

Zero-Shot Image Anomaly Detection Using Generative Foundation Models

Lemar Abdi, Amaan Valiuddin, Francisco Caetano, Christiaan Viviers, Fons van der Sommen
Eindhoven University of Technology
The Netherlands

l.abdi@tue.nl

Abstract

Detecting out-of-distribution (OOD) inputs is pivotal for deploying safe vision systems in open-world environments. We revisit diffusion models, not as generators, but as universal perceptual templates for OOD detection. This research explores the use of score-based generative models as foundational tools for semantic anomaly detection across unseen datasets. Specifically, we leverage the denoising trajectories of Denoising Diffusion Models (DDMs) as a rich source of texture and semantic information. By analyzing Stein score errors, amplified through the Structural Similarity Index Metric (SSIM), we introduce a novel method for identifying anomalous samples without requiring re-training on each target dataset. Our approach improves over state-of-the-art and relies on training a single model on one dataset — CelebA — which we find to be an effective base distribution, even outperforming more commonly used datasets like ImageNet in several settings. Experimental results show near-perfect performance on some benchmarks, with notable headroom on others, highlighting both the strength and future potential of generative foundation models in anomaly detection.

1. Introduction

Foundation models are large-scale general-purpose machine learning models trained on vast datasets to encapsulate broad, fundamental information applicable for multiple downstream tasks. One such downstream application is anomaly detection, which focuses on identifying data points that deviate from a preset in-distribution set (conventionally the training distribution). Often, the representational power of deep neural networks is utilized for this. Although traditionally framed as the identification of samples that deviate from the training distribution, anomaly detection has evolved with the rise of expressive models and large-scale datasets. This shift has moved the field away from task-specific feature engineering toward foundation models capable of detecting distributional shifts even between entirely

unseen datasets. As a result, foundation models can be used as dataset-agnostic detectors to prevent overconfident predictions across a wide range of downstream tasks.

Since anomaly detection is largely unsupervised, generative models are commonly employed for this task. In particular, latent-variable models are a popular choice due to their ability to capture rich representations of the data [11]. Recently, Denoising Diffusion Models (DDMs) have gained immense popularity due to their flexibility in general image generation. The strength of these architectures can be attributed to the finely discretized trajectory from latent space to data space, enabling the encapsulation of both a broad range of semantics as well as subtleties in the data [10, 15, 24]. As such, the merits of this model design have been recognized in the field of anomaly detection [8, 17, 18, 20, 30, 34]. However, much of the existing literature remains narrowly focused on specific datasets, highlighting the need for further research into more generalizable, foundation model-based approaches.

In this work, we leverage DDMs as a generative foundation model (GFM) for semantic anomaly detection by exploiting the statistics of the induced diffusion path. Specifically, we show that the data distributions can be distinguished by the error in the predicted Stein scores responsible for the diffusion trajectory. Additionally, we weigh the Stein scores by the Structural Similarity Index Measure (SSIM), emphasizing crucial, potentially anomalous areas that contribute to the score predictions. Our work can be considered an improvement to DiffPath [9], and achieves state-of-the-art performance on various natural anomaly detection benchmarks. In contrast to conventional semantic anomaly detection, the proposed technique does not require any retraining or fine-tuning on the training set, truly leveraging the power of generative foundation models. Our contributions can be summarized as follows.

- We propose DiffPathV2, a method that advances semantic anomaly detection using generative foundation models.
- We provide further insights on the Stein scores responsible for the induced diffusion trajectory.

- We achieve state-of-the-art performance on several natural image anomaly detection benchmarks.

The paper is structured as follows. We discuss related work in Section 2 and theoretical background in Section 3. We propose our new methodology in Section 4 followed by extensive experiments with ablations in Section 5. We finally conclude in Section 6.

2. Related Work

2.1. Unsupervised Image Anomaly Detection

Out-of-Distribution (OOD) Detection aims to identify inputs that deviate from the distribution p_{ID} of a given in-distribution (ID) dataset. Such problems are mostly unsupervised; no labels or examples of OOD data are available during training. This makes the task inherently challenging and highlights the importance of modeling the data distribution accurately. Existing OOD detection methods in vision can be broadly categorized into three methodological families:

Distance-based methods assume that OOD samples lie further from the ID learned embedding manifold. Hence, test samples are compared to in-distribution representations using metrics such as cosine similarity, Euclidean distance, or Mahalanobis distance. Recent work has improved robustness via ensemble representations or contrastive objectives [36], although they often require storing feature banks or class prototypes, which makes scalability a concern. Furthermore, methods are often excessively sensitive and dependent on the learned feature space and especially unreliable in near-OOD settings [27].

Density-based methods attempt to explicitly estimate the data log-likelihood using generative models. Approaches based on Normalizing Flows (NFs), and Energy-based models (EBMs) often fall in this category. Although intuitive and theoretically grounded, such methods often exhibit counterintuitive behavior, assigning higher likelihoods to OOD data [3, 12, 22]. To address these issues, recent work has proposed various alternatives. One direction involves using log-likelihood ratios (LLR), which contrast the model likelihood under different hypotheses or reference distributions to mitigate the bias toward OOD regions. Another line of research reframes the detection task from evaluating the likelihood to the information-theoretic notion of ‘typicality’. In this view, samples are considered OOD if they are ‘atypical’ with respect to the training distribution — even if their likelihood is high. Typicality can be quantified using distance-based approximations [22, 25] or via the Stein score of the log-likelihood of the data [1, 2, 33].

Reconstruction-based methods can be considered as the

implicit counterpart of density-based approaches. Instead of relying on likelihoods, generative models, such as a DDM or Variational Autoencoders (VAEs), are evaluated on the reconstruction error subject to an OOD input signal. The reconstruction error is often quantified using the pixel-level MSE error, SSIM, or LPIPS. Image-level metrics can often be insufficient to quantify semantic data shifts. In fact, anomalous inputs have been shown to be reconstructed with high fidelity [37], especially using natural imaging benchmarks [8].

