Medical Image Analysis 93 (2024) 103088

Contents lists available at ScienceDirect

Medical Image Analysis

journal homepage: www.elsevier.com/locate/media

Diffusion models for out-of-distribution detection in digital pathology

Jasper Linmans [a,*], Gabriel Raya [b], Jeroen van der Laak [a,c], Geert Litjens [a]

[a] Department of Pathology, RadboudUMC Graduate School, Radboud University Medical Center, Nijmegen, The Netherlands
[b] Jheronimus Academy of Data Science, 's-Hertogenbosch, The Netherlands
[c] Center for Medical Image Science and Visualization, Linköping University, Linköping, Sweden

ARTICLE INFO

Keywords:
Denoising diffusion probabilistic models
Out-of-distribution detection
Deep learning
Histopathology
Unsupervised learning

ABSTRACT

The ability to detect anomalies, i.e. anything not seen during training or out-of-distribution (OOD), in medical imaging applications is essential for successfully deploying machine learning systems. Filtering out OOD data using unsupervised learning is especially promising because it does not require costly annotations. A new class of models called AnoDDPMs, based on denoising diffusion probabilistic models (DDPMs), has recently achieved significant progress in unsupervised OOD detection. This work provides a benchmark for unsupervised OOD detection methods in digital pathology. By leveraging fast sampling techniques, we apply AnoDDPM on a large enough scale for whole-slide image analysis on the complete test set of the Camelyon16 challenge. Based on ROC analysis, we show that AnoDDPMs can detect OOD data with an AUC of up to 94.13 and 86.93 on two patch-level OOD detection tasks, outperforming the other unsupervised methods. We observe that AnoDDPMs alter the semantic properties of inputs, replacing anomalous data with more benign-looking tissue. Furthermore, we highlight the flexibility of AnoDDPM towards different information bottlenecks by evaluating reconstruction errors for inputs with different signal-to-noise ratios. While there is still a significant performance gap with fully supervised learning, AnoDDPMs show considerable promise in the field of OOD detection in digital pathology.

1. Introduction

Detecting anomalous data in medical imaging is critical for successfully deploying machine learning systems, especially in heterogeneous domains like digital pathology, where anomalies can negatively impact the performance of trained machine learning systems (Schömig-Markiefka et al., 2021; Linmans et al., 2023). Filtering out anomalous data using unsupervised learning is especially promising because this does not make assumptions about the type of anomalies. Furthermore, it does not require any known anomalies in advance. In theory, by simply training on a library of in-distribution data such as benign cases, unsupervised methods enable the detection of anything deviating from the training data. A recent leap in performance within unsupervised anomaly detection (Pinaya et al., 2022; Wyatt et al., 2022; Graham et al., 2022a), spearheaded by a new class of models referred to as denoising diffusion probabilistic models (DDPM) (Sohl-Dickstein et al., 2015; Ho et al., 2020), motivates further research within the medical domain.

In this work, we define anomalies as data points that lie outside the manifold of the learned distribution, as done in previous studies (Schlegl et al., 2019; Wyatt et al., 2022). We adopt terminology from Fernando et al. (2021) to describe anomaly detection as the process of identifying these Out-of-Distribution (OOD) instances. Consequently, we use the terms OOD and anomalies interchangeably throughout this study. To emphasize the difference between a data point and the training distribution, we frequently use the term OOD.

The criteria for classifying a data point as OOD often lack clarity, particularly in the field of digital pathology where numerous sources of variation exist (Fernando et al., 2021; Schömig-Markiefka et al., 2021). These variations may arise from less impactful sources such as demographic shifts or scanner variations, or from more significant factors such as modality shifts or the presence of adverse pathology. Generally, the classification of what is considered OOD is contingent on the training data. For instance, in the context of sentinel lymph node resections (Ehteshami Bejnordi et al., 2017; Litjens et al., 2018), metastatic tissue can be considered far less anomalous compared to rare disease events such as diffuse large B-cell lymphoma (Fox et al., 2010). However, when trained only on benign lymph node tissue, we consider metastatic tissue as OOD. In our work, we aim to identify the presence of OOD data within an image. Either within a patch, extracted from a Whole-Slide Image (WSI), or at the WSI-level to ascertain the presence of OOD data anywhere on the slide.

* Corresponding author.
E-mail address: jasper.linmans@radboudumc.nl (J. Linmans).

https://doi.org/10.1016/j.media.2024.103088
Received 4 February 2023; Received in revised form 29 December 2023; Accepted 10 January 2024
Available online 13 January 2024

Earlier work has applied a variety of techniques for anomaly detection, both supervised and unsupervised, with varying degrees of success. Supervised learning approaches can be used for anomaly detection by evaluating different statistics from the predictive distribution to identify cases that deviate from the training set (Kompa et al., 2021), which has been shown to work well in detecting anomalies in medical data (Guha Roy et al., 2022; Pocevičiūtė et al., 2022; Graham et al., 2022b; Linmans et al., 2023). Similarly, unsupervised methods trained on generative or compression tasks can identify anomalous data by detecting deviations in data characteristics (Salehi et al., 2021; Yang et al., 2021). In contrast, to supervised anomaly detection, a single unsupervised model could potentially be used as a universal anomaly detector to detect any disease event without requiring costly annotations. With a large list of potential disease events, anomalies, and sources for distribution shift (Schömig-Markiefka et al., 2021), the medical domain seems like a natural fit for unsupervised methods.

The current dominant approaches in unsupervised anomaly detection can be broadly classified into three categories (Salehi et al., 2021; Yang et al., 2021). *Density-based* methods use likelihood values or other statistics from a generative model to identify samples that deviate from the in-distribution training data. However, these methods sometimes fail dramatically, such as when a model trained on the CIFAR10 dataset assigns higher likelihoods to samples from the SVHN dataset than samples from CIFAR10 (Choi and Jang, 2019; Nalisnick et al., 2019). Several methods have been developed to improve density-based methods, but theoretical work suggests they may still perform poorly on certain anomalous data (Zhang et al., 2021).

A different class of unsupervised anomaly detection approaches uses *reconstruction-based* methods (Salehi et al., 2021; Yang et al., 2021). Here, models are trained to reconstruct in-distribution data such that the reconstruction error for a specific sample indicates the resemblance with samples from the training set. Often based on autoencoder-like architectures, these methods rely on an information bottleneck to capture essential in-distribution features resulting in lower reconstruction errors for in-distribution data than out-of-distribution (OOD) data (Denouden et al., 2018; Schlegl et al., 2019). Related *distance-based* methods often use similar network architectures and training procedures but evaluate the similarity of a new sample's lower-dimensional feature vector to feature centroids from the training set (Denouden et al., 2018; Gong et al., 2019).

Recent work introduces a new reconstruction-based method using DDPMs (Pinaya et al., 2022; Wyatt et al., 2022; Graham et al., 2022a). At its core, DDPMs are trained to reconstruct an image from corrupted versions with increasing amounts of noise (Sohl-Dickstein et al., 2015; Ho et al., 2020). Specifically, the DDPM is trained to denoise corrupted images for a fixed amount of timesteps corresponding to different levels of corruption. At the end of the diffusion process, no information from the input itself is retained; denoising a fully diffused input will generate a new sample. When a partially-diffused image is used as input instead, it will generate an image based on any remaining signal, effectively switching the image generation objective to image reconstruction. As such, reconstruction errors can be used to indicate similarity with the in-distribution training set, similar to other reconstruction-based methods. We borrow terminology from Wyatt et al. (2022) and refer to this method as AnoDDPM. Originally, the AnoDDPM is defined by a timestep associated with the partial diffusion process. We propose to analyze the signal-to-noise ratio (SNR) instead to determine the appropriate amount of added noise. See Fig. 1 for example outputs for three in-distribution and three out-of-distribution (OOD) samples for a model trained on benign lymph node tissue.

Prior work shows that the AnoDDPM can outperform other unsupervised methods on multiple classical computer vision anomaly detection tasks (Graham et al., 2022a). Here, multiple reconstructions based on different timesteps are evaluated to distinguish in-distribution from OOD data. However, the timesteps used for reconstruction are uniformly sampled along the diffusion process, including those corresponding to heavy noise and almost no noise. As such, a lot of computational resources are lost on trying to reconstruct images with too little or too much signal, which can even hurt anomaly detection performance (Graham et al., 2022a). Instead, a more careful selection of timesteps might further improve performance.

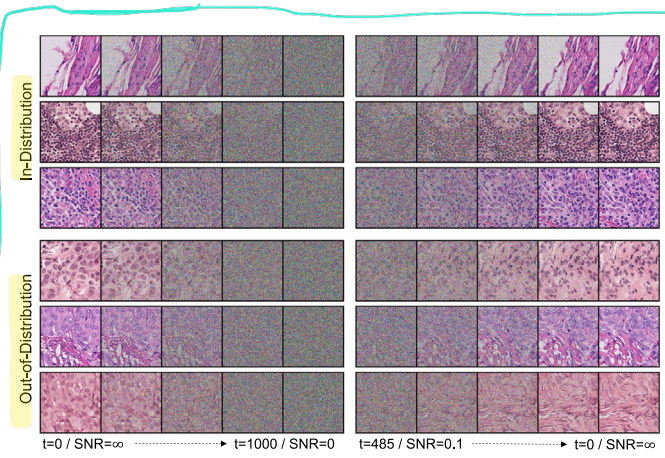

Fig. 1. The diffusion process (left) and denoising process (right) for three random patches from both the in-distribution and the out-of-distribution dataset used in this work using a DDPM model trained on benign lymph node tissue. Evenly spaced intermediate outputs are shown for the DDPM model starting from $t = 485$, corresponding to a signal-to-noise ratio (SNR) of 0.1. Visually, reconstruction errors seem lower for in-distribution data.

Unsupervised methods have already been applied to detect anomalies in medical imaging. In the field of radiology, various methods have been used for brain MRI scans (Baur et al., 2019, 2020), with recent work demonstrating the effectiveness of DDPM-based methods when trained purely on healthy patient data (Pinaya et al., 2022; Wyatt et al., 2022). This current study addresses the challenge of unsupervised anomaly detection, specifically within the heterogeneous domain of digital pathology. Unlike more conventional anomaly detection benchmark datasets that involve complete domain shifts like CIFAR10 versus SVHN, histopathology data is characterized by subtle abnormalities that pose a more significant challenge for anomaly detection (Linmans et al., 2023). Related work also evaluates methods of unsupervised anomaly detection in digital pathology (Pocevičiūtė et al., 2021; Stepec and Skocaj, 2021). However, current work is limited to methods based on generative adversarial networks (GAN) (Goodfellow et al., 2014) and only evaluates using image patches extracted from whole-slide images (WSIs). More extensive analysis on a whole-slide image level and comparisons with other methods of unsupervised OOD detection is lacking.

1.1. Contributions

In this paper, we quantitatively compare prevalent methods of unsupervised OOD detection in digital pathology. We include the recently introduced AnoDDPM, which has shown new state-of-the-art performance in earlier work in other domains (Graham et al., 2022a). Specifically, we propose to analyze the signal-to-noise ratio to determine the appropriate amount of noise for the partial diffusion process of the AnoDDPM. We aim to better understand existing SOTA models, their applicability to OOD problems in digital pathology, and the challenges related to model deployment in clinical practice.

To do so, we train different reconstruction-based and distance-based methods on purely benign lymph node tissue from the Camelyon challenge (Ehteshami Bejnordi et al., 2017; Litjens et al., 2018). Afterward, models are evaluated on the test set of the Camelyon16 challenge, which contains both benign lymph node tissue and lymph nodes with detailed, pixel-level annotations of breast cancer metastasis (OOD). To

Table 1
Summary of the histopathology datasets that are used in this study.

Dataset	Description
Training	
$\mathcal{D}_{train}$	Patches extracted from 127 fully benign WSIs from the C16 training set.
$\mathcal{D}_{valid}$	Patches extracted from 32 fully benign WSIs from the C16 training set.
Inference (WSIs)	
$\mathcal{D}_{in}$	All 80 benign WSIs from the C16 test set
$\mathcal{D}_{out}$	All 49 tumor WSIS from the C16 test set
$\mathcal{D}_{out_{macro}}$	All 22 macro-metastatic WSIs from $\mathcal{D}_{out}$
Inference (patches)	
$\mathcal{D}_{valid}$	100k patches, extracted from $\mathcal{D}_{valid}$
$\mathcal{D}_{in}$	100k patches, extracted from $\mathcal{D}_{in}$
$\mathcal{D}_{out_{100\%}}$	100k fully tumor patches, from $\mathcal{D}_{out}$
$\mathcal{D}_{out_{cp}}$	100k center-pixel tumor patches from $\mathcal{D}_{out}$

enable OOD detection using AnoDDPMs on this scale, we leverage pseudo-numerical methods to speed up the sampling process during inference significantly (Liu et al., 2022).

