion noise, and some backbones, including UNet [13], UViT [6], and prposed UDHVT with Tsimplex.

4.4. 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 the quality of recontrction which is shown in the Figure 8. Tsimplex is giving the better SSIM than other noise function.

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

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 same architecture of VAE, autoencoders based on ResNet.

(a) Brain MRI

Model	Dice (↑)	IOU (↑)	Recall (↓)	Precision (↑)
Autoencoder	0.098±0.015	0.0632 ± 0.043	0.109±0.093	0.130±0.041
VAE	0.111±0.051	0.060±0.033	0.113±0.023	0.132±0.041
Context Encoder	0.242±0.207	0.152±0.14	0.252±0.218	0.276±0.227
F-AnnoGAN	0.140±0.00	0.108±0.001	0.38±0.011	0.018±0.010
DDPM	0.017±0.001	0.006±0.016	0.013±0.021	0.042±0.025
AnoDDPM	0.334 ± 0.299	0.243 ±0.237	0.607 ± 0.452	0.263 ± 0.257
UNet (Backbone)	0.346 ± 0.224	0.234 ± 0.183	0.313 ± 0.241	**0.491 ± 0.253**
UViT (Backbone)	0.418±0.315	0.205±.174	0.297±0.253	0.421 ± 0.234
UDHVT (Ours Backbone)	**0.466 ± 0.129**	**0.364 ± 0.186**	**0.705 ± 0.131**	0.383 ± 0.197

(b) AnnoBrats2021

Model	Dice (↑)	IOU (↑)	Recall (↓)	Precision (↑)
Autoencoder	0.015±0.047	0.063±0.028	0.125±0.028	0.112±0.042
VAE	0.016±0.048	0.05±0.028	0.115±0.029	0.115±0.043
Context Encoder	0.239±0.199	0.154±0.142	0.245±0.212	0.265±0.222
F-AnnoGAN	0.135±0.001	0.098±0.003	**0.384±0.009**	0.085±0.003
DDPM	0.010±0.013	0.004±0.008	0.007±0.010	0.036±0.049
AnoDDPM	0.334±0.253	0.146±0.130	0.083±0.185	0.042±0.148
UNet (Backbone)	0.371±0.281	0.162±0.144	0.092±0.206	0.047±0.163
UViT (Backbone)	0.409±0.309	0.179±0.158	0.101±0.227	0.051±0.179
UDHVT (Ours Backbone)	**0.428±0.131**	**0.280±0.178**	0.210±0.131	**0.299±0.124**

(c) Leather

Model	Dice (↑)	IOU (↑)	Recall (↑)	Precision (↓)
Autoencoder	0.003±0.062	0.002±0.010	0.005±0.082	0.004±0.182
VAE	0.003±0.063	0.002±0.010	0.005±0.084	0.004±0.186
Context Encoder	0.141 ± 0.164	0.085 ± 0.111	0.531 ± 0.226	0.207 ± 0.323
DDPM	0.016 ± 0.020	0.008 ± 0.010	0.466 ± 0.250	0.008 ± 0.011
AnoDDPM	0.228±0.213	0.147±0.157	0.509±0.283	**0.427±0.410**
UNet (Backbone)	0.237±0.023	0.153±0.217	0.536±0.130	0.444±0.101
UViT (Backbone)	0.291± 0.244	0.1987±0.201	0.295 ±0.262	0.548±0.403
UDHVT (Ours Backbone)	**0.306±0.134**	**0.205± 0.138**	**0.585±0.165**	0.298±0.126

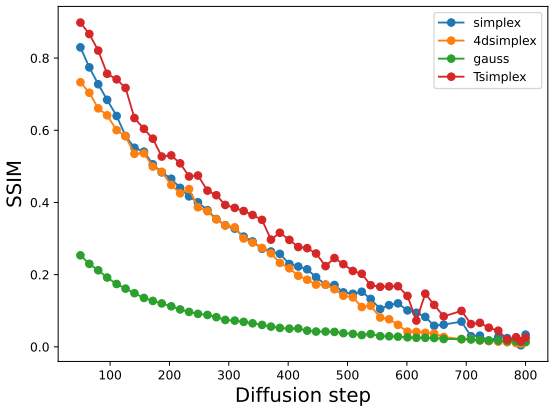


Figure 8. Effect of diffusion steps on SSIM with the backbone UDHVT and varity of noise functions.

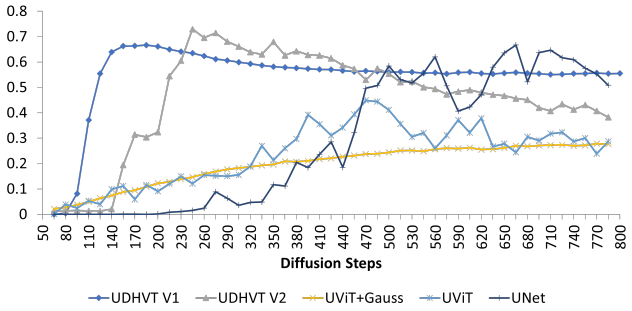


Figure 9. Diffusion steps range selection based on best Dice score for all simplex noise based model.