2.2. Anomaly detection with DDMs

Denosing diffusion models have recently gained traction for OOD detection, largely due to their strong generative fidelity and flexible, class-agnostic training paradigms. Initial diffusion-based approaches largely adhered to the reconstruction-based framework, where class-conditional models are trained to regenerate inlier examples [20, 34]. These reconstructions serve as baselines for error computation, either in pixel space or through learned representations. Other lines of work exploit the score-based nature of diffusion models to derive likelihoods through the ODE probability flow [31]. Although this enables exact likelihood computation in theory, it inherits the limitations similar to those of other likelihood-based methods, where high likelihoods can be assigned to outlier data. At the same time, such models often include high memory requirements.

The popularity of DDMs has introduced a novel perspective on anomaly detection. These methods analyze the *forward* diffusion trajectory, that is, the noising process during inference, as an implicit representation of the underlying data manifold [9, 29]. The approach is based on the hypothesis that differing data marginals lead to significantly different intermediate noise trajectories across datasets. Among these methods, Heng *et al.* [9] introduced DIFFPATH, a diffusion-based method capable of zero-shot OOD detection across natural image benchmarks without retraining or architectural changes. Their work demonstrates that the forward dynamics of a pretrained unconditional diffusion model, even when trained on unrelated data, encode sufficient statistics to discriminate between samples from different data distributions.

Our work — DIFFPATHV2 — builds upon this by extending DiffPath, which further enhances discrimination performance across natural image benchmarks. These developments underscore a broader shift in the field: from task-specific anomaly detectors toward more general, reusable models that unify generative modeling and its downstream tasks with out-of-distribution detection in a single framework.

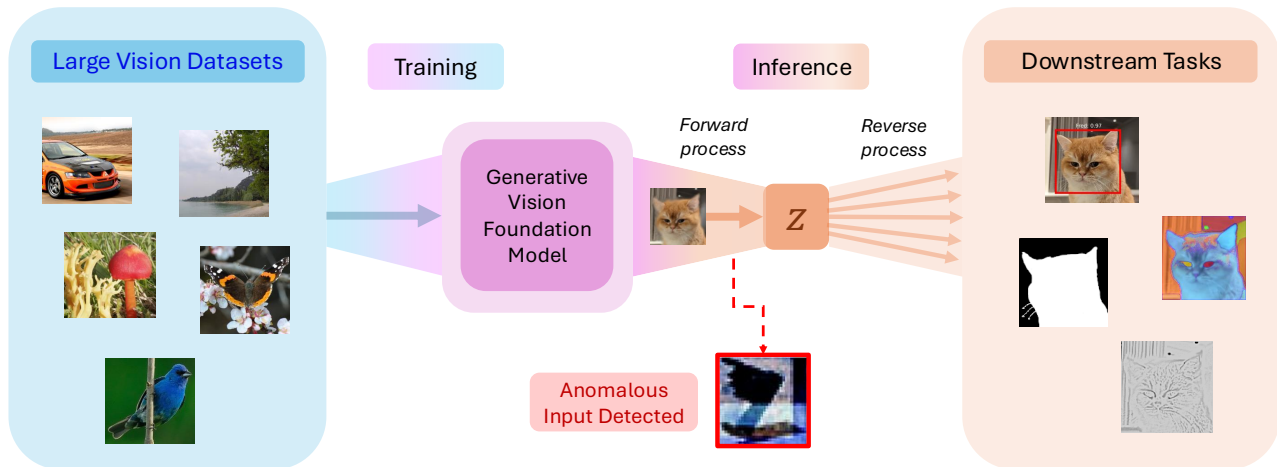


Figure 1. High-level overview of anomaly detection using generative foundation models

3. Background

3.1. Definitions

To avoid any ambiguity and maintain terminological consistency, we adopt the taxonomy introduced by Yang *et al.* [36], which frames *Generalized OOD Detection* as an umbrella task that includes three related subfields: *anomaly detection*, *open-set recognition (OSR)*, and *outlier detection*. Anomaly detection (AD) is defined as the task of detecting test-time samples that deviate from a predefined notion of normality characterized by the training data. In *sensory AD*, the aim is to detect covariate shifts. Covariate shift induces abnormalities in high-level image statistics. AD problems focus on objects with similar or identical semantics, where varying imaging conditions or surface-level defects are sought to be identified. In contrast, *semantic AD* aims to detect test samples that exhibit entirely novel semantics, such as previously unseen object classes, which are not represented in the inlier distribution.

Given the generalized OOD detection framework, this work operates under the *semantic anomaly detection* setting. Specifically, the unsupervised case is considered, where the model is trained only on unlabeled inlier data \mathbf{x} sampled from p_{ID} . The task is to detect whether a test sample \mathbf{x}^* originates from a semantically novel distribution p_{OOD} that does not share the label space with the training data. Throughout this paper, the terms *anomaly detection*, *semantic anomaly detection*, and *OOD detection* refer to this specific setting and are used interchangeably.