The key contributions of this study are:

- We compare multiple unsupervised methods on both a patch and WSI level to evaluate their performance on a scale relevant to clinical deployment in digital pathology. In both settings we evaluate two different OOD tasks with different levels of difficulty.
- We analyze the OOD detection performance of AnoDDPM for different levels of noise defining the partial diffusion process based on different signal-to-noise ratios. We use pseudo-numerical methods to speed up the denoising process and compare its performance to regular denoising.
- To better understand the denoising process, we include an in-depth comparison between the AnoDDPM on the proposed subset of signal-to-noise ratios and different ranges of uniformly sampled noise levels as done in Graham et al. (2022a), in terms of accuracy and efficiency on both a patch and WSI-level.
- We include a comparison with the performance of a regular classifier trained to discriminate in-distribution from OOD data. Doing so we gain more insights into the applicability of unsupervised OOD detection methods in relation to the current standard of supervised methods.

2. Materials

Let a dataset be noted as $\mathcal{D}_{train} = \{\mathbf{x}_n | \mathbf{x}_n \in \mathcal{X}_{in}\}_{n=1}^{N_{train}}$ where $\mathcal{X}_{in}$ defines the in-distribution image space for N_{train} amount of data samples. Here, we train various types of unsupervised models to capture different features and statistics from the underlying distribution of benign tissue samples. After training, using different similarity metrics specific to the type of unsupervised model, we evaluate the ability to discriminate between a held-out test dataset $\mathcal{D}_{test} = \{\mathbf{x}_n | \mathbf{x}_n \in \mathcal{X}_{in}\}_{n=1}^{N_{test}}$ and an Out-of-Distribution (OOD) dataset $\mathcal{D}_{out} = \{\mathbf{x}_n | \mathbf{x}_n \in \mathcal{X}_{out}\}_{n=1}^{N_{out}}$. The latter containing anomalies not seen during training from some OOD image space $\mathcal{X}_{out}$. This task is referred to as OOD detection. Our main goal is not to obtain the new state-of-the-art in image quality for reconstructed or generated images, but to evaluate the OOD detection performance of various methods of unsupervised learning on a large-scale and real-world histopathology dataset.

2.1. Training

In this work, we train models on lymph node tissue using data from the Camelyon16 (C16) challenge (Ehteshami Bejnordi et al., 2017; Litjens et al., 2018). We use the training set $\mathcal{D}_{train}$ (216 WSIs), validation set (54 WSIs), and test set (129 WSIs) as defined by the challenge organizers. Here, models are trained on randomly selected patches with sizes of either 64 × 64 or 256 × 256 from a 20× resolution with a pixel spacing of 0.48 µm. Specifically, we only train on a subset of 127 WSIs from the training set containing only benign lymph node tissue. By only sampling patches from fully benign WSIs during training, we prevent contamination from non-labeled isolated tumor cells outside the annotated regions in tumor slides (Ehteshami Bejnordi et al., 2017).

2.2. Evaluation

To evaluate the performance on the OOD detection task, we define $\mathcal{D}_{out}$ as the subset of WSIs from the test set of C16 with the presence of breast cancer metastasis (49 WSIs). Additionally, we run analysis on the subset of WSIs $\mathcal{D}_{out_{macro}}$ containing macrometastatic tumor lesions: with a tumor cell cluster diameter ≥ 2 mm (Ehteshami Bejnordi et al., 2017). Because models are only trained on benign tissue, breast cancer metastasis is considered anomalous. We use the remaining WSIs containing only benign tissue as the in-distribution test set $\mathcal{D}_{in}$ (80 WSIs). Using the exact data splits defined by the challenge organizers, we can compare the OOD performance of the different unsupervised methods with the performance of a fully supervised binary classifier trained on the C16 challenge dataset.

Specifically, we evaluate on both a patch and a WSI level. For the patch datasets, we use 100k patches per split and define two separate OOD sets. The first corresponds to a more straightforward OOD task, where the OOD patches are restricted such that 100% of the pixels fall within the annotated tumor regions. We refer to this dataset as $\mathcal{D}_{out_{100\%}}$. The second OOD set is less restrictive: patches are randomly sampled so that at least the center pixel falls within the annotated tumor regions. As such, $\mathcal{D}_{out_{cp}}$ also contains patches near tumor boundaries containing benign and tumor pixels. See Table 1 for an overview of the different histopathology datasets used throughout this work.

Throughout all experiments, we perform ROC analysis based on method-specific similarity metrics to evaluate OOD detection performance. Using the area under the ROC curve, we analyze the ability to discriminate in-distribution from OOD samples independent of the exact decision threshold. We report confidence bounds by 2000-fold bootstrapping in all experiments.

3. Methods

In this study, we evaluate different unsupervised OOD detection methods by evaluating and comparing method-specific similarity metrics on both in-distribution and out-of-distribution data. Due to the limitations set by each individual unsupervised method, some methods are trained on a lower image resolution of 64 × 64. However, when possible, we train and evaluate models on an image resolution of 256 × 256. This section describes these methods and their similarity metrics, starting with the recently introduced AnoDDPM.

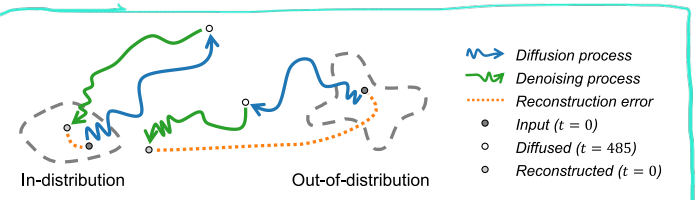

Fig. 2. Graphical representation of partial-diffusion (forward) and denoising (backward) processes of the AnoDDPM for an in-distribution and OOD sample. We hypothesize that the in-distribution image space will act as an attractor to the partially diffused image, resulting in larger reconstruction errors for OOD data.

3.1. Denoising diffusion probabilistic models (DDPM)

Central to DDPMs are both the *forward diffusion* process, gradually corrupting in-distribution data into a standard Gaussian distribution, and a learned *backward denoising* process that generates samples by turning noise back into images (Sohl-Dickstein et al., 2015; Ho et al., 2020). Samples from the training set $\mathbf{x}_0$ are gradually diffused according to a fixed process of T steps, with Gaussian noise added according to a noise variance schedule $\beta_t \in (0, 1)$ to produce noised samples $\mathbf{x}_t$, such that:

$$q(\mathbf{x}_t \mid \mathbf{x}_0) = \mathcal{N}\left(\mathbf{x}_t \mid \sqrt{\bar{\alpha}_t}\mathbf{x}_0, (1-\bar{\alpha}_t)I\right) \tag{1}$$

where $0 \leq t \leq T$, $\alpha_t := 1 - \beta_t$, and $\bar{\alpha}_t := \prod_{s=1}^{t} \alpha_s$. The schedule β_t is defined to increase with t, with $\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, I)$. Note, (1) is designed so that we can directly sample $\mathbf{x}_t$ at any timestep conditioned only on the input image $\mathbf{x}_0$ (Ho et al., 2020). In contrast, to model the backward process we learn each individual step along the Markov chain, where each step in the process depends (only) on the preceding state. To do so, we train a single network to iteratively reverse the diffusion process following:

$$p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t) = \mathcal{N}\left(\mathbf{x}_{t-1} \mid \mu_\theta(\mathbf{x}_t, t), \tilde{\beta}_t I\right) \tag{2}$$

with $\tilde{\beta}_t = \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t$ following from the fixed variance schedule. As described by Ho et al. (2020), the backward process (2) is trained by optimizing a lower bound objective on the data log-likelihood $\mathbb{E}[-\log p_\theta(\mathbf{x}_0)]$, which can ultimately be simplified into the following objective function:

$$\mathcal{L}_{simple} = \mathbb{E}_{t,\mathbf{x}_0,\epsilon}\left[\|\epsilon - \epsilon_\theta(\mathbf{x}_t, t)\|^2\right] \tag{3}$$

with $\epsilon \sim \mathcal{N}(\mathbf{0}, I)$ and ϵ_θ modeled by a network based on the U-Net architecture (Ronneberger et al., 2015). Although the network is trained to predict random noise ϵ from $\mathbf{x}_t$, we can use it to parametrize $\mu_\theta(\mathbf{x}_t, t)$ from (2) as:

$$\mu_\theta(\mathbf{x}_t, t) = \frac{1}{\sqrt{\alpha_t}}\left(\mathbf{x}_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(\mathbf{x}_t, t)\right). \tag{4}$$

Using (4) and the reparametrization trick (Kingma and Welling, 2014), we can sample $\mathbf{x}_{t-1} \sim p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t)$ with: $\mathbf{x}_{t-1} = \mu_\theta(\mathbf{x}_t, t) + \tilde{\beta}_t\epsilon$ and $\epsilon \sim \mathcal{N}(\mathbf{0}, I)$. Following principles of Langevin dynamics, data generation involves gradually moving a randomly selected initial sample, by repeated sampling $\mathbf{x}_{t-1} \sim p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t)$, towards regions of high density using the gradient of the data density as a guide (Song and Ermon, 2019).

3.2. AnoDDPM

In the context of the AnoDDPM framework, reconstructing a partially diffused image with $t = \lambda$ means evaluating (2), λ times. We hypothesize that the learned in-distribution image space will act as an attractor to the partially diffused image: denoising the image will bring it closer towards what is considered in-distribution, resulting in larger reconstruction errors for OOD data (Graham et al., 2022a). See Fig. 2 for a graphical representation of both processes for the AnoDDPM.

Similar to Graham et al. (2022a) we sample a set of $\mathbf{x}_t$ for a range of K values $t \in \{\lambda_0, \lambda_2, \ldots, \lambda_K\}$ and evaluate each reconstruction $\hat{\mathbf{x}}_{0,t} = p_\theta(\mathbf{x}_0|\mathbf{x}_t)$ by measuring the similarity with the original input $S(\hat{\mathbf{x}}_{0,t}, \mathbf{x}_0)$ using some similarity measure S. Here, we propose to select the K relevant timesteps in terms of the signal-to-noise ratio of the diffused data, as defined by Kingma et al. (2021):

$$SNR(t) = \frac{\bar{\alpha}_t}{(1-\bar{\alpha}_t)}. \tag{5}$$

Specifically, we evaluate OOD detection performance using the following subset of SNR values: [2.0, 1.5, 1.0, .75, .5, .25, .1, .05, .01]. By not including timesteps corresponding to even more noise (lower SNR) and almost no noise (higher SNR), we prevent image reconstruction of images with too little or too much signal, which can hurt OOD detection performance (Graham et al., 2022a). See the left panel in Fig. 3 for a visualization of the SNR plot and the proposed timesteps used in this work corresponding to the different SNR values.

We leverage recent advances in fast sampling methods for diffusion models to enable the application of AnoDDPMs on a large enough scale for digital pathology (Song et al., 2021; Liu et al., 2022). In particular, we use the PNDM sampler, a linear four-step pseudo numerical sampling method which has been shown to substantially reduce the number of sampling steps required during inference while maintaining or improving sample quality (Liu et al., 2022). We define the total amount of timesteps as $T = 1000$. However, during inference using the PNDM sampler, we define the total number of backward steps as 100 evenly spaced timepoints $t \in [10, 20, 30, \ldots, 990, 1000]$ to reduce the number of steps by a factor of 10. To evaluate the impact of the PNDM sampler on the OOD detection performance, we will compare the performance of the AnoDDPM-PNDM model with a regular AnoDDPM model. Specifically for AnoDDPM-PNDM, we round the corresponding timepoint associated with each SNR to the nearest ten to start the backward process.

We hypothesize that the ideal SNR is task-dependent, and finding a single optimal SNR for all possible OOD detection tasks might be unrealistic. Therefore, we not only evaluate reconstruction errors for each individual SNR value, but also evaluate the aggregate reconstruction error across the subset of nine SNR values as depicted in the left panel of Fig. 3. We use the same subset of SNR values all throughout this work and therefore simply refer to this method as AnoDDPM-PNDM (SNR_{subset}). See Section 3.7 for details regarding the exact method of aggregating reconstruction errors based on Z-score averages.

