ZERO-SHOT ANOMALY DETECTION FROM LATENT DIFFUSION MODELS

Leonardo Zavala-Jimenez, Diane Kim, Darren Choe, Cara Mann, Bowei Cheng, Om Khangaonkar, Adithi Sumitran, Chase Karlsson, Emily Mao

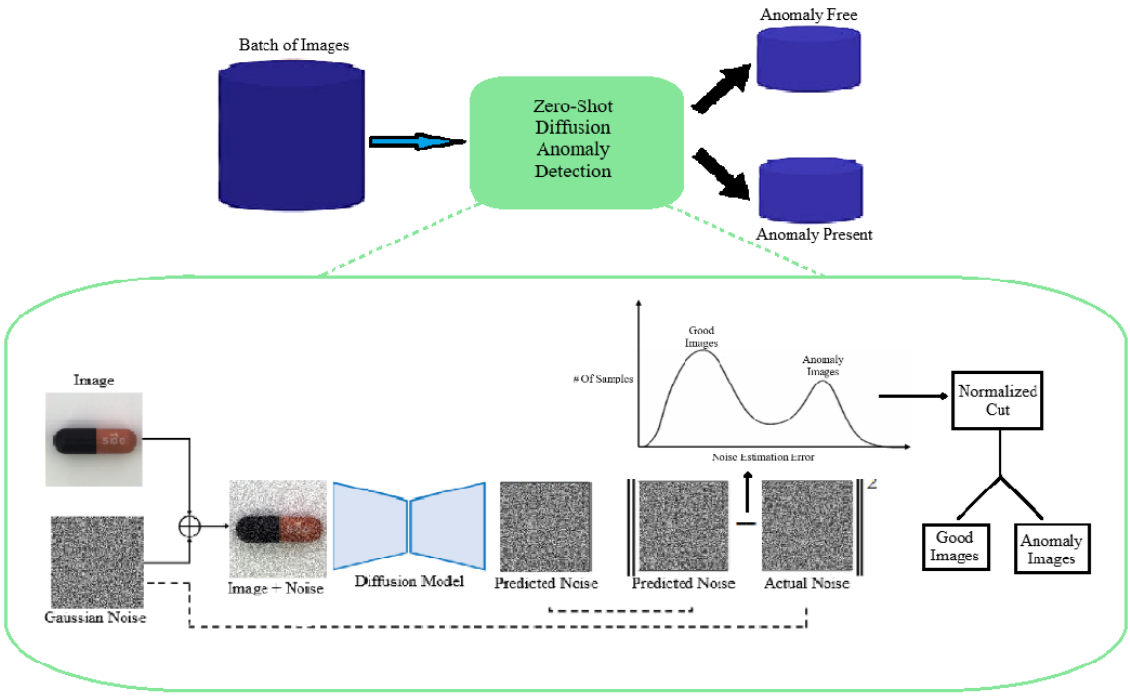

Figure 1: The top section illustrates the end-to-end pipeline where a batch of images is processed into anomaly and anomaly-free groupings. The lower section details the zero-shot diffusion anomaly detection method using a latent diffusion model and normalized cut clustering for anomaly classification.

1 INTRODUCTION

Anomaly detection is primarily defined by two seemingly opposing qualities: *context* and *detail*. To detect anomalies, one must have a strong *contextual* understanding of the object we are trying to classify and its relationship to other objects similar to it. At the same time, most anomalies present themselves in low-level *details* that do not fundamentally affect a semantic understanding of the object (i.e., its color or a scratch on its surface).

In many practical scenarios, this dual requirement poses significant challenges. Systems must balance the ability to capture high-level semantic cues with the precision needed to detect subtle, local variations. This balance becomes even more critical in fields such as industrial inspection or medical diagnostics, where missing a minor anomaly can have substantial consequences. Recent advances in machine learning, particularly in the domain of deep generative models, have paved the way for innovative solutions to these challenges.

We hypothesize that large-scale generative models are uniquely well posed to tackle anomaly detection. Not only do they need to learn and understand contextual semantic relationships (via language-image supervision), but they also need to model low-level details as well (to produce appealing images). Furthermore, these models are trained on *huge* datasets, such as Schuhmann et al. (2022), which contains over 5 billion text-image pairs. This massive amount of data enables the models to develop a rich, multifaceted representation of visual information. As a result, they are capable of discerning even subtle deviations from expected patterns, which is essential for effective anomaly detection.