3.2. Score-based diffusion models

Score-based diffusion models gradually perturb the data \mathbf{x}_0 from the data distribution p_0 into a simple prior distribu-

tion p_T , typically a standard Gaussian, through a stochastic differential equation (SDE). The forward diffusion process $\{\mathbf{x}_t\}_{t=0}^T$ is governed by the SDE

$$d\mathbf{x} = \mathbf{f}(\mathbf{x}, t)dt + g(t)d\mathbf{w}, \quad (1)$$

where $\mathbf{f}(\mathbf{x}, t)$ denotes the drift term, $g(t)$ is a scalar diffusion coefficient, and \mathbf{w} is a standard Wiener process. Sampling from the learned data distribution involves reversing this diffusion process. The reverse-time SDE is given by

$$d\mathbf{x} = [\mathbf{f}(\mathbf{x}, t) - g(t)^2 \nabla_{\mathbf{x}} \log p_t(\mathbf{x})] dt + g(t)d\bar{\mathbf{w}}, \quad (2)$$

where $\nabla_{\mathbf{x}} \log p_t(\mathbf{x})$ is the *Stein score* (or score function), and $\bar{\mathbf{w}}$ denotes the reverse Brownian motion over $[T, 0]$. This stochastic process has been shown to have a deterministic counterpart, known as the probability flow ODE, which produces the same marginal distributions as the original SDE [31]. This formulation allows the diffusion process to be expressed as

$$d\mathbf{x} = \left[\mathbf{f}(\mathbf{x}, t) - \frac{1}{2}g(t)^2 \nabla_{\mathbf{x}} \log p_t(\mathbf{x}) \right] dt. \quad (3)$$

Both the reverse-time SDE and the probability flow ODE require an accurate estimate of the score function. This is typically achieved by training a neural network $\epsilon_{\theta}(\mathbf{x}, t)$ to approximate the Stein score function, using the score matching objective [32]

$$\min_{\theta} \mathbb{E}_t \left\{ \mathbb{E}_{\mathbf{x}_0} \mathbb{E}_{\mathbf{x}_t} \left[\|\epsilon_{\theta}(\mathbf{x}_t, t) - \epsilon\|_2^2 \right] \right\}, \quad (4)$$

where $t \sim \mathcal{U}[0, T]$, $\mathbf{x}_0 \sim p_0$, and $\mathbf{x}_t \sim p_{0t}(\mathbf{x}_t|\mathbf{x}_0)$. The denoising target can be inferred by considering a linear Gaussian process at arbitrary time $t \sim \mathcal{U}[0, 1]$.

$$\mathbf{x}_t = \sqrt{\alpha} \mathbf{x}_0 + \sigma_t \epsilon \quad (5)$$

with source noise $\epsilon \sim \mathcal{N}(0, 1)$ and mixing parameter $\bar{\alpha} = \prod_s^t \alpha_s$ and $\alpha_t = 1 - \beta_t$, with $\beta_t \in (0, 1)$ being the variance schedule. Hence, the variance scheduling directly correlates with $\sqrt{\bar{\alpha}}$ and therefore the signal strength at timestep t . We can write forward noising process as $p_{0t}(\mathbf{x}_t|\mathbf{x}_0) = \mathcal{N}(\sqrt{\bar{\alpha}}\mathbf{x}_0, \sigma_t^2\mathbf{I})$, take the derivative w.r.t. \mathbf{x}_t to obtain

$$\nabla_{\mathbf{x}_t} \log p(\mathbf{x}_t|\mathbf{x}_0) = -\frac{\mathbf{x}_t - \sqrt{\bar{\alpha}}\mathbf{x}_0}{\sigma_t}, \quad (6)$$

with $\sigma_t^2 = (1 - \bar{\alpha}_{t-1})/(1 - \bar{\alpha}_t)$, following the DDPM formulation [10] for the fixed time-dependent variances. Substitution of Equation 5 followed by multiplication with σ_t results in the ground-truth score

$$\epsilon = -\sigma_t \nabla_{\mathbf{x}_t} \log p_{0t}(\mathbf{x}_t|\mathbf{x}_0). \quad (7)$$

4. Methods

As mentioned in Section 1, our framework takes inspiration from DiffPath [9] and is built upon the assumption that different data distributions will exhibit distinctive denoising dynamics along the diffusion trajectory. The established theoretical framework of score-based DDMs in the preceding section enables us to precisely define the quantification of the trajectory dynamics.

4.1. DiffPathV2

We train a time-conditional encoder-decoder denoising model, $\epsilon_\theta(\mathbf{x}_t, t)$, on a large-scale data set to predict scores at each timestep. In previous work, DiffPath extracted these predicted scores to quantify statistics in the denoising trajectory. However, we argue that it is more principled to model the estimated score *error*, rather than just the scores themselves. If a test sample \mathbf{x}_0 is drawn from p_{OOD} , its trajectory under the noising process $q(\mathbf{x}_t|\mathbf{x}_0)$ will induce greater errors when estimating the denoising score using the pre-trained model. We propose quantifying these deviations not just at the trajectory level, but also by incorporating structural relevance via pixel-level weighting.

4.1.1. Quantifying trajectory dynamics

Let \mathbf{x}_t denote the noisy image at timestep t obtained from the forward diffusion process $q(\mathbf{x}_t|\mathbf{x}_0)$, and let $\epsilon(t)$ denote the true noise. The per-timestep pixelwise error is computed as

$$\text{mse}(\mathbf{x}_0, t) = \|\epsilon_\theta(\mathbf{x}_t, t) - \epsilon(t)\|_2^2. \quad (8)$$

Similarly to previous work, we employ a six-dimensional anomaly score. However, in our case, the calculations were based upon higher-order terms and the temporal behavior of the reconstruction. We briefly recap these concepts and argue their relevance for these settings.

Higher order terms: The first, second and third norm are extracted from the estimated Stein score errors. The different norms can provide more information on the raw magnitude, energy of the squared errors and the severity of outlier errors.

Temporal behavior: The errors only provide information about the current timestep. However, the score errors over time is also a crucial component that can provide a notion of the trajectories. Hence, the time derivative per timestep, together with the first, second and third norm are evaluated.

