Unsupervised Out-of-Distribution Detection with Diffusion Inpainting

Zhenzhen Liu [*1] **Jin Peng Zhou** [*1] **Yufan Wang** [1] **Kilian Q. Weinberger** [1]

Abstract

Unsupervised out-of-distribution detection (OOD) seeks to identify out-of-domain data by learning only from unlabeled in-domain data. We present a novel approach for this task – Lift, Map, Detect (LMD) – that leverages recent advancement in diffusion models. Diffusion models are one type of generative models. At their core, they learn an iterative denoising process that gradually maps a noisy image closer to their training manifolds. LMD leverages this intuition for OOD detection. Specifically, LMD lifts an image off its original manifold by corrupting it, and maps it towards the in-domain manifold with a diffusion model. For an out-of-domain image, the mapped image would have a large distance away from its original manifold, and LMD would identify it as OOD accordingly. We show through extensive experiments that LMD achieves competitive performance across a broad variety of datasets.

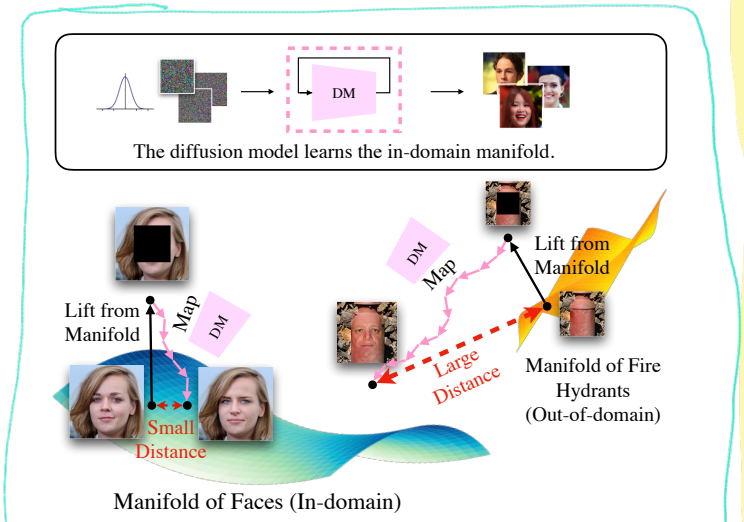


Figure 1. The pictorial intuition behind LMD for OOD detection. A diffusion model learns a mapping to the in-domain manifold. LMD lifts an image off its manifold by masking, and uses the diffusion model to move it towards the in-domain manifold. An in-domain image would have a much smaller distance between the original and mapped locations than its out-of-domain counterparts.

1. Introduction

Out-of-distribution (OOD) detection seeks to classify whether a data point belongs to a particular domain. It is especially important, because machine learning models typically assume that test-time samples are drawn from the same distribution as the training data. If the test data do not follow the training distribution, they can inadvertently produce non-sensical results. The increased use of machine learning models in high-stake areas, such as medicine (Hamet & Tremblay, 2017) and criminal justice (Rigano, 2019), amplifies the importance of OOD detection. For example, if a doctor mistakenly inputs a chest X-ray into a brain tumor detector, the model would likely still return a prediction – which would be meaningless and possibly misleading.

Previous researches have studied OOD detection under different settings: supervised and unsupervised. Within the supervised setup, the supervision can originate from different sources. In the most informed setting, one assumes access to representative out-of-domain samples. These allow one to train an OOD detector as a classifier distinguishing in-domain from out-of-domain data, and achieves high performance (Hendrycks et al., 2018; Ruff et al., 2019) – as long as the out-of-domain data do not deviate from the assumed out-of-domain distribution. In many practical applications, however, such knowledge is unattainable. In fact, out-of-domain data can be highly diverse and unpredictable. A significantly more relaxed assumption is to only require access to an in-domain classifier or class labels. Under this setting, methods such as Hendrycks & Gimpel (2016); Liang et al. (2017); Lee et al. (2018); Huang et al. (2021); Wang et al. (2022) have achieved competitive performance. Although less informed, this setting relies on two implicit assumptions: the in-domain data have well-defined classes, and there are sufficiently plenty data with class annotations. In practice, these assumptions often cannot be met. Unlabeled data do not require the expensive human annotation, and thus are often readily available in large quantity. Ideally, one would like to build an OOD detector that only requires unlabeled in-domain data during training.

[*]Equal contribution [1]Cornell University. Correspondence to: Zhenzhen Liu <zl535@cornell.edu>, Jin Peng Zhou <jz563@cornell.edu>.

Recently, a class of generative models – the diffusion models (DM) (Ho et al., 2020; Song et al., 2020) – have gained increasing popularity. DMs formulate two processes: The forward process converts an image to a sample drawn from a noise distribution by iteratively adding noise to its pixels; the backward process maps a noise image towards a specific image manifold by iteratively removing noise from the image. A dedicated neural network is trained to perform the denoising steps in the backward process.

In this paper, we argue that we can leverage the property that the diffusion model learns a mapping to a manifold, and turn it into a strong unsupervised OOD detector. Intuitively, if we lift an image from its manifold, then the lifted image can be mapped back to its original vicinity with a diffusion model trained over the same manifold. If instead the diffusion model is trained over a different manifold, it would attempt to map the lifted image towards its own training manifold, causing a large distance between the original and mapped images. Thus, we can detect out-of-domain images based on such distance.

To this end, we propose a novel unsupervised OOD detection approach called **L**ift, **M**ap, **D**etect (LMD) that captures the above intuition. We can **lift** an image from its original manifold by corrupting it. For example, a face image masked in the center clearly does not belong to the face manifold anymore. As shown by Song et al. (2020); Lugmayr et al. (2022), the diffusion model can impute the missing regions of an image with visually plausible content, a process commonly referred as inpainting, without retraining. Thus, we can **map** the lifted image by inpainting with a diffusion model trained over the in-domain data. We can then use a standard image similarity metric to measure the distance between the original and mapped images, and **detect** an out-of-domain image when we observe a large distance. Figure 1 illustrates an example: A diffusion model trained on face images maps a lifted in-domain face image closer to the original image than an out-of-domain fire hydrant counterpart.