800, all grounded in simplex noise, as illustrated in Figure 9. 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 UDHVT 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, UDHVT V1 and UDHVT V2, achieve the highest Dice scores with fewer steps. Furthermore, the stochastic nature is notably reduced in UDHVT V1 compared to the model without averaging during training (UDHVT V2). The Dice score also improve robustness in UDHVT V1, not decline as observed in UDHVT V2.

Backbone Configration: To assess the impact of all the modifications made to UDHVT, 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 2. The Leather dataset was chosen for this comparison due to its inherent difficulty in segmenting anomalies, as highlighted in Table 1c. As depicted in Figure 3a, 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, UDHVT 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.

Table 2. Comparison of variations with DICE and AUC scores.

UDHVT 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

5. Discussion

In the realm of anomaly detection, a full Markov chain is not necessarily required. However, when it comes to generating high-quality images, the full Markov chain becomes an essential component. Our observations reveal that as the number of diffusion steps increases, Tsimplex exhibits a decrease in sample quality, as depicted in Figure 8. To address this issue, a multi-section noise can specifically designed to incorporate multi-frequency noise functions, aiming to improve sampling quality and subsequently enhance the Structural Similarity Index (SSIM) which can be a versatile solution, offering potential avenues to enhance sample quality and tackle the asymmetry inherent in noise patterns. The presence of zig-zag patterns in the results can be attributed to the stochastic nature of Tsimplex noise. To mitigate this, we adopted a strategy of multiple sampling and the averaging of reconstructions, resulting in a smoother Dice score graph, as illustrated in Figure 9 (UDHVT V1). It is worth considering the integration of such strategies into the training of diffusion models in the future. Furthermore, UDHVT, as the backbone for diffusion models, has been designed to acquire a generalized self-supervised representation of healthy images. Looking ahead, this foundational structure could potentially be further explored for the generation of high-quality images, akin to the capabilities of a UNet architecture. The methodology presented here exhibits robustness when dealing with unannotated data. In the future, we envision its potential to harness additional information in three-dimensional (3D) images, particularly when coupled with the developed Tsimplex noise. This expansion into 3D data could offer exciting prospects for enhancing anomaly detection and image generation.

6. Conclusion

In this study, we develop a self-supervised anomaly detection model using partially observed diffusion steps. We incorporate simplex noise diffusion to control anomaly size and introduce Tsimplex sampling techniques for 3D and color images, reducing noise function stochasticity through output averaging. Additionally, we present UDHVT, a versatile backbone for diffusion models by incorporating the dynamical head intreted multi-head attention in vision transformers. Our method is applied to three diverse image datasets, delivering high-quality results through self-supervised learning without the need for annotated data. We aim to extend this methodology to various medical anomaly detection challenges in the future, as we believe that UDHVT holds promise for advancing backbone research in dif-

fusion models and enhancing generative modeling.

References

[1] Yuki Markus Asano, Christian Rupprecht, and Andrea Vedaldi. Self-labelling via simultaneous clustering and representation learning. *arXiv preprint arXiv:1911.05371*, 2019. 2

[2] Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15619–15629, 2023. 2

[3] Alexei Baevski, Wei-Ning Hsu, Qiantong Xu, Arun Babu, Jiatao Gu, and Michael Auli. Data2vec: A general framework for self-supervised learning in speech, vision and language. In *International Conference on Machine Learning*, pages 1298–1312. PMLR, 2022. 2

[4] 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. 6

[5] Fan Bao, Chongxuan Li, Jiacheng Sun, Jun Zhu, and Bo Zhang. Estimating the optimal covariance with imperfect mean in diffusion probabilistic models. *arXiv preprint arXiv:2206.07309*, 2022. 2, 3

[6] 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. 3, 4, 7

[7] 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

[8] 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. 2

[9] 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. 2

[10] 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

[11] Emily L. Denton, Soumith Chintala, Rob Fergus, and et al. Deep generative image models using a laplacian pyramid of adversarial networks. In *Advances in Neural Information Processing Systems*, pages 1486–1494, 2015. 3

[12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 2

[13] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021. 2, 3, 4, 7

[14] 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

[15] Tim Dockhorn, Arash Vahdat, and Karsten Kreis. Score-based generative modeling with critically-damped langevin diffusion. *arXiv preprint arXiv:2112.07068*, 2021. 3

[16] Jeff Donahue, Philipp Krähenbühl, and Trevor Darrell. Adversarial feature learning. *arXiv preprint arXiv:1605.09782*, 2016. 3

[17] 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. 4, 5

[18] 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. 4

[19] 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, 2

[20] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, pages 2672–2680, 2014. 2, 3