6-dimensional score: DiffPathV2 builds directly upon the original DiffPath [9], which computes a six-dimensional anomaly score from the predicted score ϵ_θ . Unlike the original formulation, which relies solely on ϵ_θ , instead we compute the 6D score based on the *mean squared error* (MSE) between the predicted and ground-truth noise ϵ , offering a more principled signal to assess the accuracy of the score. To capture both the error magnitudes and their temporal dynamics, we construct the following six-dimensional score

$$\mathbf{s}_{6\text{D}}(\mathbf{x}_0) = \begin{bmatrix} \sum_t \|\text{mse}(\mathbf{x}_0, t)\|_1 \\ \sum_t \|\text{mse}(\mathbf{x}_0, t)\|_2 \\ \sum_t \|\text{mse}(\mathbf{x}_0, t)\|_3 \\ \sum_t \|\partial_t \text{mse}(\mathbf{x}_0, t)\|_1 \\ \sum_t \|\partial_t \text{mse}(\mathbf{x}_0, t)\|_2 \\ \sum_t \|\partial_t \text{mse}(\mathbf{x}_0, t)\|_3 \end{bmatrix}. \quad (9)$$

The first three terms capture the aggregated p -norms of the prediction error across timesteps, while the remaining three quantify the curvature (i.e., rate of change) of the error along the diffusion path. This formulation provides a richer trajectory signature than previous approaches.

4.1.2. Incorporating structure

Although the proposed method improves the fidelity of error estimation, it treats all pixels equally when aggregating the MSE statistics. However, not all regions of the image carry equal semantic weight: fine-grained structures and high-frequency textures are often more indicative of semantic content, and are also harder to model during the denoising process. We address this by incorporating a spatial weighting mechanism based on the Structural Similarity Index Measure (SSIM). Specifically, we compute a pixel-wise SSIM map between the input image \mathbf{x}_0 and the accumulated predicted scores $\sum_t \epsilon_\theta(\mathbf{x}_t, t)$. The SSIM map acts as a structural relevance filter, or amplifier, localizing regions where the noise estimation might be poor. We then re-weight the 6D score components by modulating the MSE with the inverted SSIM map as

$$s(\mathbf{x}_0) = \mathbf{s}_{6\text{D}}(\mathbf{x}_0) \cdot (1 - \text{SSIM}(\mathbf{x}_0, \sum_t \epsilon_\theta(\mathbf{x}_t, t))). \quad (10)$$

Algorithm 1 Inference procedure of DIFFPATHV2

Require: Test samples $\mathcal{X}_{\text{test}}$, ID val set \mathcal{X}_{val} , trained DM g_θ

Ensure: OOD scores for test samples $S_\theta(\mathcal{X}_{\text{test}})$

```
1: function COMPUTESCORE( $\mathbf{x}_0, \epsilon, \epsilon_\theta$ )
2:    $\triangleright$  Inputs are assumed to be normalized
3:    $\text{MSE}_t \leftarrow \text{MSE}(\epsilon, \epsilon_\theta)$   $\triangleright$  shape  $(N, T, C, H, W)$ 
4:    $\text{SSIM} \leftarrow \text{SSIM}_{\text{MAP}}(\mathbf{x}_0, \sum_t \epsilon_\theta)$   $\triangleright$  shape  $(N, C, H, W)$ 
5:   for  $p \in \{1, 2, 3\}$  do
6:      $\mathbf{s}_p \leftarrow \sum_{c,h,w} (\sum_t \text{MSE}_t^p \cdot (1 - \text{SSIM}))$ 
7:      $\mathbf{s}_{p+3} \leftarrow \sum_{c,h,w} (\sum_t (\partial_t \text{MSE}_t^p) \cdot (1 - \text{SSIM}))$ 
8:   end for
9:   return  $\text{stack}([\mathbf{s}_1, \dots, \mathbf{s}_6])$   $\triangleright$  shape  $(N, 6)$ 
10: end function
11: for  $\mathbf{x}_0$  in  $\mathcal{X}_{\text{val}}$  do
12:    $\{\epsilon_\theta(\mathbf{x}_t, t), \epsilon(t)\} \leftarrow \text{DDIM}_{\text{Inversion}}(\mathbf{x}_0, g_\theta)$ 
13:    $\mathbf{s} \leftarrow \text{ComputeScore}(\mathbf{x}_0, \epsilon, \epsilon_\theta)$ 
14:   Append  $\mathbf{s}$  to  $L_{\text{val}}$ 
15: end for
16: Fit GMM:  $p_{\text{val}} \leftarrow \text{GMM}(L_{\text{val}})$ 
17: for  $\mathbf{x}_0$  in  $\mathcal{X}_{\text{test}}$  do
18:   Repeat lines 12–13 to compute  $\mathbf{s}$  and append to  $L_{\text{test}}$ 
19: end for
20: return  $p_{\text{val}}(L_{\text{test}})$ 
```

This yields a final anomaly score that emphasizes structurally relevant error regions, thus increasing sensitivity to subtle but semantically meaningful deviations. Importantly, this modulation occurs before global pooling into the 6D summary, ensuring that the influence of salient regions is preserved throughout the scoring process.

4.2. Benchmarks

To evaluate the anomaly detection capabilities of our proposed method, we adopt five publicly available image datasets: CIFAR-10 [14], CIFAR-100 [14], SVHN [23], CelebA [16], and Textures [4]. Each benchmark is constructed by designating one dataset as in-distribution (ID) and treating the others as sources of out-of-distribution (OOD) anomalies. Specifically, we consider CIFAR-10, SVHN, and CelebA as inlier datasets for their respective benchmarks. These combinations are chosen to represent both *near* and *far* semantic shifts. For instance, CIFAR-10 vs. CIFAR-100 represents a near-semantic shift within natural images, while SVHN vs. CelebA exemplifies a far-semantic shift (digits vs. faces). As the primary evaluation metric, we report the Area Under the Receiver Operating Characteristic curve (AUROC), which measures the ability to distinguish between ID and OOD samples across all thresholds.