To summarize our contributions:

1. We propose a novel approach LMD for unsupervised OOD detection, which directly leverages the diffusion model's manifold mapping ability without retraining. We introduce design choices that improve the separability between in-domain and out-of-domain data.

2. We show that LMD is versatile through experiments on datasets with different coloring, variability and resolution.

3. We provide qualitative visualization and quantitative ablation results that verify the basis of our approach and our design choices.

2. Background

Unsupervised OOD Detection. We formalize the unsupervised OOD detection task as follows: Given a distribution of interest $\mathcal{D}$, one would like to build a detector that decides whether a data point $\mathbf{x}$ is drawn from $\mathcal{D}$. The detector is only built upon unlabeled in-distribution samples $\mathbf{x_1}, \cdots, \mathbf{x_n} \sim \mathcal{D}$. Given a test data point $\mathbf{x}$, the detector outputs an OOD score $s(\mathbf{x})$, where a higher $s(\mathbf{x})$ signifies that $\mathbf{x}$ is more likely *not* to be sampled from $\mathcal{D}$.

Existing works can be roughly divided into three categories: likelihood-based, reconstruction-based, and feature-based. Likelihood-based approaches date back to Bishop (1994). At a high level, one fits the in-domain distribution with a model, and evaluates the likelihood of the test data under the model. Recent approaches often employ a deep generative model that supports likelihood computation, such as PixelCNN++ (Salimans et al., 2017) or Glow (Kingma & Dhariwal, 2018). However, several works (Choi et al., 2018; Nalisnick et al., 2018; Kirichenko et al., 2020) have found that generative models sometimes assign higher likelihood to out-of-domain data.

This issue can be alleviated in various ways. One line of work adopts a likelihood ratio approach: Ren et al. (2019) trains a semantic model and a background model, and takes the ratio of the likelihoods from the two models. Serrà et al. (2019) observes a negative correlation between an image's complexity and its likelihood, and adjusts the likelihood by the compression size. Xiao et al. (2020) optimizes the model configuration to maximize a test image's likelihood, and measures the amount of likelihood improvement. Another line of work adopts a typicality test approach (Nalisnick et al., 2019; Morningstar et al., 2021; Bergamin et al., 2022). They examine the distribution of in-domain likelihood or other model statistics, and evaluate the typicality of the test data model statistics through hypothesis testing or density estimation. Lastly, several works (Maaløe et al., 2019; Kirichenko et al., 2020) seek to improve the design choices of generative models.

Reconstruction-based approaches evaluate how well a data point can be reconstructed by a model learned over the in-domain data. Our approach LMD falls into this category. Within this line of work, Sakurada & Yairi (2014); Xia et al. (2015); Zhou & Paffenroth (2017); Zong et al. (2018) encode and decode data with autoencoders. Schlegl et al. (2017); Li et al. (2018) perform GAN (Goodfellow et al., 2014) inversion for a data point, and evaluate its reconstruction error and discriminator confidence under the inverted latent variable. Additionally, concurrent to our work, Graham et al. (2022) leverages diffusion models to reconstruct images at varied diffusion steps, while we mask and inpaint an image repeatedly with fixed steps. The two approaches are complementary to each other.

Feature-based approaches featurize data in an unsupervised manner, and fit a simple OOD detector like a Gaussian Mixture Model over the in-domain features. Denouden et al. (2018) leverages the latent variables of an autoencoder, and evaluates the Mahalanobis distance in the latent space along with the data reconstruction error. Ahmadian & Lindsten (2021) extracts low-level features from the encoder of an invertible generative model. Hendrycks et al. (2019); Bergman & Hoshen (2020); Tack et al. (2020); Sehwag et al. (2021) learn a representation over the in-domain data through self-supervised training; Xiao et al. (2021) further shows that one can instead use a strong pretrained feature extractor while maintaining comparable performance.

Diffusion Models. In this section, we provide a brief overview of the diffusion models (DM). It is a type of generative model that learns the distribution of its training data. DM formulates a forward process of corrupting data by adding noise to them, commonly referred as diffusion. It learns the reverse process of gradually producing a less noisy sample, commonly referred as denoising. One classic formulation of DM is called Denoising Diffusion Probabilistic Models (DDPMs) (Sohl-Dickstein et al., 2015; Ho et al., 2020). Specifically, starting from a data sample x_0, each step $t = 1, 2, \cdots, T$ of the diffusion process injects Gaussian noise given by

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

where β_t follows a fixed variance schedule. The DDPM with a prior distribution $x_T \sim \mathcal{N}(0, 1)$ learn the denoising process given by

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t)) \tag{2}$$

where both $\mu_\theta(x_t, t)$ and $\Sigma_\theta(x_t, t)$ are learned by a neural network parametrized by θ. Note that other formulations of DMs, such as score-based generative models (Song & Ermon, 2019) and stochastic differential equations (Song et al., 2020), also support diffusion and denoising processes. Since LMD is agnostic to different formulations of DMs, we refer the reader to Yang et al. (2022) for a more detailed mathematical description of the other formulations.

3. Lift, Map, Detect

The intuition behind our algorithm, Lift, Map, Detect (LMD), is illustrated in Figure 1. In a nutshell, we employ a diffusion model learned over the in-domain data, which provides a mapping towards the underlying in-domain image manifold. To test whether an image is in-domain or out-of-domain, we lift the image off its original manifold through corruption, and map the lifted image to the in-domain manifold with the trained DM. If the original image is in-domain, it is mapped back to its manifold, near its original location. If it is out-of-domain, the image is mapped to a different manifold, likely leaving a large distance between the original image and the mapped image. Figure 2 shows the high-level workflow of LMD. Algorithm 2 summarizes the key steps of LMD in pseudocode.

Lifting Images. To lift an image off its manifold, we need to corrupt the image so that it no longer appears to be from its original manifold. Concretely, we apply a mask to the image so that part of it is completely removed. Since various mask patterns and sizes can be used, masking provides a direct and flexible way to lift the image from the manifold. For example, it is intuitive to see that the larger the mask is, the further away the image is lifted from the manifold.