[21] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022. 2

[22] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 2, 3, 4, 7

[23] Diederik Kingma, Tim Salimans, Ben Poole, and Jonathan Ho. Variational diffusion models. *Advances in neural information processing systems*, 34:21696–21707, 2021. 2, 3

[24] Diederik P Kingma and Prafulla Dhariwal. Glow: Generative flow with invertible 1x1 convolutions. In *Advances in neural information processing systems*, pages 10215–10224, 2018. 2

[25] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 2

[26] Komal Kumar, Snehashis Chakraborty, and Sudipta Roy. Self-supervised diffusion model for anomaly segmentation in medical imaging. In *International Conference on Pattern Recognition and Machine Intelligence*, Kolkata, India, 2023. 3, 6

[27] 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, 2023. IEEE. 1

[28] 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

[29] 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

[30] 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

[31] 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. 4, 5

[32] Namuk Park and Songkuk Kim. How do vision transformers work? *arXiv preprint arXiv:2202.06709*, 2022. 4, 6

[33] 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. 2, 7

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

[35] 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. 2

[36] 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

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

[38] 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. 4

[39] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1 (2):3, 2022. 4

[40] 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. 4

[41] 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

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

[43] 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. 4

[44] 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, 7

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

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

[47] 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

[48] 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.

[49] 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

[50] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017. 2

[51] 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

[52] 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. 2, 3, 4, 5, 7

[53] 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

Robust Self-Supervised Anomaly Detection with Diffusion Models

Supplementary Material

Algorithm 2 Simplex Noise

Input: Input point $\mathbf{P} = (x, y, z)$
Output: Noise value $N(\mathbf{P})$
Initialize $N(\mathbf{P}) \leftarrow 0$
Step 1: Grid Setup
Define a regular grid of points by dividing the space into a grid of cells: $\mathbf{G} = \{\mathbf{G}_1, \mathbf{G}_2, \ldots, \mathbf{G}_n\}$.
Step 2: Grid Positioning
Determine the position of $\mathbf{P}$ within the grid by finding the closest grid points to $\mathbf{P}$.
$\mathbf{V} = \{\mathbf{V}_1, \mathbf{V}_2, \ldots, \mathbf{V}_n\}$, where $\mathbf{V}_i$ is the position of the i-th closest grid point to $\mathbf{P}$.
Step 3: Gradient Vectors
Calculate the pseudo-random gradient vector $\mathbf{G}_i$ for each grid point: $\mathbf{G}_i = (g_{ix}, g_{iy}, g_{iz})$.
Step 4: Dot Products
Calculate the dot products between $\mathbf{G}_i$ and vectors from the closest grid points to $\mathbf{P}$: $D_i = \mathbf{G}_i \cdot (\mathbf{P} - \mathbf{V}_i)$.
Step 5: Interpolation
Interpolate the dot products using a smooth function to obtain $N(\mathbf{P})$:
$N(\mathbf{P}) = \sum_{i=1}^{n} F(D_i)$, where $F(D_i)$ is the interpolation function.
Step 6: Octaves
Repeat Steps 1-5 for multiple octaves with different frequencies and amplitudes, accumulating the results in $N(\mathbf{P})$.
return $N(\mathbf{P})$

Table 3. Hyperparameters

Parameter	Value
Image Settings	
Img_size	$(224, 224)$
Batch_Size	32
Epochs	3000
Time_step	1000
Model Configuration	
channels	1 or 3
beta_schedule	cos
loss-type	$l2 - norm$
learning rate	1e-4
patch_size	16
embed_dim	384
depth	6
num_heads	6
mlp_ratio	4
num_class	null or 2
EMA rate	0.9999
Other Parameters	
octave	6
frequency	64
persistence	0.9

7. Further Noise function

We have endeavored to enhance the sample quality by introducing two new noise sources: Tsimplex Gauss (Tsg) and Komal-Tsimplex (KTs) noise. Gaussian noise, known for its high-quality samples, is integrated into Tsg, which is formulated as follows:

$$\text{Tsg}(x) = \alpha \cdot \text{GenNoise}(S, t) + (1 - \alpha) \cdot \text{gauss}(x), \quad (6)$$

Here, $\alpha = \left(1 - \frac{t}{\chi}\right)$, with t representing the current diffusion step and χ denoting the total diffusion steps. For KTs noise, we introduce the multi-frequency ($F_{i,j}$) for simplex noise, where the position of a pixel (i, j), and described by the following equation:

$$F_{i,j} = \delta \times e^{-\left(\frac{(i-\frac{n+1}{2})^2}{2\sigma^2} + \frac{(j-\frac{m+1}{2})^2}{2\sigma^2}\right)} \quad (7)$$

In this equation, δ represents the desired average frequency value, n and m are the dimensions of the grid, and σ controls the spread of the Gaussian distribution. A larger value of σ results in a Komal (smoother) distribution.