To get a deeper understanding of the denoising process of the AnoDDPM we compare the efficiency and OOD detection performance of the proposed SNR subset model with the uniform sampling approach as proposed by Graham et al. (2022a). In the method proposed by Graham et al. (2022a), the timesteps used for reconstruction are uniformly sampled along the diffusion process, including those corresponding to heavy noise and almost no noise (i.e. all $t \in [10, 20, 30, \ldots, 990, 1000]$). To put the additional computational costs in perspective, and to compare the efficiency of using the PNDM sampler with the regular denoising process, we provide exact inference costs in the right panel of Fig. 3. Here we note that the uniform backward process as proposed by Graham et al. (2022a), which we denote as AnoDDPM-PNDM ($U_{[0,1000]}$), demands significantly more computational resources compared to the proposed backward process based on a subset of nine SNR values, increasing the time required in the backward process by a factor of fifteen.

We use the exact DDPM model architecture and training procedures as previously described to model the CelebA-HQ dataset (Karras et al., 2018) on a resolution of 256 × 256 in earlier work (Ho et al., 2020). Specifically, we use a 4-layer U-Net with [128, 256, 256, 256] channels, with two residual blocks per layer and a single-headed attention block after each residual block with a corresponding spatial dimension of 16. The timestep is encoded using a transformer sinusoidal position embedding (Vaswani et al., 2017), which is added as a bias term in

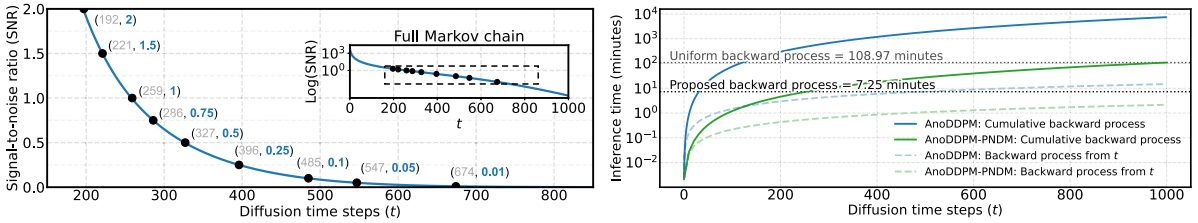

Fig. 3. Left: Illustrating the timesteps used in the AnoDDPM framework for partial diffusion and denoising. The proposed timesteps used in this work, corresponding to the different SNR values are included as well as the context of the zoomed-in part in relation to the log(SNR) values of the entire Markov chain. Right: The exact inference costs measured for various settings of the AnoDDPM. Each process is measured in minutes (log scale), based on a batch size of 128 using a single A100 GPU. The dashed lines represent the time in minutes required for the backward process from some diffused state t, which increases linearly over time. The solid lines represent the uniform process as defined by Graham et al. (2022a) up until diffused state t. We highlight the uniform backward process of the full Markov chain Graham et al. (2022a), and the proposed backward process corresponding to the subset of nine SNR values as well.

each residual block. We define a linear noise schedule with β_t varying between 1e-4 and 0.02 as described in Ho et al. (2020). Similar to Ho et al. (2020), we train the DDPM model for 1.3M iterations with a batch size of 28 using the Adam optimizer (Kingma and Ba, 2015), limited by the available computational resources.

3.3. (Denoising) autoencoders

As a baseline, we train and evaluate a simple autoencoder (AE) on inputs of size 64 × 64 where the encoder and decoder are defined by plain convolutional neural networks following the architectural design of Gong et al. (2019). We first define Conv2(k, s, c) to denote a 2D convolution layer, where k, s, and c are the kernel size, stride size, and the number of kernels respectively. We implement the encoder using four convolutional layers: Conv2(3, 2, 128)-Conv2(3, 2, 128)-Conv2(3, 2, 128)-Conv2(3, 2, 256). The decoder is symmetrical but replaces the convolutional layers with deconvolutional layers. Except for the last deconvolutional layers, each layer is followed by batch normalization and a ReLU activation.

In an effort to improve the OOD detection performance of the autoencoder baseline, we also train and evaluate a denoising autoencoder (DAE) (Vincent et al., 2010). Here the objective is similar to the objective of the AnoDDPM. However, instead of iteratively removing noise, a DAE is trained to remove a predetermined amount of noise in a single forward pass. For a fair comparison, we also include a DAE using the exact U-Net architecture used in the AnoDDPM model (UNet-DAE) and train it to remove the same amount of noise. Here, we remove the time-conditional layers specific to the DDPM model. With similar objectives, it is noteworthy that current work on AnoDDPM does not compare against DAEs (Pinaya et al., 2022; Wyatt et al., 2022; Graham et al., 2022a).

3.4. Generative adversarial networks

Similar to DDPMs, GANs are generative models that can be applied to OOD detection tasks (Schlegl et al., 2019; Karras et al., 2020; Pocevičiūtė et al., 2021; Stepec and Skocaj, 2021). In this work, we include the F-AnoGAN (Schlegl et al., 2019) and the StyleGANv2 (Karras et al., 2020) methods. The F-AnoGAN model is based on a Wasserstein GAN (WGAN) (Arjovsky et al., 2017), which is limited to an image resolution of 64 × 64. Here, an encoder is trained to map data to the WGAN's latent space in a single step after training the WGAN. The encoder's purpose is to map images to specific locations in the latent space so that the latent vector can be used to reconstruct the original input using the generator. We follow the exact WGAN architecture and training procedures as proposed by the original authors (Schlegl et al., 2019) and train a ResNet-based WGAN with gradient penalty (Wei et al., 2018). Following Schlegl et al. (2019), we copy the architecture of the discriminator to define the encoder network and optimize it using the RMSprop optimizer for 50.000 iterations while keeping the parameters of the discriminator and generator fixed.

Similarly, StyleGANv2 enables mapping images to the latent space of a trained GAN. However, it is designed to work with high-resolution images and is based on a simple iterative optimization procedure instead of a separate encoder network (Karras et al., 2020). We train a StyleGANv2 model using the original authors' network architecture and training procedure on the same image resolution as the DDPM. Following Karras et al. (2020), the optimization procedure starts by defining the initial latent code $\mathbf{w}$ as $\mathbf{w} = \mathbb{E}_{\mathbf{z}}[f(\mathbf{z})]$ based on the mapping network f and $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. Afterward, both the latent code $\mathbf{w}$ and per-layer noise maps $\mathbf{n}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ for all i layers, together forming the input to the generator network, are optimized for 1000 iterations using the Adam optimizer. Here, the objective function is composed of a regularization term and an image reconstruction loss based on the Learned Perceptual Image Patch Similarity (LPIPS) distance metric (Zhang et al., 2018). The iterative optimization procedure is required for each image during inference, making this especially computationally expensive.

3.5. Adversarial autoencoders

We also train and evaluate an Adversarial AutoEncoder (AAE) (Makhzani et al., 2016). The AAE combines the reconstruction capabilities of an AE with the adversarial training of a GAN. The training process of the AAE involves two objectives: first, the encoder and decoder are trained to minimize the reconstruction error, and then the discriminator is trained to distinguish between the encoded data and a prior distribution. This adversarial training encourages the encoder to produce more realistic representations of the input data, while the decoder learns to generate samples that are more similar to the original data distribution. Here, we copy the architecture of the previously defined AE for inputs of size 64 × 64 and define the discriminator by three linear layers followed by dropout and ReLU activation layers, except for the final layer which is defined by a Sigmoid activation.

3.6. Supervised learning baseline

To enable comparisons between the different unsupervised OOD detection methods and the current standard on the task of breast cancer metastasis in Camelyon16, we include a fully supervised patch-based classifier. Specifically, we adopt a lightweight DenseNet architecture (Huang et al., 2017), which has shown near SOTA performance on Camelyon16 in prior work (Linmans et al., 2023). This proven network architecture contains three dense blocks with eight valid padded convolutional layers. In total: 27 convolutional layers with 32 initial filters and a growth rate of 32. We follow the exact training procedures as outlined in Linmans et al. (2023).

3.7. Similarity metrics

A common method measures the mean-squared error (MSE) between the input and the reconstruction to evaluate similarity. Here, the difference between the input $\mathbf{x}$ and its reconstruction $\hat{\mathbf{x}}$ is given

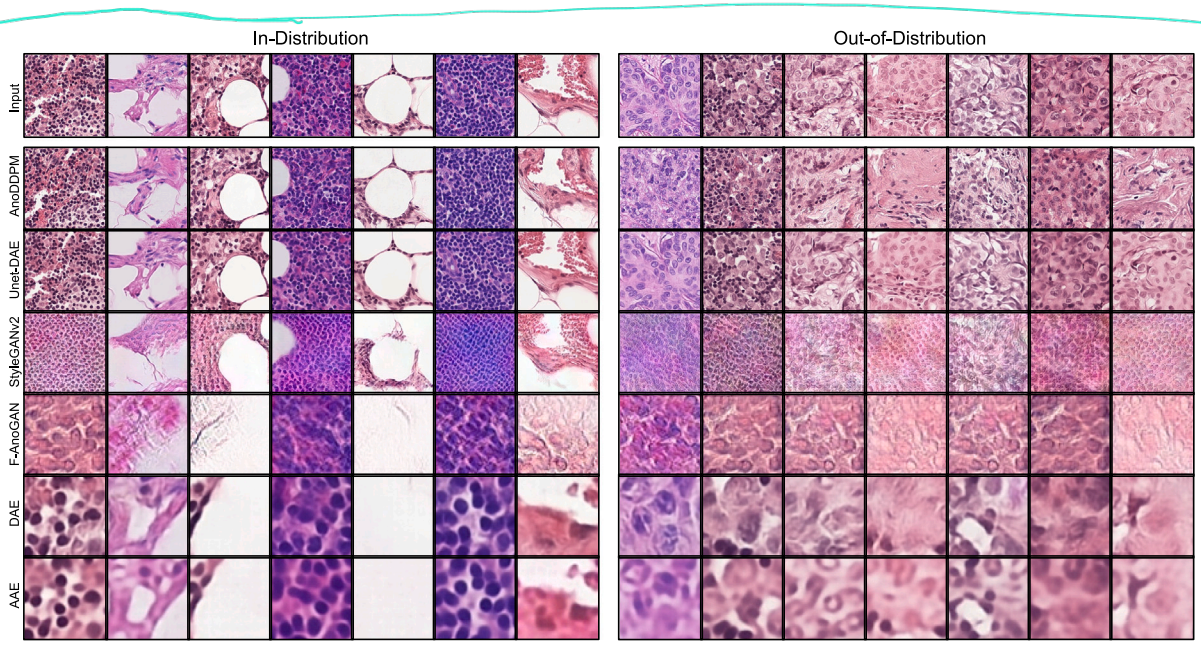

Fig. 4. Example input and reconstructions of in-distribution benign lymph node tissue patches and cancerous OOD tissue patches for the different reconstruction-based methods used in this work. The output for the AnoDDPM and the UNet-DAE are shown after denoising the image patches corresponding to an SNR of 0.1. The F-AnoGAN, DAE, and AAE are trained and evaluated on an image resolution of 64 × 64 (the center of the input patch), the other methods are trained on 256 × 256. AnoDDPM appears to alter the semantic properties of OOD tissue by replacing tumor cells with more benign-appearing tissue. Best viewed in a digital version with zoom.

by $\|\mathbf{x} - \hat{\mathbf{x}}\|_2$. A higher MSE indicates a larger discrepancy between the input and its reconstruction. In this work, we also use the Structural Similarity Index Measure (SSIM) (Wang et al., 2004). In contrast to the MSE, which is based on absolute errors between individual pixels, SSIM compares local patterns of pixel intensities that have been normalized for luminance and contrast. Higher values indicate a greater similarity between the reconstruction and the input. Finally, we evaluate the LPIPS (Zhang et al., 2018) reconstruction error, which uses the distance between the deep features of a pre-trained network (in this case, Alexnet (Krizhevsky et al., 2012)) from two inputs to measure their perceptual similarity. LPIPS has been shown to correlate well with human evaluations of image similarity (Zhang et al., 2018). Similar to MSE, a higher LPIPS indicates larger differences between the input and the reconstruction.