Mapping the Lifted Images. Since diffusion models (DM) can perform inpainting without retraining (Song et al., 2020; Lugmayr et al., 2022) (see Algorithm 1), we naturally employ a DM and use inpainting to map the lifted images. Specifically, we employ a DM parametrized by θ_{in} that is trained on the in-domain data. This DM can model the in-domain distribution well enough to map a lifted in-domain image back to its original vicinity. Meanwhile, the DM should have almost no knowledge about the out-of-domain manifold. Thus, it naturally maps a lifted out-of-domain image towards the DM's training manifold, which is the in-domain manifold. This phenomenon leads to a larger distance between the original and mapped images for out-of-domain images than the in-domain ones. For ease of reference, we also refer these mapped images as reconstructed images or simply reconstructions.

Algorithm 1 Inpaint

Input: original image x_{orig}, binary mask M where 0 indicates region to be inpainted, diffusion model θ
Output: inpainted image x_{inp}
for $i = T$ **to** 1 **do**
  if $i == T$ **then**
    $x_{inp} \leftarrow$ sample from noise distribution
  end if
  $x'_{orig} \leftarrow \texttt{diffuse}(x_{orig};\ \theta)$ to step $i-1$
  $x_{inp} \leftarrow \texttt{denoise}(x_{inp};\ \theta)$ to step $i-1$
  $x_{inp} \leftarrow x'_{orig} \cdot M + x_{inp} \cdot (1-M)$
end for
return x_{inp}

Reconstruction Distance Metric. We adopt the Learned Perceptual Image Patch Similarity (LPIPS) (Zhang et al., 2018) metric, a standard and strong metric that captures the perceptual difference between images. Since LPIPS assigns higher values to more dissimilar images, we compute the LPIPS between original and reconstructed images, and use it directly as the OOD score. We perform detailed ablation on different reconstruction distance metrics in Section 4.4.

It is worth noting that mapping lifted images is the most cru-

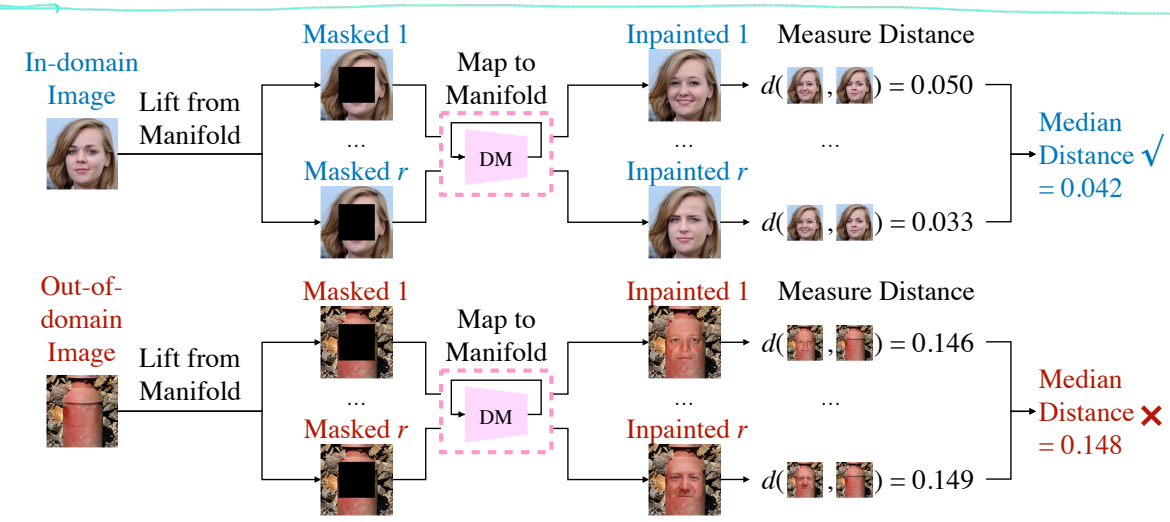


Figure 2. High-level workflow of LMD. LMD employs a diffusion model learned over the in-domain manifold. It first repeatedly lifts an image from its manifold by masking it, and maps it towards the diffusion model's training manifold by inpainting. Then, it inspects the median distance between the original image and each mapped image to detect out-of-domain images. As out-of-domain images cannot be mapped back to their own manifolds, they have larger distances.

cial component. The hypothesis – in-domain reconstructions are closer to their original images than the out-of-domain ones – ensures the effectiveness of LMD. With this in mind, we now discuss two simple and yet effective ways that can further improve detection performance consistently: multiple reconstructions and novel masking strategy.

Multiple Reconstructions. The DM inpainting process inherently involves multiple sampling steps. Occasionally, due to randomness, DM could provide dissimilar reconstructions for in-domain data, or similar reconstructions for out-of-domain data. This could make the reconstruction distance of the in-domain and out-of-domain images less separable, and hence leads to suboptimal OOD detection performance. To reduce the randomness, we perform multiple lifting and mapping attempts for each image. We calculate the OOD score from each attempt, and take the median[1] OOD score as the final OOD score for an image. As shown in 4.3, the simple median aggregation already provides strong performance. For further improvement, one may use a parameterized model to learn the distribution of the reconstruction distance across multiple attempts. We leave this to future work.

Novel Masking Strategy. The extent to which we mask an image is crucial to the detection performance. If the size of the mask is too large (or too small), the reconstruction distance for both in-domain and out-of-domain images would be very large (or very small). Indeed, if the mask covers the entire image, the reconstruction will be independent of the original image. In this case, if the in-domain manifold contains diverse images, an in-domain reconstruction can be far from its original image despite still being on the same manifold. Therefore, a suitable masking strategy should leave enough context to allow in-domain reconstructions to be similar to the original ones. To this end, we propose to use a checkerboard mask pattern. It divides an image into an $N \times N$ grid of image patches independent of the image size, and masks out half of the patches similar to a checkerboard. When multiple reconstruction attempts are performed, we also invert the masked and unmasked regions at each attempt. We call this masking strategy *alternating checkerboard* $N \times N$ (see Figure 3). Alternating checkerboard ensures all regions of the image to be masked with just two attempts. This avoids situations in which the distinguishing features of an out-of-domain image is never masked. LMD by default sets $N = 8$; ablation study on different mask choices can be found in Table 2.