We follow standard conventions established in previous work and briefly outline the procedure for completeness. We used the train-validation-test splits provided by each

dataset. The images are resized to 64×64 for experiments involving the ImageNet pre-trained model and to 32×32 for other models. During inference, anomaly scores are computed in the test set, while the validation set is used to fit a lightweight GMM, as described in Algorithm 1. The GMM enables assignment of log-likelihood scores to test samples to obtain the final anomaly score. The optimal parameters for the GMM are found through a grid search. All experiments are run on an NVIDIA H100 GPU.

4.3. Baselines

We compare our proposed method, DIFFPATHV2, against a diverse set of state-of-the-art (SOTA) approaches that span several modeling paradigms. Baselines can be categorized into two types: foundational models and non-foundational models, requiring dataset-specific training.

4.3.1. Non-foundational

We can further categorize the non-foundational approaches into three distinct baselines: energy-based, flow-based, diffusion-based, and foundation model-based methods.

Energy-based Models: This approach trains a model to assign low energy (or high likelihood) to in-distribution samples. *IGEBM* [6] and *VAEBM* [35] combine energy modeling with generative inference, while *Improved CD* [7] stabilizes contrastive divergence training to improve anomaly detection performance.

Flow-based Models: The methods in this group compute exact likelihoods using bijective invertible transformations, also known as Normalizing Flows [5, 28]. *IC* [30] calculates the input complexity. *DoS* [19] uses non-parametric density estimators to measure model statistics. *WAIC* [3] leverages model variance across ensemble samples to penalize unreliable likelihood estimates. *TT* [21] takes the information-theoretic perspective, arguing that the entropy of the samples compared to the generating distribution (i.e., the ‘typicality’) can indicate whether a sample is anomalous. Finally, *LR* [26] is a log-likelihood ratio baseline.

Diffusion-based Models: Such methods adapt score-based or DDPMs. *NLL* computes exact log-likelihoods using probability flow ODEs [31]. *IC (DM)* [30] applies input complexity scores to diffusion models. *MSMA* [18] utilizes multi-scale score matching. *DDPM-OOD* [8] uses a conditional reconstruction-based approach. *LMD* [17] introduces a masking and inpainting strategy for an improved distance-based method.

4.3.2. Foundational

These methods leverage pre-trained generative models without task-specific retraining. *ITD* [13] computes the ex-

act NLL in terms of an optimal denoiser as a function of the signal-to-noise ratio. *DiffPath* [9] is the baseline method discussed in the theoretical background of this paper. To the best of our knowledge, DiffPath is state-of-the-art on foundational anomaly detection at the time of writing this manuscript.

All baselines are evaluated according to a consistent protocol across benchmarks. For fairness, we reimplement ITD using the same model architecture and pre-processing as DiffPath-based methods.

5. Experiments

5.1. Main results

The results of our experiments on each dataset are presented in Table 1. The quantitative scores evaluate the distinctive power of each model between several datasets. Each value indicates the AUROC and bold numbers indicate the best score per dataset combination. At first glance, it is clear that on average, over all combinations, our proposed improvement methodology performs best. Furthermore, compared to DiffPath, DIFFPATHV2 achieves notably higher performance in the near-OOD setting, specifically when distinguishing CIFAR-10 inliers from CIFAR-100 anomalies, where semantic overlap between classes makes detection particularly challenging. Interestingly enough, the runner-up model MSMA, indicated with an underscore, is also a score-based diffusion model. Although MSMA is not foundational, the results still confirm the fact that Stein scores are an appropriate source of information about the datasets. Overall, our results confirm the hypothesis that the trajectory of the diffusion process contains sufficient information to near-perfectly distinguish datasets. Moreover, the proposed method shows that this paradigm translates to even unseen datasets.

5.2. Ablation on Stein scores

To what extent do Stein *errors* and SSIM influence the OOD performance? We conduct an ablation study comparing four variants: the original DIFFPATH baseline using estimated scores, the same baseline modulated by $(1 - \text{SSIM})$, MSE-based error signals between predicted and true noise, and finally our full method combining error signals with SSIM weighting. Results across all benchmarks are shown in Table 2, and the corresponding histograms of anomaly scores are visualized in Figure 2 for the near-OOD setting of CIFAR-10 vs. CIFAR-100.

The baseline without error computation or SSIM already performs competitively on most benchmarks, except under near-semantic shifts as shown in Figure 2a. However, directly incorporating score errors — i.e., using the mean squared error between ϵ_θ and ϵ — generally improves performance, particularly on datasets such as CelebA and

SVHN, where semantic anomalies are further from the inlier distribution. Interestingly, modulating the original baseline with the inverse SSIM ($\text{Baseline} \cdot (1 - \text{SSIM})$) degrades performance, suggesting that SSIM alone does not suffice as a spatial discriminator without complementary error signals. Moreover, it shifts both inlier and OOD distributions toward higher anomaly scores, as seen in Figure 2b. In contrast, combining the Stein error with SSIM ($\text{Error} \cdot (1 - \text{SSIM})$) variant sharply shifts only the OOD distribution to the right while preserving the inlier distribution, resulting in a clear distinction between the two datasets. This method yields the strongest results overall, with an average AUROC of 94.9. These findings confirm the hypothesis that SSIM helps localize regions with meaningful semantic deviations during the denoising process, thereby amplifying anomalies that would otherwise be diluted in global averaging.

Taken together, these results validate that (1) computing the error between predicted and ground-truth noise estimates provides a more informative signal than raw scores, and (2) that spatial modulation via SSIM further enhances discriminative power by highlighting perceptually significant regions.