Using all three metrics, we have multiple estimates of similarity for each input and the reconstructed image. Similar to Graham et al. (2022a), we convert all measurements into a single anomaly score by evaluating the average Z-score using statistics measured on a validation set for each metric separately. Specifically, we evaluate image reconstructions on all 100k patches in $\mathcal{D}_{valid}$ and measure the mean μ_m and standard deviation σ_m across all patches for all M similarity metrics. During inference on the in-distribution or OOD test sets, we evaluate the average Z-score:

$$Z - score = \frac{1}{M}\sum_{m=1}^{M}\frac{S_m(\hat{\mathbf{x}}, \mathbf{x}) - \mu_m}{\sigma_m}. \tag{6}$$

with $S_m(\hat{\mathbf{x}}, \mathbf{x})$ the output of the m'th similarity metric comparing the input $\mathbf{x}$ with its reconstruction $\hat{\mathbf{x}}$. Note that SSIM is inversely proportional to MSE and LPIPS and has an upper bound of 1. Therefore, to summarize all three similarity values in a single Z-score, we use the inverse SSIM (i.e., 1 - SSIM) in contrast with earlier work, which only evaluates MSE and LPIPS (Graham et al., 2022a). As a result, higher Z-scores (6) correspond to larger differences between the input and the reconstruction.

For AnoDDPM-PNDM settings where the backward process is defined by multiple timesteps, e.g. (SNR_{subset}) and ($U_{[0,1000]}$), we evaluate all reconstruction errors (i.e., $M = 3K$) corresponding to the total set of K different SNR values, similar to Graham et al. (2022a). To do so, we first evaluate the Z-score for each image and SNR value. Then we report OOD detection results per similarity metric using the average Z-score. Afterward, we take the average Z-score (aggregated over all similarity metrics, following (6)) and perform ROC analysis.

To include a comparison with a distance-based OOD detection method, we also evaluate ROC analysis for the encoder using the Mahalanobis distance given by $\alpha D_M(E(\mathbf{x})) + \beta\|\mathbf{x} - \hat{\mathbf{x}}\|_2$ (Denouden et al., 2018). Here, $D_M(E(\mathbf{x}))$ is the Mahalanobis distance from the latent space of the encoded input $E(\mathbf{x})$ to the latent space of the training set, and α and β are constants set to the reciprocal of the standard deviation of $D_M(E(\mathbf{x}))$ and $\|\mathbf{x} - \hat{\mathbf{x}}\|_2$ respectively, as evaluated on the validation set.

Finally, we include a methods which utilizes deep pre-trained features, which showed state-of-the-art in similar computer vision tasks, such as industrial inspection (Roth et al., 2021; Bergmann et al.). Here, we follow the exact implementation as presented in Salehi et al. (2020), which applied multiresolution knowledge distillation for anomaly detection in medical imaging.

3.8. Whole-slide image level evaluation

By evaluating the Z-score (6) value for every patch in a whole-slide image exhaustively, we end up with an anomaly score heatmap. We determined, through visual inspection of whole-slide level anomaly heatmaps from the validation set, that minimal postprocessing using greyscale erosion should be applied in order to reduce small false positive regions. Specifically, we use the standard implementation of SciPy with an erosion size of three (Virtanen et al., 2020). The resulting heatmap can help identify ambiguous regions within a WSI. However, discriminating between in-distribution and OOD data on a whole-slide level requires translating the heatmap to a slide-level anomaly score. To this end, we propose to perform ROC analysis using either the maximum Z-score value or the average Z-score of all values exceeding the 99'th percentile of the anomaly heatmap, similar to Linmans et al. (2023).

In an effort to further improve WSI-level anomaly detection, we also implement a simple one-class SVM (Schölkopf et al., 1999) using the default implementation in scikit-learn (Pedregosa et al., 2011) and train

Table 2
OOD detection performance on two separate detection tasks: in-distribution benign patches from $\mathcal{D}_{in}$ vs tumor patches from out-of-distribution data with either full tumor coverage or center-pixel tumor labels, $\mathcal{D}_{out_{100\%}}$ and $\mathcal{D}_{out_{cp}}$ respectively. AUC values are reported using each individual similarity metric, and the Z-score average across all metrics. Confidence bounds are reported for all methods using 2000-fold bootstrapping. The AnoDDPM and DAE models are trained to denoise images with a SNR of 0.1. The AnoDDPM method based on the proposed subset of nine SNR values is included, as well as the AnoDDPM based on the uniform denoising process of the entire Markov chain. Both methods report the average Z-score across all K timesteps for each individual metric.

Model	$\mathcal{D}_{in}$ vs. $\mathcal{D}_{out_{100\%}}$				$\mathcal{D}_{in}$ vs. $\mathcal{D}_{out_{cp}}$			
	MSE	SSIM	LPIPS	Z-score	MSE	SSIM	LPIPS	Z-score
AE	66.44 [66.2, 66.7]	41.79 [41.6, 42.0]	32.05 [31.9, 32.3]	51.51 [51.3, 51.7]	69.03 [68.8, 69.2]	49.39 [49.2, 49.6]	46.42 [46.2, 46.6]	59.49 [59.3, 59.7]
DAE ($SNR_{0.1}$)	72.81 [72.6, 73.0]	88.96 [88.8, 89.1]	61.21 [61.0, 61.4]	85.51 [85.4, 85.7]	69.06 [68.9, 69.3]	84.23 [84.1, 84.4]	49.43 [49.2, 49.6]	75.60 [75.4, 75.8]
AAE	72.58 [72.4, 72.8]	89.35 [89.2, 89.5]	**84.63** [**84.5**, **84.8**]	86.53 [86.3, 86.7]	67.95 [67.7, 68.2]	81.41 [81.2, 81.5]	**73.44** [**73.3**, **73.6**]	75.81 [75.6, 76.0]
UNet-DAE ($SNR_{0.1}$)	**75.85** [**75.7**, **76.0**]	93.87 [93.8, 93.9]	70.31 [70.1, 70.5]	93.37 [93.2, 93.4]	71.37 [71.2, 71.6]	**88.65** [**88.5**, **88.7**]	60.55 [60.3, 60.7]	84.90 [84.7, 85.0]
F-AnoGAN	64.39 [64.2, 64.6]	73.79 [73.6, 74.0]	43.30 [43.1, 43.5]	65.41 [65.2, 65.7]	67.64 [67.4, 67.9]	73.63 [73.4, 73.8]	41.01 [40.8, 41.2]	66.77 [66.6, 67.0]
StyleGANv2	68.85 [68.3, 69.4]	72.67 [72.1, 73.2]	49.37 [48.8, 50.0]	85.21 [84.8, 85.6]	**73.32** [**72.8**, **73.8**]	73.11 [72.6, 73.7]	41.99 [41.4, 42.6]	83.45 [83.0, 83.8]
AnoDDPM ($SNR_{0.1}$)	68.82 [68.6, 69.0]	**94.56** [**94.5**, **94.6**]	60.66 [60.4, 60.9]	**94.13** [**94.1**, **94.2**]	68.75 [68.5, 69.0]	88.16 [88.0, 88.3]	51.16 [50.9, 51.4]	**86.93** [**86.8**, **87.1**]
AnoDDPM-PNDM ($SNR_{0.1}$)	68.34 [68.1, 68.6]	93.85 [93.8, 93.9]	67.23 [67.0, 67.4]	93.67 [93.6, 93.8]	67.97 [67.8, 68.2]	87.34 [87.2, 87.5]	56.13 [55.9, 56.3]	85.47 [85.3, 85.6]
AnoDDPM-PNDM (SNR_{subset})	71.74 [71.5, 71.9]	91.88 [91.8, 92.0]	73.21 [73.0, 73.4]	93.33 [93.2, 93.4]	69.71 [69.5, 79.0]	86.45 [86.3, 86.6]	60.81 [60.6, 61.0]	85.87 [85.7, 86.0]
AnoDDPM-PNDM ($U_{[0,1000]}$)	69.40 [68.7, 70.0]	93.13 [92.8, 93.3]	79.47 [78.9, 80.0]	93.04 [92.7, 93.2]	69.15 [68.4, 69.7]	87.98 [87.7, 88.3]	65.56 [64.9, 66.1]	85.04 [84.6, 85.4]

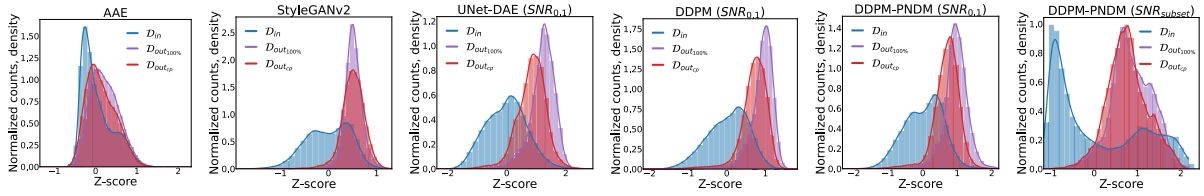

Fig. 5. Distributions of anomaly scores on benign and tumor patches from in-distribution $\mathcal{D}_{in}$ and out-of-distribution data with full tumor coverage and center-pixel tumor labels, $\mathcal{D}_{out_{100\%}}$ and $\mathcal{D}_{out_{cp}}$ respectively. A selection of methods and their Z-scores (higher values indicating greater reconstruction loss) are shown.

it using simple statistics from the anomaly heatmap. As input to the SVM, we create a feature vector containing the length of the major and minor axes and the mean Z-score value of the top 20 largest connected components in the heatmap, sorted by their mean Z-score value.

Similar to the patch-level experiments, we evaluate the task of OOD detection on a whole-slide level in two different scenarios: using all available tumor slides ($\mathcal{D}_{in}$ vs $\mathcal{D}_{out}$) and using only the subset of macrometastatic lesions ($\mathcal{D}_{in}$ vs $\mathcal{D}_{out_{macro}}$). To accommodate for a class imbalance (e.g. when evaluating the ability to detect macrometastatic tissue slides), we also include the results when evaluating area under the precision–recall curve (AUPR) for all WSI-level experiments.

4. Results

This section presents the results of the different OOD detection experiments. We start the analysis using patches extracted from whole-slide images before analyzing performance on a whole-slide level.

4.1. Patch-level OOD detection

In Fig. 4, we present example inputs and reconstructions of both in-distribution benign lymph node tissue patches and cancerous OOD tissue patches. Here, samples are taken from $\mathcal{D}_{in}$ and $\mathcal{D}_{out_{100\%}}$. We have left out the reconstructions for the regular AE for visualization's sake. The bottleneck for the AE is too large: the model was able to learn the identity function and reconstruct both in-distribution data and OOD samples with low reconstruction errors. A similar effect can be observed for both the DAE and UNet-DAE models, trained to denoise images with a noise level similar to the AnoDDPM model. The effect is best observed in the reconstructions of the UNet-DAE model, trained to denoise patches on a resolution of 256 × 256. Here, the model even faithfully denoises tumor cells in OOD samples. In contrast, the AnoDDPM model appears to alter the semantic properties of OOD tissue by replacing tumor cells with more benign-looking tissue. Although the stochastic backward process alters the content of in-distribution data slightly more than reconstructions of the UNet-DAE, the shift in semantics for OOD data supports the usefulness of AnoDDPM for detecting OOD data. The reconstructions by both GAN-based methods demonstrate significantly lower variability in their visual appearance and seem to suffer from mode collapse (Goodfellow et al., 2014). Furthermore, StyleGANv2 seems to suffer from color artifacts when reconstructing OOD samples. Overall, the quality of the reconstructed images of StyleGANv2 seems considerably lower compared to the image quality of fully synthesized images, where StyleGANv2 is comparable to the DDPM, see Fig. 6.