Algorithm 2 Lift, Map, Detect (LMD)

Input: test image x, in-domain diffusion model θ_{in}
Output: OOD score of test image x
for $i = 1$ **to** r **do**
  $M_r \leftarrow$ `Get_Mask` (i)
  $x'_r \leftarrow$ `Inpaint` (x, M_r, θ_{in})
  $d_r \leftarrow$ `Distance` (x, x'_r)
end for
return `Aggregate` $(d_1, \ldots, d_r)$

[1] In our preliminary experiments, we find that median works the best than other simple aggregation methods such as mean.

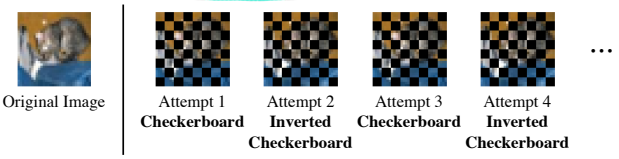


Figure 3. The alternating checkerboard mask pattern. We invert regions that are masked and unmasked at each reconstruction attempt. The example in the figure is 8×8.

4. Experiments

4.1. Experiment Settings

Evaluation Metric. LMD outputs an OOD score for each input, so in practice we need to apply a threshold to binarize the decision. In the experiments, we follow Hendrycks & Gimpel (2016); Ren et al. (2019); Xiao et al. (2021), and use the area under Receiver Operating Characteristic curve (ROC-AUC) as our quantitative evaluation metric.

Baselines. We compare our methods with four existing baselines: Likelihood (Bishop, 1994), Input Complexity (Serrà et al., 2019), Likelihood Regret (Xiao et al., 2020), and Pretrained Feature Extractor + Mahalanobis Distance (Xiao et al., 2021). Likelihood is obtained from the DM using the implementation from Song et al. (2020) [2]. For both Input Complexity and Likelihood Regret, we adapt the official GitHub repository of Likelihood Regret[3]. Specifically, to compute the Input Complexity, we use the likelihood calculated from the DM for a fair comparison, and convert the compression size to bits per dimension; we use the PNG compressor, because it yields the best performance among all available compressors in the GitHub repository. Pretrained Feature Extractor + Mahalanobis Distance is implemented by ourselves, as there is no existing publicly available implementations to our best knowledge.

Datasets. We perform OOD detection pairwise among CIFAR10 (Krizhevsky, 2009), CIFAR100 (Krizhevsky, 2009) and SVHN (Netzer et al., 2011), and pairwise among MNIST (LeCun et al., 2010), KMNIST (Clanuwat et al., 2018) and FashionMNIST (Xiao et al., 2017). For LMD and all the baselines, we use the training set of the in-domain dataset to train the model if needed, and evaluate the performance on the full test set of in-domain and out-of-domain datasets. Additionally, to demonstrate our performance on higher resolution images, we show qualitative results on CelebA-HQ (Karras et al., 2017) as in-domain and ImageNet (Russakovsky et al., 2015) as out-of-domain.

[2]https://github.com/yang-song/score_sde_pytorch
[3]https://github.com/XavierXiao/Likelihood-Regret

4.2. Implementation Details of LMD

We adapt the diffusion model implementation from Song et al. (2020). For experiments in Table 1, we use Song et al. (2020)'s pretrained checkpoint for CIFAR10, and we train DMs on the training set of the in-domain dataset for all the other datasets. We evaluate the OOD scores of the full in-domain and out-of-domain test sets. The inpainting reconstruction is repeated 10 times with alternating checkerboard 8×8 masks (Figure 3). For CelebA-HQ vs. ImageNet, we observe that CelebA-HQ does not have a train/test set split, and its pretrained checkpoint is trained over the full dataset. Thus, to avoid potential memorization issue, we use the pretrained FFHQ (Karras et al., 2019) checkpoint instead. We randomly sample a subset of size 100 from each dataset, and standardize all images to 256×256. We explore three mask choices: checkerboard 4×4, checkerboard 8×8, and a square centered mask. We reconstruct each image only *once*. We use LPIPS as the reconstruction distance metric to calculate the OOD score for all the experiments.

4.3. Experimental Results

Table 1 shows the performance of LMD and the baselines on various pairs of datasets. LMD achieves the highest performance on five pairs, with a maximum improvement of 10% (CIFAR100 vs. SVHN). LMD also achieves the second highest performance on five other pairs, and has the highest average ROC-AUC. This shows that LMD is consistent and versatile. We observe that the performance of the baselines are competitive on some pairs but limited on the others.

Figure 5 shows examples of the original, masked and inpainted images for three pairs. We show four reconstruction examples for each image, two with checkerboard mask and two with inverted checkerboard mask. The diffusion models reconstruct the in-domain images relatively accurately, while introducing a lot of artifacts in the out-of-domain inpaintings. For example, when SVHN is out-of-domain, the noise almost overwhelms the signals in the inpaintings.

Figure 6 shows the qualitative results and the ROC-AUC for CelebA-HQ vs. ImageNet. Checkerboard 8×8 performs competitively, achieving an ROC-AUC of 0.991 without any repeated reconstructions. Visually, the in-domain inpaintings look almost identical to the original images, while the out-of-domain inpaintings are locally incoherent. In this specific setting, checkerboard 4×4 and center masks yield slightly better performance. This is probably because faces are highly structured and provide strong inductive bias. Thus, with larger contiguous masked regions, the DM can still produce reasonably authentic reconstructions for the in-domain images, while being able to introduce more obvious artifacts for the out-of-domain images. Consequently, the reconstruction qualities of the two domains are more distinguishable. More discussion on mask choices can be

Table 1. ROC-AUC performance of LMD against various baselines on 12 pairs of datasets. Higher is better.