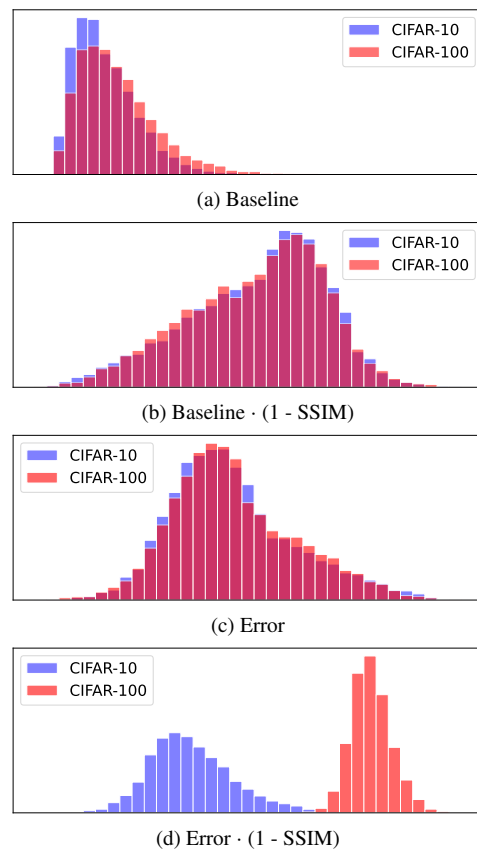


Figure 2. Qualitative comparison of Stein score errors and SSIM on Near-OOD (CIFAR10 vs. CIFAR100) performance.

Table 1. Comparison between SOTA methods in semantic AD benchmarks. The Generative Foundational Models do not require the model to retrain on each inlier dataset, contrary to the other methods listed. (*) denotes re-implementation.

Method	C10 vs.				SVHN vs.				CelebA vs.				Average
	SVHN	CelebA	C100	Textures	C10	CelebA	C100	Textures	C10	SVHN	C100	Textures	
<i>Energy-based</i>													
IGEBM	63.0	70.0	50.0	48.0	-	-	-	-	-	-	-	-	-
VAEBM	83.0	77.0	62.0	-	-	-	-	-	-	-	-	-	-
Improved CD	91.0	-	83.0	88.0	-	-	-	-	-	-	-	-	-
<i>Flow-based</i>													
IC	95.0	86.3	73.6	-	-	-	-	-	-	-	-	-	-
DoS	95.5	<u>99.5</u>	57.1	-	96.2	100	96.5	-	94.9	99.7	95.6	-	92.8
WAIC	14.3	92.8	53.2	-	80.2	99.1	83.1	-	50.7	13.9	53.5	-	60.1
TT	87.0	84.8	54.8	-	97.0	100	96.5	-	63.4	98.2	67.1	-	83.2
LR	06.4	91.4	52.0	-	81.9	91.2	77.9	-	32.3	02.8	35.7	-	52.4
<i>Diffusion-based</i>													
NLL	09.1	57.4	52.1	60.9	<u>99.0</u>	<u>99.9</u>	<u>99.2</u>	98.3	81.4	10.5	78.6	80.9	68.9
IC	92.1	51.6	51.9	55.3	08.0	02.8	10.0	17.4	48.5	97.2	51.0	55.9	45.1
MSMA	95.7	100	61.5	<u>98.6</u>	97.6	99.5	98.0	<u>99.6</u>	91.0	99.6	92.7	99.9	<u>94.5</u>
DDPM-OOD	39.0	65.9	53.6	59.8	95.1	98.6	94.5	91.0	79.5	63.6	77.8	77.3	74.6
LMD	99.2	55.7	60.4	66.7	91.9	89.0	88.1	91.4	<u>98.9</u>	100	<u>97.9</u>	<u>97.2</u>	86.5
<i>Generative Foundation Models</i>													
ITD*	<u>99.0</u>	100	100	99.0	1.2	76.8	57.9	58.6	91.0	62.4	52.4	58.1	76.2
DiffPath	91.0	89.7	59.0	92.3	93.9	97.9	95.3	98.1	99.8	100	99.8	99.9	93.1
DiffPathV2 (Ours)	94.9	<u>99.5</u>	<u>99.4</u>	90.5	100	100	100	100	91.1	100	79.8	80.2	94.9

Table 2. Effect of the Stein score errors and SSIM on AD performance.

Method	C10 vs.				SVHN vs.				CelebA vs.				Average
	SVHN	CelebA	C100	Textures	C10	CelebA	C100	Textures	C10	SVHN	C100	Textures	
Baseline	91.0	89.7	<u>59.0</u>	92.3	93.9	97.9	95.3	98.1	<u>99.8</u>	100	<u>99.8</u>	99.9	<u>93.1</u>
Baseline · (1-SSIM)	92.1	77.0	58.9	53.7	<u>95.7</u>	<u>99.2</u>	<u>95.8</u>	91.4	64.4	99.8	68.5	63.6	80.0
Error	93.4	<u>96.2</u>	52.7	<u>91.0</u>	94.4	95.8	94.6	<u>98.8</u>	99.9	100	99.9	99.9	93.0
Error · (1-SSIM)	94.9	99.5	99.4	90.5	100	100	100	100	91.1	100	79.8	80.2	94.9

5.3. Ablation on Foundational dataset

Which dataset serves as the best foundational dataset?
To assess how the base distribution influences anomaly

detection performance, we ablate the effect of the dataset used to pre-train the GFM. Specifically, we compare ImageNet and CelebA as base distributions across multiple

Table 3. Comparison between Generative Foundation Models and the effect of the training dataset.