Table 2 reports the OOD detection performance for the different reconstruction-based methods. We measure the area under the ROC curve using every individual similarity metric and the Z-score average (6). Here, we evaluate the performance of two separate OOD detection tasks with different difficulty levels. The first four columns of Table 2 refer to the task of discriminating between patches from $\mathcal{D}_{in}$ and $\mathcal{D}_{out_{100\%}}$, with the OOD patches fully covered by tumor tissue. The remaining columns report the results for the slightly more difficult task of discriminating between in-distribution data and data from $\mathcal{D}_{out_{cp}}$. See Fig. 5 for a visualization of the corresponding Z-score distributions for a selection of methods. We observe that the baseline AE cannot effectively discriminate between benign and tumor patches in either experiment. In contrast, the DAE model based on the same architecture improves results significantly, increasing the AUC from 51.51 to 85.51 and 59.49 to 75.60 for both tasks respectively, using Z-score values. The AAE model reaches similar performances levels compared to the DAE in both experiments, but significantly outperforms the DAE when evaluating the LPIPS reconstruction error. In fact, the AAE reaches the highest performance values for this metric compared to any of the implemented methods. The DAE, based on the exact architecture of the DDPM (UNet-DAE), further improves results of the DAE with AUC values of 93.37 and 84.90 for both experiments respectively.

The performance of the two GAN-based methods is lower compared to the performance of the DAEs, except when evaluating the MSE on the second experiment. Here, the StyleGANv2 outperforms the F-AnoGAN model mainly when evaluating ROC analysis based on the Z-score values. However, the AUC values based on the individual metrics are similar between both GAN-based methods. These results show that aggregating errors of individual metrics in a Z-score can improve the OOD detection performance of the StyleGANv2 method more than

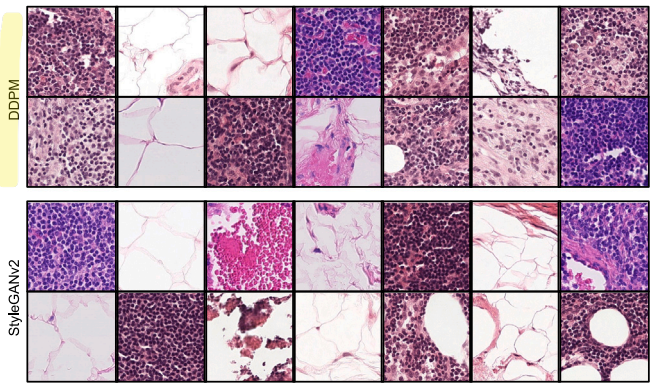

Fig. 6. Generated samples from the DDPM and the StyleGANv2 models.

Table 3
OOD detection performance on the two separate patch-level tasks: $\mathcal{D}_{in}$ vs $\mathcal{D}_{out_{100\%}}$ and $\mathcal{D}_{in}$ vs $\mathcal{D}_{out_{cp}}$. AUC values and corresponding confidence bounds are reported for each method using 2000-fold bootstrapping.

Model	$\mathcal{D}_{in}$ vs $\mathcal{D}_{out_{100\%}}$	$\mathcal{D}_{in}$ vs $\mathcal{D}_{out_{cp}}$
Fully supervised		
DenseNet	99.23 [99.2, 99.3]	98.91 [98.9, 98.9]
Distance-based		
AE-Mahalanobis	73.57 [73.4, 73.8]	71.53 [71.3, 71.7]
DAE-Mahalanobis	68.48 [68.3, 68.7]	71.52 [71.3, 71.7]
Deep pre-trained features		
Knowledge Distillation	78.70 [78.5, 78.9]	81.71 [81.5, 81.9]

the F-AnoGAN method, suggesting complementary errors across the individual similarity metrics.

Overall, these results show that the AnoDDPM slightly outperforms the other methods in our patch-level experiments. Regarding AUC values based on the Z-score metric, the AnoDDPM ($SNR_{0.1}$) model reaches the highest AUC values for both experiments: 94.13 and 86.93 respectively. However, for the second experiment, the UNet-DAE model reaches the highest AUC value of 88.65 when only considering SSIM. To analyze the impact of the pseudo-numerical sampling method, we evaluate the AnoDDPM model with and without using the PNDM sampler. Both results are included in Table 2, based on denoising images with a signal-to-noise ratio of 0.1. Results demonstrate that the PNDM sampler does not influence the OOD detection performance to a significant degree while substantially improving computational efficiency.

To include a comparison with a distance-based unsupervised OOD detector, we report the performance of the AE using the distance-based anomaly score in Table 3. We observe that the anomaly score based on a weighted sum of the Mahalanobis distance and the MSE improves the previously reported results of the AE in both experiments with AUC scores of 73.57 and 71.51 respectively. Here, using the autoencoder trained to denoise inputs instead, does not improve the OOD detection performance. The results of the multiresolution knowledge distillation method are also included in Table 3. We observe that the model, trained to distill knowledge from an ImageNet pre-trained expert network, outperforms the two distance based methods with AUC scores of 78.70 and 81.71.

To enable a comparison with the current standard for detecting breast cancer metastasis in Camelyon16, we evaluate a fully supervised DenseNet on the same data. The results are included in Table 3. We observe a near-perfect patch-level classification for the DenseNet trained to discriminate benign tissue from breast cancer metastasis with an AUC of up to 99.23.

The increased efficiency of the PNDM sampler allowed us to repeat both experiments and evaluate the OOD detection performance for AnoDDPM-PNDM settings where the backward process is defined by multiple timesteps, such as AnoDDPM-PNDM (SNR_{subset}) which defines the backward process based on the proposed subset of SNR values as depicted in Fig. 3. Similarly, we also evaluate the performance of the uniform sampling approach as proposed by Graham et al. (2022a), AnoDDPM-PNDM ($U_{[0,1000]}$). The results for both these methods are included in Table 2. Although the uniform sampling approach takes fifteen times more computational resources during inference (see Fig. 3), performance differences are mostly negligible. Here, the uniform sampling based method outperforms the AnoDDPM based on the proposed subset of SNR values when evaluating SSIM and LIPS reconstruction errors, however, these differences do not lead to a better AUC score based on the overall Z-scores.

To provide further insights into the inner workings of the AnoDDPM, we include a detailed ablation study to verify its performance for all possible values of t for the AnoDDPM-PNDM as well as different intervals to evaluate the uniform sampling approach. Note, throughout this ablation study we evaluate on a subset of 20k patches from each individual dataset as summarized in Table 1, reducing the total computational costs to a significant but more manageable 1135 GPU hours on a single A100 GPU. The resulting AUC values based on each individual similarity metric and the Z-score aggregate are plotted against the corresponding diffusion timesteps in the central panel of Fig. 7. Note, because the AnoDDPM-PNDM initiates the backward process at t as well as three subsequent timesteps in the diffusion process, we are only able to evaluate the AnoDDPM-PNDM up until the 970'th timestep. Overall we observe that the OOD detection performance, based on the SSIM value, benefits from adding more noise before denoising up to a maximum around a SNR of 0.05 at $t = 550$. ROC analysis based on MSE achieves the highest AUC values when denoising inputs with less noise. Performance levels based on the LIPS reconstruction error shows a similar trend, but this evaluation metrics seems to benefit from adding more noise in the forward process, with a maximum at $t = 970$. To evaluate the impact of the amount of remaining signal in the diffused image for the DAE baseline model and to compare it against the AnoDDPM, we retrain the DAE a total of 39 times, for all SNR values corresponding to $t \in [25, 50, 75, \ldots, 950, 975]$. The resulting AUC values based on each individual similarity metric and the Z-score aggregate are plotted against the corresponding diffusion timesteps in the left panel of Fig. 7. Similar to the AnoDDPM model, the DAE shows increased AUC values when increasing the amount of distortion in the early phase of the diffusion process ($t \leq 200$) for most metrics except the MSE. However, in contrast to the AnoDDPM-PNDM method, the DAE shows a more stabilized performance for any $t \geq 200$.

Overall, we observe that the Z-score for the AnoDDPM is most discriminative in both experiments at an SNR of 0.1, corresponding to $t = 485$. Therefore, we include the results based on denoising inputs corresponding to a SNR of 0.1 for the remainder of the experiments for both the AnoDDPM and DAEs.

Lastly, we evaluate the uniform sampling approach from Graham et al. (2022a) at different intervals for the AnoDDPM-PNDM ($U_{[t_{min},t_{max}]}$), with t_{min} and t_{max} defining the interval of the uniform sampling approach. The results for all possible intervals AnoDDPM-PNDM ($U_{[0,t]}$) are shown in the right panel of Fig. 7. Here, we observe that the OOD detection performance of the AnoDDPM-PNDM method generally benefits from adding multiple timesteps t in the partial denoising process. We observe the largest gains in performance in the early phase of the diffusion and denoising process ($t \leq 400$). Furthermore, we see that the uniform sampling approach can be impacted by the lower performance of the individual timesteps $t \geq 600$, shown in the central panel of Fig. 7.

Table 4 includes the results for all evaluation metrics across the two patch-level OOD detection experiments for five non-overlapping intervals and five overlapping cumulative intervals starting from $t_{min} = 0$. We include the previously reported results for the other AnoDDPM

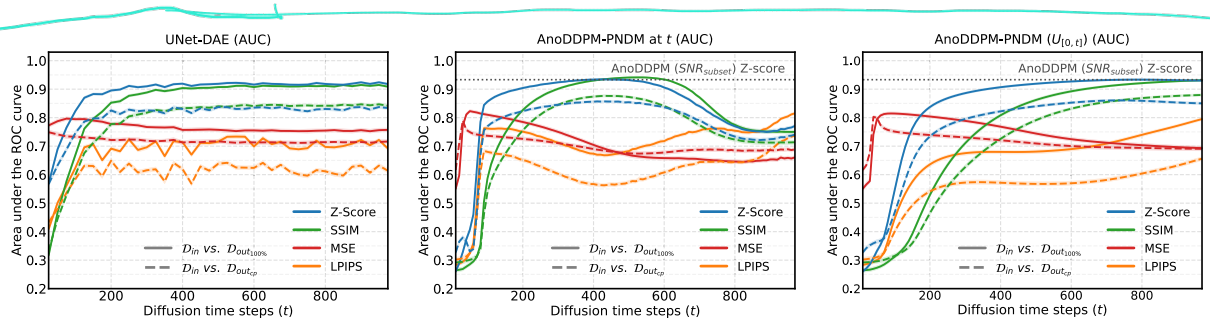

Fig. 7. AUC values of the $\mathcal{D}_{in}$ vs. $\mathcal{D}_{out_{cp}}$ (dashed lines) and the $\mathcal{D}_{in}$ vs. $\mathcal{D}_{out_{100\%}}$ (solid lines) patch-level OOD detection experiments for every individual similarity metric and the average Z-scores for the UNet-DAE (left), the AnoDDPM-PNDM for all possible signal-to-noise ratios (center), and the uniform denoising process of the entire Markov chain up until time t (right). Left: visualizes the results for independently trained DAEs, at time intervals of $t = 25$. Center, right: included are the results based on the Z-score metric of the AnoDDPM-PNDM (SNR_{subset}).

Table 4
Ablation experiments for various SNR ranges of the AnoDDPM. We repeat the two separate patch-level OOD detection tasks. AUC values are reported using each individual similarity metric, and the Z-score average across all metrics. Confidence bounds are reported for all methods using 2000-fold bootstrapping. We include the previously reported results for the fixed SNR of 0.1, as well as the proposed subset of SNR values.