ID	OOD	LIKELIHOOD	IC	LR	PRETRAINED	LMD
CIFAR-10	CIFAR-100	0.520	0.568	0.546	**0.806**	0.607
	SVHN	0.180	0.870	0.904	0.888	**0.992**
CIFAR100	CIFAR-10	0.495	0.468	0.484	0.543	**0.568**
	SVHN	0.193	0.792	0.896	0.776	**0.985**
SVHN	CIFAR-10	0.974	0.973	0.805	**0.999**	0.914
	CIFAR-100	0.970	0.976	0.821	**0.999**	0.876
MNIST	KMNIST	0.948	0.903	**0.999**	0.887	0.984
	FASHIONMNIST	0.997	**1.000**	0.999	0.999	0.999
KMNIST	MNIST	0.152	0.951	0.431	0.582	**0.978**
	FASHIONMNIST	0.833	**0.999**	0.557	0.993	0.993
FASHIONMNIST	MNIST	0.172	0.912	0.971	0.647	**0.992**
	KMNIST	0.542	0.584	**0.994**	0.730	0.990
AVERAGE		0.581	0.833	0.783	0.821	**0.907**

found in Section 4.4.

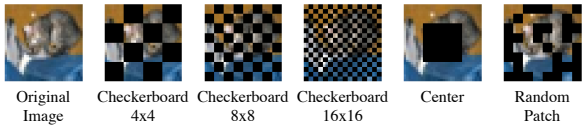


Figure 4. Visualization of the masks used in the mask ablation. For the random patch mask, this figure only shows one example; we sample a different pattern at each reconstruction attempt.

4.4. Ablation

Effects of Mask Choices. Table 2 shows ablation results on different types of mask patterns (see Figure 4). Specifically, we examine the following patterns: alternating checkerboard 4×4 and 16×16, a fixed non-alternating 8×8 checkerboard, a square centered mask covering one-fourth of an image (*center*), and a random patch mask covering 50% of an 8×8 patch grid (*random patch*) introduced in Xie et al. (2022)[4].

Alternating checkerboard 8×8 performs consistently across the three datasets, while other patterns have fluctuation in their performance. In particular, alternating checkerboard 4×4 performs poorly on MNIST vs. KMNIST. This suggests that if the masked patches are too large, both in-domain and out-of-domain reconstructions may be dissimilar from the original images. Fixed checkerboard 8×8 performs only slightly worse than its alternating counterpart, usually with a performance drop of less than 0.01. This may be because for these datasets, the distinguishing features of the in-domain and out-of-domain images exist in many patches. Thus, they can already be captured well enough by the fixed 8×8 mask. Nevertheless, the alternating checkerboard pattern should still be preferred, since it can mask the entire image across multiple reconstruction attempts. Not surprisingly, the center mask exhibits very poor performance (0.444) on MNIST vs. KMNIST, as it removes too much information from the images. Lastly, we find that the random patch mask also performs competitively.

Effects of Reconstruction Distance Metrics. LMD needs to assess the reconstruction distance of the inpaintings from the DM, so we explore three off-the-shelf reconstruction distance metrics: Mean Squared Error (MSE), Structural Similarity Index Measure (SSIM) (Wang et al., 2003), and Learned Perceptual Image Patch Similarity (LPIPS) (Zhang et al., 2018). Additionally, Xiao et al. (2021) demonstrates strong performance using SimCLRv2 (Chen et al., 2020) representations, so we experiment with a SimCLRv2-based error metric too. Specifically, we calculate the cosine distance between the SimCLRv2 representations of the original and reconstructed images, which we simply refer as SimCLRv2. These four reconstruction distance metrics range from shallow reference based to deep feature based metrics.

We summarize the results of three dataset pairs in Table 3. LPIPS is competitive on all three dataset pairs, while MSE, SSIM and SimCLRv2 fluctuate in their performance. Interestingly, on CIFAR10 vs. CIFAR100, SimCLRv2 outperforms other metrics significantly, with an improvement of 0.09. In Table 1, Xiao et al. (2021) also outperforms all other methods on CIFAR10 vs. CIFAR100 using SimCLRv2 representations. This indicates that some reconstruction distance metrics can be particularly suitable for specific domains.

Number of Reconstruction Attempts per Image. We also

[4]https://github.com/microsoft/SimMIM

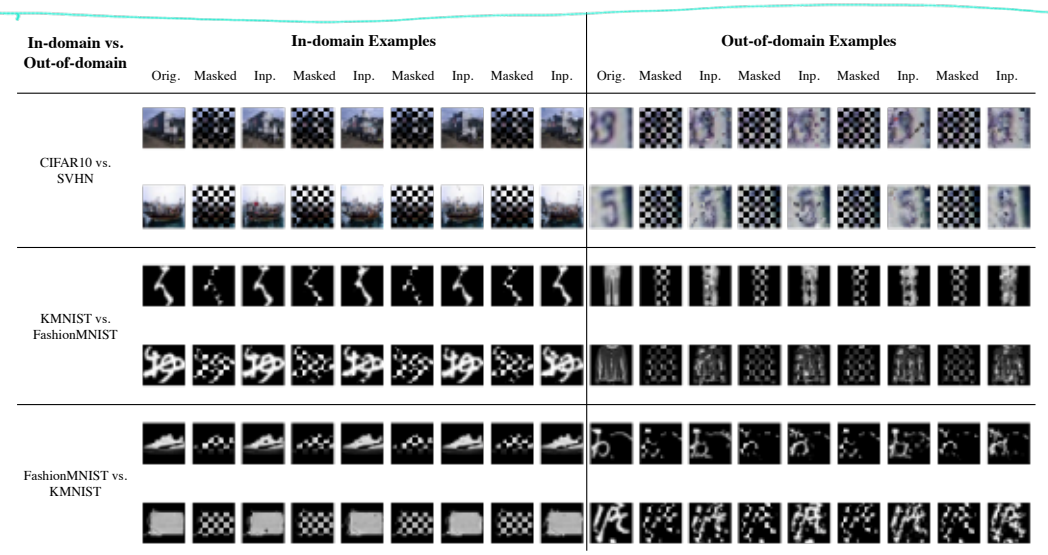


Figure 5. Reconstruction examples from three dataset pairs. "Orig." stands for the original image; "Inp." stands for the inpainted image. In general, the in-domain reconstructions are close to their original images, while the out-of-domain reconstructions are noisy and different from the original ones.