Model / $p_{\text{train}}(\mathbf{x})$	C10 vs.				SVHN vs.				CelebA vs.				Average
	SVHN	CelebA	C100	Textures	C10	CelebA	C100	Textures	C10	SVHN	C100	Textures	
ITD / ImageNet	99.0	100	100	99.0	1.2	76.8	57.9	58.6	91.0	62.4	52.4	58.1	76.2
ITD / CelebA	1.2	0.0	0.0	0.6	<u>98.9</u>	20.1	27.6	45.4	100.0	79.9	59.7	77.3	42.6
DiffPath / ImageNet	85.6	50.2	58.0	84.1	94.3	96.4	<u>95.4</u>	96.9	80.7	98.1	84.3	<u>96.4</u>	85.0
DiffPath / CelebA	91.0	89.7	59.0	<u>92.3</u>	93.9	<u>97.9</u>	95.3	98.1	<u>99.8</u>	100	99.8	<u>99.9</u>	<u>93.1</u>
DiffPathV2 / ImageNet	08.6	63.8	00.1	96.0	100	100	100	100	74.4	100	68.9	100	76.0
DiffPathV2 / CelebA	94.9	<u>99.5</u>	<u>99.4</u>	90.5	100	100	100	100	91.1	100	79.8	80.2	94.9

GFM. All datasets are resized to a consistent resolution, the higher resolution datasets are downscaled to 32×32 , before upscaling all datasets to match the 64×64 ImageNet model using bilinear pixel interpolation.

Heterogeneity is not always better: The field often enjoys prototypical large-scale and diverse datasets such as ImageNet, which offers broad generalization. However, we find that training on CelebA consistently yields better average anomaly detection scores for Diffpath(V2), as reported in Table 3. Our findings suggest that increasing the heterogeneity of the foundational distribution does not always lead to better performance, particularly when the anomaly score is tied to fine-grained deviations along the generative trajectory. These findings validate previous experiments in the literature [9], which evidently translate to the proposed method. CelebA may offer a form of structural consistency that improves the sensitivity of the model to perturbations in the diffusion process, despite its narrower semantic scope. Our method explicitly measures deviations in the trajectory curvature and amplifies this via the SSIM. Therefore, such semantic coherence appears to reinforce the scoring mechanism. This indicates that certain datasets, although semantically narrower, might serve as a better base distribution.

Foundational datasets are objective-dependent: In contrast, the Information-Theoretic Diffusion (ITD) model performs best with the ImageNet base and shows limited generalization when trained on CelebA. This is likely due to its direct optimization of the exact log-likelihood $\log p(\mathbf{x})$ as a function of the signal-to-noise ratio (SNR), which is tightly coupled to the distributional complexity of the training domain. While this formulation achieves SOTA results for density estimation [13], it may be less robust to anomaly detection across multiple benchmarks.

These findings highlight that the choice of pretraining data not only defines the generative capacity of the model

but also determines how well samples from different distributions can be detected in its diffusion trajectory. Rather than simply suggesting that narrower datasets are better, we emphasize that there is a complex interplay between the diversity of the base distribution and the nature of the anomaly detection objective.

6. Conclusion

In this work, we demonstrate that score-based models can serve as foundational generative models for semantic anomaly detection across datasets never seen during training. Specifically, we show that the denoising trajectory of Denoising Diffusion Models (DDMs) encodes rich information about both texture and semantics. We extract and analyze statistical signals from Stein score errors, amplified using the Structural Similarity Index Metric (SSIM), and leverage these to detect anomalous samples and improve over the state-of-the-art. The implemented methods are truly foundational, as they are trained on solely a single dataset and only utilizes an in-distribution validation set for determining an appropriate error threshold. Contrary to the common reliance on ImageNet, we find that CelebA provides a surprisingly effective base distribution for a variety of anomaly detection baselines. This work underscores the feasibility and promise of using generative foundation models for anomaly detection. While near-perfect scores are achieved on some datasets, significant room for improvement remains on others. We encourage further exploration of anomaly detection using (larger) generative foundation models, particularly by leveraging the denoising trajectories of score-based models.