Model	$\mathcal{D}_{in}$ vs. $\mathcal{D}_{out_{100\%}}$				$\mathcal{D}_{in}$ vs. $\mathcal{D}_{out_{cp}}$			
	MSE	SSIM	LPIPS	Z-score	MSE	SSIM	LPIPS	Z-score
AnoDDPM ($\text{SNR}_{0.1}$)	68.82 [68.6, 69.0]	**94.56** [**94.5**, **94.6**]	60.66 [60.4, 60.9]	**94.13** [**94.1**, **94.2**]	68.75 [68.5, 69.0]	**88.16** [**88.0**, **88.3**]	51.16 [50.9, 51.4]	**86.93** [**86.8**, **87.1**]
AnoDDPM-PNDM ($\text{SNR}_{0.1}$)	68.34 [68.1, 68.6]	93.85 [93.8, 93.9]	67.23 [67.0, 67.4]	93.67 [93.6, 93.8]	67.97 [67.8, 68.2]	87.34 [87.2, 87.5]	56.13 [55.9, 56.3]	85.47 [85.3, 85.6]
AnoDDPM-PNDM ($U_{[0,200]}$)	**80.40** [**79.8**, **81.0**]	52.87 [52.1, 53.6]	63.60 [62.8, 64.2]	83.73 [83.2, 84.2]	**73.90** [**73.2**, **74.6**]	46.49 [45.7, 47.1]	54.11 [53.4, 54.9]	73.42 [72.8, 74.]
AnoDDPM-PNDM ($U_{[200,400]}$)	74.58 [73.9, 75.2]	90.03 [89.6, 90.4]	71.77 [71.1, 72.3]	92.56 [92.2, 92.9]	70.85 [70.1, 71.5]	84.43 [83.9, 84.9]	60.49 [59.8, 61.2]	84.91 [84.5, 85.3]
AnoDDPM-PNDM ($U_{[400,600]}$)	67.22 [66.5, 67.9]	94.15 [93.8, 94.4]	70.82 [70.2, 71.4]	93.96 [93.6, 94.3]	67.69 [66.9, 68.3]	87.66 [87.2, 88.1]	58.02 [57.2, 58.7]	86.44 [86.0, 86.8]
AnoDDPM-PNDM ($U_{[600,800]}$)	65.12 [64.3, 65.7]	89.37 [89.0, 89.7]	**81.76** [**81.2**, **82.2**]	85.21 [84.8, 85.7]	68.25 [67.5, 68.9]	78.17 [77.5, 78.7]	68.61 [68.0, 69.2]	79.23 [78.7, 79.7]
AnoDDPM-PNDM ($U_{[800,1000]}$)	64.69 [64.0, 65.3]	75.35 [74.6, 76.0]	81.24 [80.7, 81.7]	76.05 [75.7, 76.7]	68.53 [67.8, 69.1]	71.37 [70.6, 72.0]	**70.85** [**70.2**, **71.4**]	72.87 [72.2, 73.6]
AnoDDPM-PNDM ($U_{[0,200]}$)	**80.40** [**79.8**, **81.0**]	52.87 [52.1, 53.6]	63.60 [62.8, 64.2]	83.73 [83.2, 84.2]	**73.90** [**73.2**, **74.6**]	46.49 [45.7, 47.1]	54.11 [53.4, 54.9]	73.42 [72.8, 74.]
AnoDDPM-PNDM ($U_{[0,400]}$)	76.83 [76.2, 77.4]	80.83 [80.2, 81.3]	67.98 [67.3, 68.6]	90.68 [90.3, 91.0]	72.07 [71.3, 72.7]	72.56 [71.9, 73.1]	57.16 [56.4, 57.9]	82.22 [81.7, 82.6]
AnoDDPM-PNDM ($U_{[0,600]}$)	72.45 [71.8, 73.0]	89.24 [88.8, 89.7]	68.97 [68.2, 69.6]	92.70 [92.3, 93.0]	70.04 [69.3, 70.7]	83.15 [82.6, 83.6]	56.98 [56.3, 57.7]	85.26 [84.8, 85.6]
AnoDDPM-PNDM ($U_{[0,800]}$)	70.34 [69.7, 71.0]	92.11 [91.8, 92.4]	73.98 [73.3, 74.5]	93.33 [93.0, 93.6]	69.37 [68.6, 70.1]	86.87 [86.4, 87.3]	60.38 [59.7, 61.0]	85.92 [85.5, 86.3]
AnoDDPM-PNDM ($U_{[0,1000]}$)	69.40 [68.7, 70.0]	93.13 [92.8, 93.3]	79.47 [78.9, 80.0]	93.04 [92.7, 93.2]	69.15 [68.4, 69.7]	87.98 [87.7, 88.3]	65.56 [64.9, 66.1]	85.04 [84.6, 85.4]
AnoDDPM-PNDM (SNR_{subset})	71.74 [71.5, 71.9]	91.88 [91.8, 92.0]	73.21 [73.0, 73.4]	93.33 [93.2, 93.4]	69.71 [69.5, 79.0]	86.45 [86.3, 86.6]	60.81 [60.6, 61.0]	85.87 [85.7, 86.0]

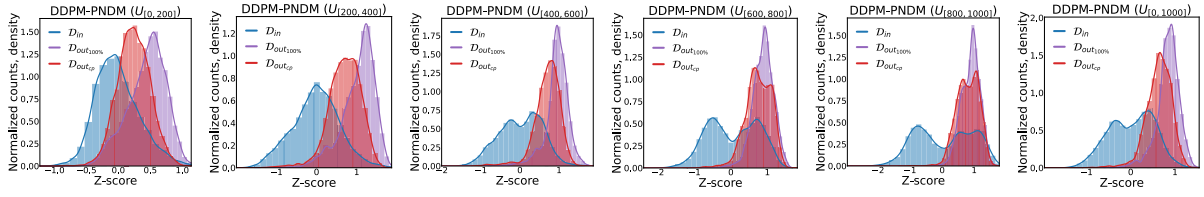

Fig. 8. Distributions of anomaly scores on datasets: $\mathcal{D}_{in}$, $\mathcal{D}_{out_{100\%}}$ and $\mathcal{D}_{out_{cp}}$ for each non-overlapping uniform time interval as well as the full Markov chain.

models for comparison. See Fig. 8 for a visualization of the corresponding Z-score distributions for a selection of time intervals and the full interval $U_{[0,1000]}$. For both patch-level experiments, we observe the best OOD performance based on the SSIM and Z-score for the intervals $U_{[200,400]}$ and $U_{[400,600]}$. Consequently, including these two regions also results in the largest relative performance gains in the cumulative interval $U_{[0,t]}$. Similar to the results observed when evaluating the OOD performance based on individual timesteps (see Fig. 7), the MSE and LPIPS values are more effective in earlier and later time intervals respectively.

4.2. WSI-level OOD detection

To evaluate the full capability of unsupervised approaches in digital pathology, we will evaluate a selection of methods on a whole-slide image level and compare their performances with the current standard: supervised learning. Table 6 reports the main results.

Although the PNDM sampler improves the computational efficiency of the AnoDDPM model, inference on all 129 WSIs from the Camelyon16 test set remains challenging. Table 5 shows the computational budget required to evaluate the entire Camelyon16 test set for the different settings of AnoDDPM, in relation to the current standard based on supervised learning. Here, we report the required GPU days when maximizing the utility of a single A100 GPU. We do so by extrapolating

Table 5
Inference time per method in GPU days, measured by extrapolating the time for a single batch to the total number of batches required to evaluate all tissue patches in the Camelyon16 test set, independent from the overhead of reading tissue patches from all slides. Estimates are included for a sliding window approach with overlapping and non-overlapping inputs, resulting in output heatmaps with pixel resolutions of 7.24 µm/px and 115.87 µm/px respectively.

Model	7.24 µm/px	115.87 µm/px
Fully supervised		
DenseNet	**1.473**	**0.006**
Unsupervised AnoDDPM		
AnoDDPM ($\text{SNR}_{0.1}$)	34519	135
AnoDDPM-PNDM ($\text{SNR}_{0.1}$)	502	**1.96**
AnoDDPM-PNDM ($U_{[0,1000]}$)	529408	2068
AnoDDPM-PNDM (SNR_{subset})	35223	138

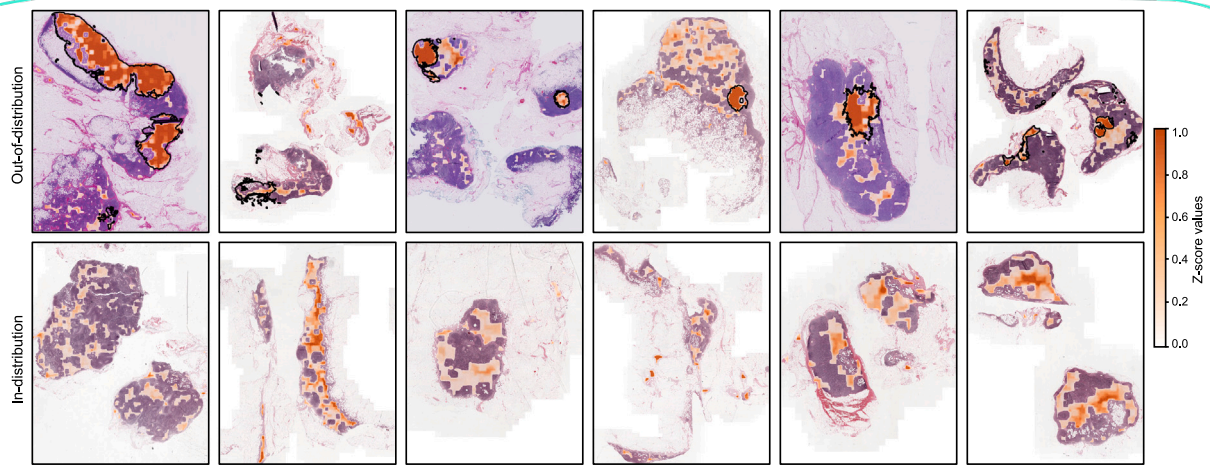

Fig. 9. Z-score heatmaps from AnoDDPM-PNDM ($SNR_{0.1}$) on a whole-slide image level for six random samples from $\mathcal{D}_{out_{macro}}$ (top) and $\mathcal{D}_{in}$ (bottom). Pixel-level ground truth annotations are available and included (top). Most macro-metastasis are flagged with high Z-scores. However, the second column contains an example where a big tumor lesion remains largely undetected. Furthermore, false-positive detections seem to influence the slide-level OOD detection performance.

Table 6
Detecting the presence of breast cancer metastasis on a WSI level. Here, the unsupervised methods are compared to a fully supervised network (evaluated on two different output resolutions) trained to detect breast cancer metastasis on a patch level. Both AUC and AUPR values are included with confidence bounds using 2000-fold bootstrapping. The AnoDDPM and DAE models are trained to denoise images with a SNR of 0.1. Results are included for a sliding window approach with overlapping and non-overlapping input patches, resulting in output heatmaps with pixel resolutions of 7.24 µm/px and 115.87 µm/px respectively.

Model	$\mathcal{D}_{in}$ / $\mathcal{D}_{out}$		$\mathcal{D}_{in}$ / $\mathcal{D}_{out_{macro}}$	
	AUROC	AUPR	AUROC	AUPR
Fully supervised				
DenseNet (output @ 7.24 µm/px)	97.37 [94.3, 99.6]	97.12 [93.6, 99.5]	100.0 [100, 100]	100.0 [100, 100]
DenseNet (output @ 115.87 µm/px)	94.86 [90.2, 98.6]	94.87 [90.0, 98.3]	97.51 [91.3, 100]	100.0 [100, 100]
Unsupervised OOD detection (output @ 115.87 µm/px)				
UNet-DAE ($SNR_{0.1}$) (max z-score)	47.28 [39.5, 56.6]	39.19 [30.1, 48.7]	61.47 [51.1, 71.1]	29.72 [19.8, 42.4]
UNet-DAE ($SNR_{0.1}$) (99th percentile z-score)	59.44 [50.8, 68.1]	50.65 [39.2, 62.5]	78.05 [69.4, 86.0]	50.18 [34.2, 67.1]
UNet-DAE ($SNR_{0.1}$) (OneClassSVM)	63.48 [55.1, 71.8]	71.19 [61.8, 80.5]	68.15 [55.3, 79.8]	84.92 [76.6, 92.9]
AnoDDPM-PNDM ($SNR_{0.1}$) (max z-score)	55.24 [46.6, 63.7]	46.13 [35.2, 57.3]	75.31 [66.6, 82.8]	42.87 [28.6, 58.7]
AnoDDPM-PNDM ($SNR_{0.1}$) (99th percentile z-score)	58.24 [49.6, 67.1]	50.69 [39.3, 62.9]	**83.28** [**75.8**, **90.1**]	56.61 [39.6, 74.6]
AnoDDPM-PNDM ($SNR_{0.1}$) (OneClassSVM)	**67.21** [**59.4**, **75.0**]	**78.47** [**70.1**, **85.5**]	78.16 [67.0, 88.4]	**90.01** [**82.6**, **96.5**]

the costs of evaluating a single batch to the total batches required to evaluate all tissue patches in the Camelyon16 test set. Specifically, we include the costs for two settings resulting in different pixel resolutions of the predicted output heatmap. In the first setting, the model is used in a sliding-window approach with overlapping windows, as is normal for patch-based methods on the Camelyon16 dataset (Linmans et al., 2023), resulting in a pixel resolution of 7.24 µm/px. The second scenario corresponds to a sliding window approach, where the stride is equal to the size of the input patch, resulting in non-overlapping windows and an output pixel resolution of 115.87 µm/px. As shown, the PNDM sampler significantly reduces the computational budget required to evaluate on the full test set of Camelyon16. Due to the significant costs of evaluating AnoDDPM on the full test set using a denoising process defined by multiple timesteps, we limit the analysis by only evaluating the model based on inputs with an SNR of 0.1 for non-overlapping input patches. Doing so, the input at a pixel spacing of 0.48 µm is reduced to an anomaly score heatmap with a pixel spacing of 115.87 µm, making micrometastatic lesions difficult to detect.