Table 2. ROC-AUC on three dataset pairs with different mask types. Alternating checkerboard 8×8 shows strong and consistent performance.

MASK TYPE	CIFAR10 VS. CIFAR100	CIFAR10 VS. SVHN	MNIST VS. KMNIST
ALTERNATING CHECKERBOARD 4×4	0.594	0.987	0.923
ALTERNATING CHECKERBOARD 8×8	**0.607**	**0.992**	0.984
ALTERNATING CHECKERBOARD 16×16	0.597	0.981	**0.997**
FIXED CHECKERBOARD 8×8	0.601	0.990	0.974
CENTER	0.570	0.978	0.479
RANDOM PATCH	0.591	0.990	0.912

Table 3. ROC-AUC on three dataset pairs with different reconstruction distance metrics. LPIPS shows strong and consistent performance.

RECON. METRIC	CIFAR10 VS. CIFAR100	MNIST VS. KMNIST	KMNIST VS. MNIST
MSE	0.548	**0.998**	0.835
SSIM	0.624	0.997	0.922
LPIPS	0.607	0.984	**0.978**
SIMCLRV2	**0.713**	0.983	0.920

study the effect of the number of reconstruction attempts on the performance. Figure 7 shows the ROC-AUC from one attempt to ten attempts per image for two pairs of datasets. In both dataset pairs, increasing the number of attempts almost always improves the ROC-AUC. The improvement is especially significant initially, and saturates at around ten attempts. The improvement is consistent across all four error metrics, further supporting the effectiveness of LMD's multiple reconstructions approach.

Table 4. ROC-AUC performance of our OOD detection framework with an alternative lifting and mapping instantiation – diffusion and denoising. It shows strong performance, although it is slightly outperformed by our default choice of masking and inpainting.

MAPPING METHOD	CIFAR10 VS. CIFAR100	CIFAR10 VS. SVHN	FASHIONMNIST VS. MNIST
DENOISING	0.600	0.976	0.941
INPAINTING	**0.607**	**0.992**	**0.992**

Alternative Way of Lifting and Mapping. Alternatively, we can lift an image by diffusion, and map it by denoising. Table 4 shows the performance of our OOD detection framework under this instantiation on three dataset pairs. In our experiments, we add noise to step $t = 500$ in each attempt, where $T = 1000$, as it generally yields good results. Similar

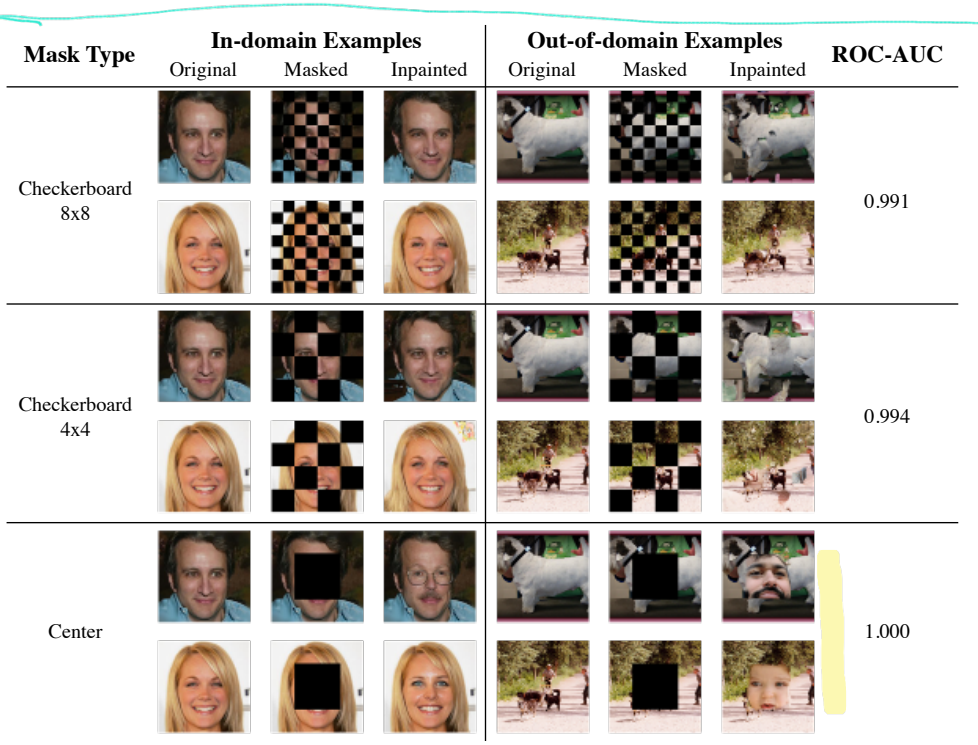
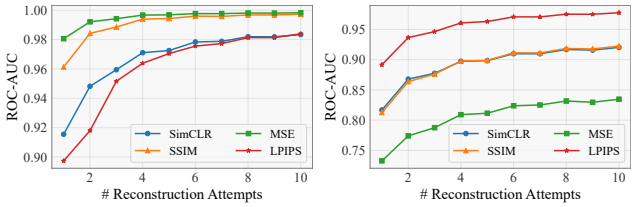


Figure 6. Reconstruction examples from CelebA-HQ (in-domain) and ImageNet (out-of-domain) using different masks. For out-of-domain inpaintings, the checkerboard masks introduce locally incoherent artifacts, while the center mask introduces face-like artifacts. This makes the out-of-domain images highly distinguishable.