References

- [1] Lemar Abdi, MM Amaan Valiuddin, Christiaan GA Viviers, Peter HN de With, and Fons van der Sommen. Typicality excels likelihood for unsupervised out-of-distribution detection in medical imaging. In *Uncertainty for Safe Utilization of Machine Learning in Medical Imaging: 6th International Workshop, Held in Conjunction with MICCAI 2024, Proceedings*, page 149. Springer, 2025. 2
- [2] Samy Chali, Inna Kucher, Marc Duranton, and Jacques-Olivier Klein. Improving normalizing flows with the approximate mass for out-of-distribution detection. In *Proceedings of the IEEE/CVF Conference on CVPR*, pages 750–758, 2023. 2
- [3] Hyunsun Choi, Eric Jang, and Alexander A. Alemi. WAIC, but Why? Generative Ensembles for Robust Anomaly Detection, 2019. arXiv:1810.01392 [cs, stat]. 2, 5
- [4] M. Cimpoi, S. Maji, I. Kokkinos, S. Mohamed, , and A. Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2014. 5
- [5] Laurent Dinh, David Krueger, and Yoshua Bengio. Nice: Non-linear independent components estimation. *arXiv preprint arXiv:1410.8516*, 2014. 5
- [6] Yilun Du and Igor Mordatch. Implicit generation and modeling with energy based models. *Advances in neural information processing systems*, 32, 2019. 5
- [7] Yilun Du, Shuang Li, Joshua Tenenbaum, and Igor Mordatch. Improved contrastive divergence training of energy-based models. In *International Conference on Machine Learning*, pages 2837–2848. PMLR, 2021. 5
- [8] Mark S Graham, Walter HL Pinaya, Petru-Daniel Tudosiu, Parashkev Nachev, Sebastien Ourselin, and Jorge Cardoso. Denoising diffusion models for out-of-distribution detection. In *Proceedings of the IEEE/CVF CVPR*, pages 2948–2957, 2023. 1, 2, 5
- [9] Alvin Heng, Harold Soh, et al. Out-of-distribution detection with a single unconditional diffusion model. *Advances in NeurIPS*, 37:43952–43974, 2024. 1, 2, 4, 6, 8
- [10] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in NeurIPS*, 33:6840–6851, 2020. 1, 4
- [11] Diederik P Kingma and Max Welling. Auto-encoding variational bayes, 2022. 1
- [12] Polina Kirichenko, Pavel Izmailov, and Andrew G Wilson. Why Normalizing Flows Fail to Detect Out-of-Distribution Data. In *NIPS*, pages 20578–20589. Curran Associates, Inc., 2020. 2
- [13] Xianghao Kong, Rob Brekelmans, and Greg Ver Steeg. Information-theoretic diffusion. In *The Eleventh International Conference on Learning Representations*, 2023. 5, 8
- [14] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 5
- [15] Luping Liu, Yi Ren, Zhijie Lin, and Zhou Zhao. Pseudo numerical methods for diffusion models on manifolds. In *International Conference on Learning Representations*, 2022. 1
- [16] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, 2015. 5
- [17] Zhenzhen Liu, Jin Peng Zhou, Yufan Wang, and Kilian Q Weinberger. Unsupervised out-of-distribution detection with diffusion inpainting. In *International Conference on Machine Learning*, pages 22528–22538. PMLR, 2023. 1, 5
- [18] Ahsan Mahmood, Junier Oliva, and Martin Andreas Styner. Multiscale score matching for out-of-distribution detection. In *International Conference on Learning Representations*, 2021. 1, 5
- [19] Warren Morningstar, Cusuh Ham, Andrew Gallagher, Balaji Lakshminarayanan, Alex Alemi, and Joshua Dillon. Density of states estimation for out of distribution detection. In *International Conference on Artificial Intelligence and Statistics*, pages 3232–3240. PMLR, 2021. 5
- [20] Arian Mousakhan, Thomas Brox, and Jawad Tayyub. Anomaly detection with conditioned denoising diffusion models. In *DAGM German Conference on Pattern Recognition*, pages 181–195. Springer, 2024. 1, 2
- [21] Eric Nalisnick, Akihiro Matsukawa, Yee Whye Teh, and Balaji Lakshminarayanan. Detecting out-of-distribution inputs to deep generative models using typicality. *arXiv preprint arXiv:1906.02994*, 2019. 5
- [22] Eric T. Nalisnick, Akihiro Matsukawa, Yee Whye Teh, Dilan Görür, and Balaji Lakshminarayanan. Do deep generative models know what they don’t know? In *ICLR 2019*, 2019. 2
- [23] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bisacco, Baolin Wu, Andrew Y Ng, et al. Reading digits in natural images with unsupervised feature learning. In *NIPS workshop on deep learning and unsupervised feature learning*, page 4. Granada, 2011. 5
- [24] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International conference on machine learning*, pages 8162–8171. PMLR, 2021. 1
- [25] G. Osada, T. Takahashi, and T. Nishide. Understanding likelihood of normalizing flow and image complexity through the lens of out-of-distribution detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(19):21492–21500, 2024. 2
- [26] Jie Ren, Peter J Liu, Emily Fertig, Jasper Snoek, Ryan Poplin, Mark Depristo, Joshua Dillon, and Balaji Lakshminarayanan. Likelihood ratios for out-of-distribution detection. *Advances in neural information processing systems*, 32, 2019. 5
- [27] Jie Ren, Stanislav Fort, Jeremiah Liu, Abhijit Guha Roy, Shreyas Padhy, and Balaji Lakshminarayanan. A simple fix to mahalanobis distance for improving near-ood detection. *arXiv preprint arXiv:2106.09022*, 2021. 2
- [28] Danilo Rezende and Shakir Mohamed. Variational inference with normalizing flows. In *International conference on machine learning*, pages 1530–1538. PMLR, 2015. 5
- [29] Shunsuke Sakai and Tatsuhito Hasegawa. Reconstruction-free anomaly detection with diffusion models via direct la-

- tent likelihood evaluation. *arXiv preprint arXiv:2504.05662*, 2025. [2](#)
- [30] Joan Serrà, David Álvarez, Vicenç Gómez, Olga Slizovskaia, José F. Núñez, and Jordi Luque. Input complexity and out-of-distribution detection with likelihood-based generative models. In *International Conference on Learning Representations*, 2020. [1](#), [5](#)
- [31] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *ICLR*, 2021. [2](#), [3](#), [5](#)
- [32] Pascal Vincent. A connection between score matching and denoising autoencoders. *Neural computation*, 23(7):1661–1674, 2011. [3](#)
- [33] Christiaan Viviers, Amaan Valiuddin, Francisco Caetano, Lemar Abdi, Lena Filatova, Fons van der Sommen, et al. Can your generative model detect out-of-distribution covariate shift? In *ECCV*, pages 184–201. Springer, 2025. [2](#)
- [34] Julia Wolleb, Florentin Bieder, Robin Sandkühler, and Philippe C Cattin. Diffusion models for medical anomaly detection. In *MICCAI*, pages 35–45. Springer, 2022. [1](#), [2](#)
- [35] Zhisheng Xiao, Karsten Kreis, Jan Kautz, and Arash Vahdat. {VAEBM}: A symbiosis between variational autoencoders and energy-based models. In *International Conference on Learning Representations*, 2021. [5](#)
- [36] Jingkang Yang, Kaiyang Zhou, Yixuan Li, and Ziwei Liu. Generalized out-of-distribution detection: A survey. *International Journal of Computer Vision*, 132(12):5635–5662, 2024. [2](#), [3](#)
- [37] Yibo Zhou. Rethinking reconstruction autoencoder-based out-of-distribution detection. In *Proceedings of the IEEE/CVF CVPR*, pages 7379–7387, 2022. [2](#)