Fig. 9 presents a visual overview of the Z-score anomaly heatmaps from AnoDDPM-PNDM ($SNR_{0.1}$) for 12 random whole-slide images from both $\mathcal{D}_{out_{macro}}$ (top) and $\mathcal{D}_{in}$ (bottom). We observe that most macrometastatic tissue regions are flagged with high Z-scores. However, most smaller and some larger tumor lesions remain largely undetected. Furthermore, the set of benign WSIs also shows that false-positive detections impact performance.

Table 6 reports the slide-level OOD detection performance of the AnoDDPM model and the fully supervised DenseNet. We also include the best-performing baseline: the denoising autoencoder based on the exact same architecture and noise levels as the DDPM model (UNet-DAE). We report AUC and AUPR, the area under the precision–recall curve, values using two slide-level statistics and the output of an auxiliary one-class SVM (Section 3.8). We observe that the AnoDDPM can discriminate between macrometastatic WSIs and benign WSIs with an AUPR of up to 90.01 and a corresponding AUC of 78.16. Although confidence intervals overlap, the UNet-DAE performs slightly worse with an AUPR of 84.92 and a corresponding AUC of 78.05. When using the output of the auxiliary one-class SVM, the AnoDDPM achieves an AUC of 67.21 and a corresponding AUPR of 78.47 for the complete Camelyon16 test set, slightly outperforming the UNet-DAE baseline. Differences between the AUC and AUPR values seems largest on the task of detecting macrometastatic tissue slides, which can be explained by the more skewed distribution of positives for this task: only ≈ 17% of the test set contains macrometastatic lesions. Overall, the performance gap with the fully supervised DenseNet model remains considerable, even when evaluating the DenseNet model with an output pixel spacing of 115.87 µm.

5. Discussion

In our experiments, we have demonstrated that denoising diffusion probabilistic models can be used for out-of-distribution detection to discriminate breast cancer metastasis from benign lymph node tissue when trained purely on benign data. By leveraging pseudo-numerical

sampling methods, we were able to apply AnoDDPM on a large enough scale for whole-slide image analysis on the complete test set of the Camelyon16 challenge. This allowed us to compare the performance of the unsupervised anomaly detectors with the current standard for this task, which is based on supervised learning.

Our results show that reconstruction-based methods that aim to denoise inputs achieve the best unsupervised OOD detection performance in the setting of the Camelyon16 challenge. When comparing the AnoDDPM method with the denoising autoencoder based on the same architecture, the AnoDDPM only marginally outperforms the computationally efficient UNet-DAE. However, the advantage of the AnoDDPM is that a single trained model can evaluate reconstruction errors for all SNR values corresponding to all $T = 1000$ timesteps. In other words, the AnoDDPM is flexible towards different information bottlenecks. In contrast, the UNet-DAE is trained to denoise a specific amount of noise.

In an extensive ablation study, we have shown that the optimal signal-to-noise ratio for the AnoDDPM is specific for each similarity metric. Future work might even shown that the optimal SNR is dependent on the size of the artifacts. However, the ability of AnoDDPM to evaluate reconstruction errors for inputs with different SNRs removes the need to find unknown SNR-specific optima for each similarity metric. Instead, we have shown that we can obtain approximate optimal discriminative power of the individual similarity metrics by evaluating the reconstruction errors at multiple SNRs. As we did in our work, limiting the SNR to values close to 1.0 might be vital in finding optimal OOD detection strategies based on denoising methods that generalize to other OOD detection tasks.

The UNet-DAE and the AnoDDPM both show good unsupervised OOD detection performance, but there are distinct differences when visualizing the reconstructed images. The UNet-DAE has powerful denoising capabilities; it even reconstructs tumor cells. In contrast, the AnoDDPM alters the semantics of the reconstructed tumor patches by replacing tumor cells with benign-looking tissue. Although this may not be unique to the AnoDDPM model, other types of reconstruction based methods that are not included in this work might also be able to replace anomalies with more benign looking tissue. Based on the visuals included in this work, it is perhaps surprising that performance is similar between the UNet-DAE and the AnoDDPM. This raises questions about how to evaluate the similarities between the input and the reconstruction effectively. To address this, future work should focus on incorporating other similarity metrics that better capture the semantic properties of images. The LPIPS reconstruction error based on features from a pretrained network seems like a natural fit (Zhang et al., 2018). However, in our results, the LPIPS metric using pretrained Imagenet weights (Krizhevsky et al., 2012) performed poorly in terms of its discriminative power. This could be because the pretrained Imagenet weights might not be optimal for detecting small deviations between different tissue types in digital pathology. Therefore, future work should consider using the LPIPS metric with weights pretrained on a pathology dataset. Similarly, any other general tissue classifier or cell segmentation method could be used in future work as a basis for a more discriminative similarity metric.

In our analysis of GAN-based reconstruction methods, we observed that the reconstruction quality was inferior to that of the AnoDDPM, even though the quality of fully synthesized images is similar. Additionally, the reconstructions from both GAN-based methods suffered from mode collapse (Goodfellow et al., 2014). These observations are in line with earlier work (Karras et al., 2020). The original authors showed that although images generated by StyleGANv2 can be projected almost perfectly back into generator inputs, projected real images (from the training set) show apparent differences from the originals (Karras et al., 2020). Furthermore, although f-AnoGAN leverages the Wasserstein formula to reduce issues of mode collapse, collapse can still occur as observed in earlier work (Sayeri et al., 2018). Perhaps other strategies presented in Sayeri et al. (2018) can further reduce these issues in future work.

To get a more realistic estimate of the performance of unsupervised OOD detection in digital pathology, we also evaluate WSI-level tasks. Here, results demonstrate that there is still a significant performance gap between unsupervised and supervised methods. The UNet-DAE and AnoDDPM methods showed lower AUC values for the slide-level experiments than the patch-level experiments. It is worth noting that patch-based performance on a random subset of patches does not perfectly translate to slide-level performance, as the slide-level anomaly score can easily be influenced by small false positive or false negative regions. Our results indicate that while the AnoDDPM method performed well in detecting macrometastatic lesions, it struggled with false positive detections on benign WSIs and failed to detect most micrometastases. However, the AUC and AUPR analysis based on discriminating between benign slides and macrometastatic WSIs resulted in AUC and AUPR values up to 78.16 and 90.01 respectively.

To further improve our results, future work could reduce false positive detections by fine-tuning the AnoDDPM model on hard negative regions in the training set. Additionally, incorporating more SNRs when evaluating the AnoDDPM on whole-slide images may enhance performance. Finally, increasing the output resolution of the sliding-window approach may improve the detection of smaller tumor lesions. To address the increased computational demands, reducing the number of timesteps in the backward process with the PNDM sampler may be necessary. Future work should analyze the impact of further reducing the timestep schedule. Recent work has shown that diffusion based anomaly detection methods can be applied to the latent representations learnt by autoencoders, making them more scalable for application to high dimensional data, such as medical images (Pinaya et al., 2022). Future work should try to combine the use of latent representations with the use of the PNDM sampler to further improve computational efficiency of the AnoDDPM. By implementing these improvements, we may be able to narrow the performance gap with current supervised learning methods in digital pathology.

6. Conclusion

This article extensively evaluates unsupervised out-of-distribution detection methods in digital pathology, analyzing their performance on patch-level and whole-slide image level tasks. We include a method based on denoising diffusion probabilistic models where we effectively replace the image generation objective with a reconstruction objective: we begin the backward process with a partially-diffused image and then generate an image based on any remaining signal. Based on the observation that the AnoDDPM is flexible towards different information bottlenecks, we evaluate the reconstruction errors for inputs with different signal-to-noise ratios.

Throughout different experiments, we gain insights into the feasibility of building a universal anomaly detector trained only on benign data without requiring disease-specific annotations. Our results showed important limitations that resulted in a significant performance gap compared to the current standard in breast cancer metastasis detection. Although the visual reconstructions of the AnoDDPM indicate that it can replace out-of-distribution tissue with benign-looking tissue, we currently lack a similarity metric that can detect these subtle semantic changes in the input. Despite these limitations, our results demonstrate that the AnoDDPM can detect OOD data on both a patch and WSI level. To improve the results, we suggest further exploring techniques of fast sampling methods to enable a more extensive analysis at higher output resolutions. Similarly, hard negative mining might be crucial to reduce the effect of false positive detections on a slide-level scale. Ideally, the performance gap with supervised methods is further reduced. If so, unsupervised anomaly detectors may eventually complement or replace the standard of supervised methods by simply training on a library of benign tissue and detecting anything that deviates from normal.

CRediT authorship contribution statement

Jasper Linmans: Conceptualization, Formal analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – original draft. **Gabriel Raya:** Conceptualization, Methodology, Software. **Jeroen van der Laak:** Data curation, Funding acquisition, Writing – review & editing, Project administration, Supervision. **Geert Litjens:** Conceptualization, Data curation, Supervision, Writing – review & editing.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Jeroen van der Laak reports a relationship with Philips that includes: board membership and funding grants. Jeroen van der Laak reports a relationship with ContextVision AB that includes: board membership and funding grants. Jeroen van der Laak reports a relationship with Sectra AB that includes: funding grants. Jeroen van der Laak reports a relationship with Aiosyn BV that includes: employment and equity or stocks. Geert Litjens reports a relationship with Philips that includes: funding grants. Geert Litjens reports a relationship with Canon Health Informatics (USA) that includes: consulting or advisory. Geert Litjens reports a relationship with Aiosyn BV that includes: equity or stocks.

Data availability

Data will be made available on request.

Acknowledgments

We would like to thank Dejan Štepec for the help with the GAN-based models. Jeroen van der Laak and Geert Litjens have received funding from the Innovative Medicines Initiative 2 Joint Undertaking under grant agreement No 945358. This Joint Undertaking receives support from the European Union's Horizon 2020 research and innovation program and EFPIA. Geert Litjens also received funding by a grant from the Dutch Cancer Society (KWF), grant number KUN 2015-7970 and the Dutch Research Council (NWO) under Veni grant number 91618152. Gabriel Raya was funded by the Dutch Research Council (NWO) as part of the CERTIF-AI project (file number 17998). The Knut and Alice Wallenberg foundation is acknowledged for generous support.

References

Arjovsky, M., Chintala, S., Bottou, L., 2017. Wasserstein GAN. In: Proc. of the 34th Int. Conf. on Mach. Learn. ICML, http://dx.doi.org/10.48550/ARXIV.1701.07875.

Baur, C., Graf, R., Wiestler, B., Albarqouni, S., Navab, N., 2020. Steganomaly: Inhibiting cyclegan steganography for unsupervised anomaly detection in brain mri. In: Proc. Int. Conf. on Med. Im. Compt. and Compt-Assist. Interv. pp. 718–727. http://dx.doi.org/10.1007/978-3-030-59713-9_69.

Baur, C., Wiestler, B., Albarqouni, S., Navab, N., 2019. Deep autoencoding models for unsupervised anomaly segmentation in brain MR images. In: International MICCAI Brainlesion Workshop. pp. 161–169. http://dx.doi.org/10.1007/978-3-030-11723-8_16.

Bergmann, P., Fauser, M., Sattlegger, D., Steger, C., Uninformed students: Student-teacher anomaly detection with discriminative latent embeddings. In: IEEE Conf. on Comp. Vis. and Pat. Recog. Workshops. CVPR, URL: http://arxiv.org/abs/1911.02357.

Choi, H., Jang, E., 2019. Generative ensembles for robust anomaly detection. http://dx.doi.org/10.48550/ARXIV.1810.01392, arXiv preprint arXiv:1810.01392.

Denouden, T., Salay, R., Czarnecki, K., Abdelzad, V., Phan, B., Vernekar, S., 2018. Improving reconstruction autoencoder out-of-distribution detection with mahalanobis distance. http://dx.doi.org/10.48550/ARXIV.1812.02765, arXiv preprint arXiv:1812.02765.

Ehteshami Bejnordi, B., Veta, M., Johannes van Diest, P., van Ginneken, B., Karssemeijer, N., Litjens, G., van der Laak, J., the CAMELYON16 Consortium, 2017. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. JAMA 318 (22), 2199–2210. http://dx.doi.org/10.1001/jama.2017.14585.