(a) MNIST vs. KMNIST (b) KMNIST vs. MNIST

Figure 7. ROC-AUC against number of reconstruction attempts on two pairs of datasets. As the number of reconstruction attempts increases, the OOD detection performance improves regardless of the choice of the reconstruction distance metric.

to our inpainting setting, we perform 10 attempts per image, and use the median reconstruction error under LPIPS as the OOD score. We observe that diffusion/denoising is also competitive, although it is slightly outperformed by masking/inpainting in some cases. This indicates that our framework is generally applicable in OOD detection, and supports various promising alternative instantiations.

5. Discussion and Conclusion

One limitation of the vanilla diffusion model is that the denoising process involves many iterations and is thus slow. Consequently, like many DM-based algorithms in other applications (Meng et al., 2021; Lugmayr et al., 2022), LMD is hard to be applied to real-time OOD detection at the current stage. Recently, there has been a popular line of work on speeding up diffusion models without retraining. For example, Nichol & Dhariwal (2021) re-scales the noise schedule to skip sampling steps, Liu et al. (2022) proposes pseudo numerical methods for diffusion models, and Watson et al. (2022) optimizes fast samplers that enable sampling with only 10-20 steps. This opens up a promising direction for future work to integrate these methods into LMD.

In conclusion, we leverage the diffusion model's manifold mapping ability, and propose a method – Lift, Map, Detect (LMD) – for unsupervised OOD detection. We show that it is competitive and versatile through our experiments.

6. Acknowledgement

This research is supported by grants from DARPA AIE program, Geometries of Learning (HR00112290078), the Natural Sciences and Engineering Research Council of Canada (NSERC) (567916), the National Science Foundation NSF (IIS-2107161, III1526012, IIS-1149882, and IIS-1724282), and the Cornell Center for Materials Research with funding from the NSF MRSEC program (DMR-1719875).

References

Ahmadian, A. and Lindsten, F. Likelihood-free out-of-distribution detection with invertible generative models. In *IJCAI*, pp. 2119–2125, 2021.

Bergamin, F., Mattei, P.-A., Havtorn, J. D., Senetaire, H., Schmutz, H., Maaløe, L., Hauberg, S., and Frellsen, J. Model-agnostic out-of-distribution detection using combined statistical tests. In *International Conference on Artificial Intelligence and Statistics*, pp. 10753–10776. PMLR, 2022.

Bergman, L. and Hoshen, Y. Classification-based anomaly detection for general data. *arXiv preprint arXiv:2005.02359*, 2020.

Bishop, C. M. Novelty detection and neural network validation. *IEE Proceedings-Vision, Image and Signal processing*, 141(4):217–222, 1994.

Chen, T., Kornblith, S., Swersky, K., Norouzi, M., and Hinton, G. E. Big self-supervised models are strong semi-supervised learners. *Advances in neural information processing systems*, 33:22243–22255, 2020.

Choi, H., Jang, E., and Alemi, A. A. Waic, but why? generative ensembles for robust anomaly detection. *arXiv preprint arXiv:1810.01392*, 2018.

Clanuwat, T., Bober-Irizar, M., Kitamoto, A., Lamb, A., Yamamoto, K., and Ha, D. Deep learning for classical japanese literature. 2018.

Denouden, T., Salay, R., Czarnecki, K., Abdelzad, V., Phan, B., and Vernekar, S. Improving reconstruction autoencoder out-of-distribution detection with mahalanobis distance. *arXiv preprint arXiv:1812.02765*, 2018.

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial networks, 2014. URL `https://arxiv.org/abs/1406.2661`.

Graham, M. S., Pinaya, W. H., Tudosiu, P.-D., Nachev, P., Ourselin, S., and Cardoso, M. J. Denoising diffusion models for out-of-distribution detection. *arXiv preprint arXiv:2211.07740*, 2022.

Hamet, P. and Tremblay, J. Artificial intelligence in medicine. *Metabolism*, 69:S36–S40, 2017.

Hendrycks, D. and Gimpel, K. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv preprint arXiv:1610.02136*, 2016.

Hendrycks, D., Mazeika, M., and Dietterich, T. Deep anomaly detection with outlier exposure. *arXiv preprint arXiv:1812.04606*, 2018.

Hendrycks, D., Mazeika, M., Kadavath, S., and Song, D. Using self-supervised learning can improve model robustness and uncertainty. *Advances in neural information processing systems*, 32, 2019.

Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

Huang, R., Geng, A., and Li, Y. On the importance of gradients for detecting distributional shifts in the wild. *Advances in Neural Information Processing Systems*, 34: 677–689, 2021.

Karras, T., Aila, T., Laine, S., and Lehtinen, J. Progressive growing of gans for improved quality, stability, and variation. *CoRR*, abs/1710.10196, 2017. URL `http://arxiv.org/abs/1710.10196`.

Karras, T., Laine, S., and Aila, T. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4401–4410, 2019.

Kingma, D. P. and Dhariwal, P. Glow: Generative flow with invertible 1x1 convolutions. *Advances in neural information processing systems*, 31, 2018.

Kirichenko, P., Izmailov, P., and Wilson, A. G. Why normalizing flows fail to detect out-of-distribution data. *Advances in neural information processing systems*, 33: 20578–20589, 2020.

Krizhevsky, A. Learning multiple layers of features from tiny images. Technical report, 2009.

LeCun, Y., Cortes, C., and Burges, C. Mnist handwritten digit database. *ATT Labs [Online]. Available: http://yann.lecun.com/exdb/mnist*, 2, 2010.

Lee, K., Lee, K., Lee, H., and Shin, J. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. *Advances in neural information processing systems*, 31, 2018.

Li, D., Chen, D., Goh, J., and Ng, S.-k. Anomaly detection with generative adversarial networks for multivariate time series. *arXiv preprint arXiv:1809.04758*, 2018.

Liang, S., Li, Y., and Srikant, R. Enhancing the reliability of out-of-distribution image detection in neural networks. *arXiv preprint arXiv:1706.02690*, 2017.

Liu, L., Ren, Y., Lin, Z., and Zhao, Z. Pseudo numerical methods for diffusion models on manifolds. *arXiv preprint arXiv:2202.09778*, 2022.

Lugmayr, A., Danelljan, M., Romero, A., Yu, F., Timofte, R., and Van Gool, L. Repaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11461–11471, 2022.

Maaløe, L., Fraccaro, M., Liévin, V., and Winther, O. Biva: A very deep hierarchy of latent variables for generative modeling. *Advances in neural information processing systems*, 32, 2019.

Meng, C., He, Y., Song, Y., Song, J., Wu, J., Zhu, J.-Y., and Ermon, S. Sdedit: Guided image synthesis and editing with stochastic differential equations. In *International Conference on Learning Representations*, 2021.

Morningstar, W., Ham, C., Gallagher, A., Lakshminarayanan, B., Alemi, A., and Dillon, J. Density of states estimation for out of distribution detection. In *International Conference on Artificial Intelligence and Statistics*, pp. 3232–3240. PMLR, 2021.

Nalisnick, E., Matsukawa, A., Teh, Y. W., Gorur, D., and Lakshminarayanan, B. Do deep generative models know what they don't know? *arXiv preprint arXiv:1810.09136*, 2018.

Nalisnick, E. T., Matsukawa, A., Teh, Y. W., and Lakshminarayanan, B. Detecting out-of-distribution inputs to deep generative models using a test for typicality. 2019.

Netzer, Y., Wang, T., Coates, A., Bissacco, A., Wu, B., and Ng, A. Y. Reading digits in natural images with unsupervised feature learning. 2011.

Nichol, A. Q. and Dhariwal, P. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*, pp. 8162–8171. PMLR, 2021.

Ren, J., Liu, P. J., Fertig, E., Snoek, J., Poplin, R., Depristo, M., Dillon, J., and Lakshminarayanan, B. Likelihood ratios for out-of-distribution detection. *Advances in neural information processing systems*, 32, 2019.

Rigano, C. Using artificial intelligence to address criminal justice needs. *National Institute of Justice Journal*, 280: 1–10, 2019.

Ruff, L., Vandermeulen, R. A., Görnitz, N., Binder, A., Müller, E., Müller, K.-R., and Kloft, M. Deep semi-supervised anomaly detection. *arXiv preprint arXiv:1906.02694*, 2019.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., and Fei-Fei, L. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015. doi: 10.1007/s11263-015-0816-y.

Sakurada, M. and Yairi, T. Anomaly detection using autoencoders with nonlinear dimensionality reduction. In *Proceedings of the MLSDA 2014 2nd workshop on machine learning for sensory data analysis*, pp. 4–11, 2014.

Salimans, T., Karpathy, A., Chen, X., and Kingma, D. P. Pixelcnn++: Improving the pixelcnn with discretized logistic mixture likelihood and other modifications. *arXiv preprint arXiv:1701.05517*, 2017.

Schlegl, T., Seeböck, P., Waldstein, S. M., Schmidt-Erfurth, U., and Langs, G. Unsupervised anomaly detection with generative adversarial networks to guide marker discovery. In *International conference on information processing in medical imaging*, pp. 146–157. Springer, 2017.

Sehwag, V., Chiang, M., and Mittal, P. Ssd: A unified framework for self-supervised outlier detection. *arXiv preprint arXiv:2103.12051*, 2021.

Serrà, J., Álvarez, D., Gómez, V., Slizovskaia, O., Núñez, J. F., and Luque, J. Input complexity and out-of-distribution detection with likelihood-based generative models. *arXiv preprint arXiv:1909.11480*, 2019.

Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., and Ganguli, S. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pp. 2256–2265. PMLR, 2015.

Song, Y. and Ermon, S. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32, 2019.

Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., and Poole, B. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

Tack, J., Mo, S., Jeong, J., and Shin, J. Csi: Novelty detection via contrastive learning on distributionally shifted instances. *Advances in neural information processing systems*, 33:11839–11852, 2020.

Wang, H., Li, Z., Feng, L., and Zhang, W. Vim: Out-of-distribution with virtual-logit matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4921–4930, 2022.

Wang, Z., Simoncelli, E. P., and Bovik, A. C. Multiscale structural similarity for image quality assessment. In *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*, volume 2, pp. 1398–1402. Ieee, 2003.

Watson, D., Chan, W., Ho, J., and Norouzi, M. Learning fast samplers for diffusion models by differentiating through sample quality. In *International Conference on Learning Representations*, 2022.

Xia, Y., Cao, X., Wen, F., Hua, G., and Sun, J. Learning discriminative reconstructions for unsupervised outlier removal. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1511–1519, 2015.

Xiao, H., Rasul, K., and Vollgraf, R. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *CoRR*, abs/1708.07747, 2017. URL http://arxiv.org/abs/1708.07747.

Xiao, Z., Yan, Q., and Amit, Y. Likelihood regret: An out-of-distribution detection score for variational auto-encoder. *Advances in neural information processing systems*, 33: 20685–20696, 2020.

Xiao, Z., Yan, Q., and Amit, Y. Do we really need to learn representations from in-domain data for outlier detection? *arXiv preprint arXiv:2105.09270*, 2021.

Xie, Z., Zhang, Z., Cao, Y., Lin, Y., Bao, J., Yao, Z., Dai, Q., and Hu, H. Simmim: A simple framework for masked image modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9653–9663, 2022.

Yang, L., Zhang, Z., Song, Y., Hong, S., Xu, R., Zhao, Y., Shao, Y., Zhang, W., Cui, B., and Yang, M.-H. Diffusion models: A comprehensive survey of methods and applications. *arXiv preprint arXiv:2209.00796*, 2022.

Zhang, R., Isola, P., Efros, A. A., Shechtman, E., and Wang, O. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 586–595, 2018.

Zhou, C. and Paffenroth, R. C. Anomaly detection with robust deep autoencoders. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 665–674, 2017.

Zong, B., Song, Q., Min, M. R., Cheng, W., Lumezanu, C., Cho, D., and Chen, H. Deep autoencoding gaussian mixture model for unsupervised anomaly detection. In *International conference on learning representations*, 2018.