Fernando, T., Gammulle, H., Denman, S., Sridharan, S., Fookes, C., 2021. Deep learning for medical anomaly detection – A survey. ACM Comput. Surv. 54 (7), http://dx.doi.org/10.1145/3464423.

Fox, J.P., Grignol, V.P., Gustafson, J., Cheng, P., Weighall, R., Ouellette, J., Hellan, M., Dowdy, Y., Termuhlen, P., 2010. Incidental lymphoma during sentinel lymph node biopsy for breast cancer. J. Clin. Oncol. 28 (15), e11083. http://dx.doi.org/10.1200/jco.2010.28.15_suppl.e11083.

Gong, D., Liu, L., Le, V., Saha, B., Mansour, M.R., Venkatesh, S., Hengel, A.v.d., 2019. Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection. In: IEEE Int. Conf. on Comp. Vis. ICCV, http://dx.doi.org/10.48550/ARXIV.1904.02639.

Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y., 2014. Generative adversarial networks. http://dx.doi.org/10.48550/ARXIV.1406.2661, arXiv preprint arXiv:1406.2661.

Graham, M.S., Pinaya, W.H.L., Tudosiu, P.-D., Nachev, P., Ourselin, S., Cardoso, M.J., 2022a. Denoising diffusion models for out-of-distribution detection. http://dx.doi.org/10.48550/ARXIV.2211.07740, arXiv preprint arXiv:2211.07740.

Graham, M.S., Tudosiu, P.D., Wright, P., Pinaya, W.H.L., U-King-Im, J.M., Mah, Y., Teo, J., Jäger, R.H., Werring, D., Nachev, P., Ourselin, S., Cardoso, M.J., 2022b. Transformer-based out-of-distribution detection for clinically safe segmentation. In: Medical Imaging with Deep Learning. URL: https://openreview.net/forum?id=En7660i-CLJ.

Guha Roy, A., Ren, J., Azizi, S., Loh, A., Natarajan, V., Mustafa, B., Pawlowski, N., Freyberg, J., Liu, Y., Beaver, Z., Vo, N., Bui, P., Winter, S., MacWilliams, P., Corrado, G.S., Telang, U., Liu, Y., Cemgil, T., Karthikesalingam, A., Lakshminarayanan, B., Winkens, J., 2022. Does your dermatology classifier know what it doesn't know? Detecting the long-tail of unseen conditions. Med. Image Anal. 75, 102274. http://dx.doi.org/10.1016/j.media.2021.102274.

Ho, J., Jain, A., Abbeel, P., 2020. Denoising diffusion probabilistic models. http://dx.doi.org/10.48550/ARXIV.2006.11239, arXiv preprint arXiv:2006.11239.

Huang, G., Liu, Z., van der Maaten, L., Weinberger, K.Q., 2017. Densely connected convolutional networks. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition. CVPR, pp. 2261–2269. http://dx.doi.org/10.1109/CVPR.2017.243.

Karras, T., Aila, T., Laine, S., Lehtinen, J., 2018. Progressive growing of GANs for improved quality, stability, and variation. In: International Conference on Learning Representations. URL: https://openreview.net/forum?id=Hk99zCeAb.

Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., Aila, T., 2020. Analyzing and improving the image quality of stylegan. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8110–8119. http://dx.doi.org/10.48550/ARXIV.1912.04958.

Kingma, D.P., Ba, J., 2015. Adam: A method for stochastic optimization. In: Proceedings of the International Conference on Learning Representations. San Diego, USA, http://dx.doi.org/10.48550/ARXIV.1412.6980.

Kingma, D.P., Salimans, T., Poole, B., Ho, J., 2021. On density estimation with diffusion models. In: Advances in Neural Information Processing Systems. URL: https://openreview.net/forum?id=2LdBqxc1Yv.

Kingma, D.P., Welling, M., 2014. Auto-encoding variational Bayes. In: Proc. Int. Conf. Learn. Represent. ICLR, Banff, AB, Canada, http://dx.doi.org/10.48550/ARXIV.1312.6114.

Kompa, B., Snoek, J., Beam, A.L., 2021. Second opinion needed: communicating uncertainty in medical machine learning. npj Digit. Med. 4, 1–6. http://dx.doi.org/10.1038/s41746-020-00367-3.

Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. ImageNet classification with deep convolutional neural networks. In: Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1. http://dx.doi.org/10.1145/3065386.

Linmans, J., Elfwing, S., van der Laak, J., Litjens, G., 2023. Predictive uncertainty estimation for out-of-distribution detection in digital pathology. Med. Image Anal. http://dx.doi.org/10.1016/j.media.2022.102655.

Litjens, G., Bandi, P., Ehteshami Bejnordi, B., Geessink, O., Balkenhol, M., Bult, P., Halilovic, A., Hermsen, M., van de Loo, R., Vogels, R., et al., 2018. H&e-stained sentinel lymph node sections of breast cancer patients: the CAMELYON dataset. GigaScience 7 (6), http://dx.doi.org/10.1093/gigascience/giy065.

Liu, L., Ren, Y., Lin, Z., Zhao, Z., 2022. Pseudo numerical methods for diffusion models on manifolds. In: Int. Conf. on Learn. Repr. URL: https://openreview.net/forum?id=PlKWVd2yBkY.

Makhzani, A., Shlens, J., Jaitly, N., Goodfellow, I., 2016. Adversarial autoencoders. In: International Conference on Learning Representations. URL: http://arxiv.org/abs/1511.05644.

Nalisnick, E.T., Matsukawa, A., Teh, Y.W., Görür, D., Lakshminarayanan, B., 2019. Do deep generative models know what they don't know? In: Proc. Int. Conf. Learn. Represent. http://dx.doi.org/10.48550/ARXIV.1810.09136.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al., 2011. Scikit-learn: Machine learning in python. J. Mach. Learn. Res. 12, 2825–2830. http://dx.doi.org/10.48550/ARXIV.1201.0490.

Pinaya, W.H.L., Graham, M.S., Gray, R., Da Costa, P.F., Tudosiu, P.-D., Wright, P., Mah, Y.H., MacKinnon, A.D., Teo, J.T., Jager, R., Werring, D., Rees, G., Nachev, P., Ourselin, S., Cardoso, M.J., 2022. Fast unsupervised brain anomaly detection and segmentation with diffusion models. http://dx.doi.org/10.48550/ARXIV.2206.03461, arXiv preprint arXiv:2206.03461.

Pocevičiūtė, M., Eilertsen, G., Jarkman, S., Lundström, C., 2022. Generalisation effects of predictive uncertainty estimation in deep learning for digital pathology. Sci. Rep. 12 (1), 1–15. http://dx.doi.org/10.1038/s41598-022-11826-0.

Pocevičiūtė, M., Eilertsen, G., Lundström, C., 2021. Unsupervised anomaly detection in digital pathology using GANs. In: 2021 IEEE 18th Int. Symp. on Biom. Imag. ISBI, IEEE, http://dx.doi.org/10.48550/ARXIV.2103.08945.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation. In: Proc. Int. MICCAI. Munich, BY, Germany, http://dx.doi.org/10.48550/ARXIV.1505.04597.

Roth, K., Pemula, L., Zepeda, J., Schölkopf, B., Brox, T., Gehler, P.V., 2021. Towards total recall in industrial anomaly detection. In: IEEE Conf. on Comp. Vis. and Pat. Recog. Workshops. CVPR, URL: https://arxiv.org/abs/2106.08265.

Salehi, M., Mirzaei, H., Hendrycks, D., Li, Y., Rohban, M.H., Sabokrou, M., 2021. A unified survey on anomaly, novelty, open-set, and out-of-distribution detection: Solutions and future challenges. http://dx.doi.org/10.48550/ARXIV.2110.14051, arXiv preprint arXiv:2110.14051.

Salehi, M., Sadjadi, N., Baselizadeh, S., Rohban, M.H., Rabiee, H.R., 2020. Multiresolution knowledge distillation for anomaly detection.

Sayeri, L., Maha, S., Anastasiya, B., Molei, L., 2018. Evaluation of mode collapse in generative adversarial networks. In: IEEE High Perf. Extr. Comp. Conf. HPEC, URL: https://api.semanticscholar.org/CorpusID:214622026.

Schlegl, T., Seeböck, P., Waldstein, S.M., Langs, G., Schmidt-Erfurth, U., 2019. F-AnoGAN: Fast unsupervised anomaly detection with generative adversarial networks. Med. Image Anal. 54, 30–44. http://dx.doi.org/10.1016/j.media.2019.01.010.

Schölkopf, B., Williamson, R.C., Smola, A., Shawe-Taylor, J., Platt, J., 1999. Support vector method for novelty detection. Adv. Neural Inf. Process. Syst. 12.

Schömig-Markiefka, B., Pryalukhin, A., Hulla, W., Bychkov, A., Fukuoka, J., Madabhushi, A., Achter, V., Nieroda, L., Büttner, R., Quaas, A., Tolkach, Y., 2021. Quality control stress test for deep learning-based diagnostic model in digital pathology. Mod. Pathol. 34, 1–11. http://dx.doi.org/10.1038/s41379-021-00859-x.

Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., Ganguli, S., 2015. Deep unsupervised learning using nonequilibrium thermodynamics. In: Proceedings of the 32nd International Conference on Machine Learning. pp. 2256–2265. http://dx.doi.org/10.48550/ARXIV.1503.03585.

Song, Y., Ermon, S., 2019. Generative modeling by estimating gradients of the data distribution. Adv. Neural Inf. Process. Syst. http://dx.doi.org/10.48550/ARXIV.1907.05600.

Song, J., Meng, C., Ermon, S., 2021. Denoising diffusion implicit models. In: International Conference on Learning Representations. URL: https://openreview.net/forum?id=St1giarCHLP.

Stepec, D., Skocaj, D., 2021. Unsupervised detection of cancerous regions in histology imagery using image-to-image translation. In: IEEE Conf. on Comp. Vis. and Pat. Recog. Workshops. CVPR Workshops 2021, Virtual, June 19-25, 2021, pp. 3785–3792. http://dx.doi.org/10.1109/CVPRW53098.2021.00419.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I., 2017. Attention is all you need. Adv. Neural Inf. Process. Syst. 30, http://dx.doi.org/10.48550/ARXIV.1706.03762.

Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y., Manzagol, P.A., Bottou, L., 2010. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. J. Mach. Learn. Res. 11 (12), URL: http://jmlr.org/papers/v11/vincent10a.html.

Virtanen, P., Gommers, R., Oliphant, T.E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., et al., 2020. SciPy 1.0: Fundamental algorithms for scientific computing in Python. Nature Methods 17, 261–272. http://dx.doi.org/10.1038/s41592-019-0686-2.

Wang, Z., Bovik, A., Sheikh, H., Simoncelli, E., 2004. Image quality assessment: from error visibility to structural similarity. IEEE Trans. Image Process. 13 (4), 600–612. http://dx.doi.org/10.1109/TIP.2003.819861.

Wei, X., Liu, Z., Wang, L., Gong, B., 2018. Improving the improved training of wasserstein GANs. In: International Conference on Learning Representations. URL: https://openreview.net/forum?id=SJx9GQb0-.

Wyatt, J., Leach, A., Schmon, S.M., Willcocks, C.G., 2022. Anoddpm: Anomaly detection with denoising diffusion probabilistic models using simplex noise. In: Proc. of the IEEE/CVF Conf. on Comp. Vis. and Pat. Recog. (CVPR) Workshops. http://dx.doi.org/10.1109/CVPRW56347.2022.00080.

Yang, J., Zhou, K., Li, Y., Liu, Z., 2021. Generalized out-of-distribution detection: A survey. http://dx.doi.org/10.48550/ARXIV.2110.11334, arXiv preprint arXiv:2110.11334.

Zhang, L.H., Goldstein, M., Ranganath, R., 2021. Understanding failures in out-of-distribution detection with deep generative models. In: Proc. Int. Conf. Mach. Learn. http://dx.doi.org/10.48550/ARXIV.2107.06908.

Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O., 2018. The unreasonable effectiveness of deep features as a perceptual metric. In: Proc. of the IEEE/CVF Conf. on Comp. Vis. and Pat. Recog. CVPR, http://dx.doi.org/10.48550/ARXIV.1801.03924.