Inspired by recent work that adapts large-scale text-to-image models for zero-shot image classification Li et al. (2023), we propose a novel method to adapt these models to zero-shot anomaly detection. Notably, unlike methods such as the one-class SVM, which requires a labeled set of "good" examples, our model is fully zero-shot, with no labels required at any point in the detection process. This zero-shot capability not only simplifies the deployment of anomaly detection systems by removing the reliance on extensive, curated datasets, but it also enhances their adaptability across different domains and applications.

Our approach leverages the inherent strengths of diffusion-based generative models. The generative features these models learn effectively capture both global semantic features and fine-grained details, which are then used to compute noise estimation errors across the reconstruction process. These errors serve as a proxy for anomaly scores, allowing the system to distinguish between normal and abnormal instances even when explicit labels are not available. Through clustering techniques, such as the normalized cut, our framework is able to segregate anomalous images based on their deviation from the learned distribution.

A critical component of our approach is the concept of score matching, which plays a pivotal role in how our generative model learns a distribution of images. In score matching, the model is trained to estimate the gradient of the log-density of the data, effectively capturing the structure and intricacies of the image distribution Song & Ermon (2019); Song et al. (2020). Regular images, which typically lie near the high-density modes of this learned distribution, yield stable and accurate score estimates. Conversely, abnormal images tend to fall in lower-density regions, leading to larger noise prediction errors during the reconstruction process. This natural discrepancy between normal and abnormal instances forms the foundation for our anomaly scoring, as higher noise prediction errors indicate deviations from the learned norm.

The remainder of the paper is organized as follows. Section 2 provides an overview of related work, highlighting advances in generative modeling and their applications in anomaly detection. Section 3 describes the dataset and experimental setup used to evaluate our method. In Section 4, we detail our proposed approach and present comprehensive experimental results that demonstrate the efficacy of our model compared to traditional methods. Finally, Section 5 concludes with a discussion of our findings and outlines potential directions for future research.

This work represents a step towards more robust and flexible anomaly detection systems that are capable of generalizing across diverse datasets and application domains. By harnessing the power of large-scale generative models and leveraging techniques like score matching, we aim to reduce the dependency on labeled data while improving the sensitivity and specificity of anomaly detection mechanisms.

2 Literature Review

2.1 Generative Models for Synthesis

Recent efforts on image synthesis have focused on diffusion-based techniques that offer superior quality and versatility. Diffusion models, first introduced by Sohl-Dickstein et al. (2015) and later refined by Ho et al. (2020) into Denoising Diffusion Probabilistic Models (DDPM), achieve impressive synthesis quality through iterative denoising. Recent works, such as Rombach et al. (2021) have enabled scalable, text-to-image generation by performing denoising in the low(er)-dimensional latent space of a VAE (Kingma & Welling (2013)), substantially reducing training and inference costs. In recent years, industry-scale efforts have pushed these models to new heights, leveraging massive datasets and extensive computational resources to produce high-quality, photorealistic images with diverse and creative outputs.

2.2 Generative Modeling and Representation Learning

Generative models, originally developed for data synthesis, have increasingly been adapted for perception tasks in computer vision. A longstanding viewpoint in the field (dating back to Hinton's early work) posits that learning to *generate* data can aid in *recognizing* it Hinton (2007). Early work on GANs (Goodfellow et al. (2014)) evaluated whether representations learned by generating images (Radford et al. (2015)) or videos (Vondrick et al. (2016)) transferred well to image classification or action recognition, respectively, but performance was always far below discriminatively pretrained models.

Recently, generative and discriminative models have also been observed to learn "Rosetta" neurons that share matching activations on similar visual concepts (Dravid et al. (2023)). This suggests, as further highlighted in Huh et al. (2024), that as networks are scaled up, similar representations are likely to emerge, regardless of pretraining task. A key advantage of large-scale diffusion models, such as Stable Diffusion, is their industry-scale pretraining; the sheer scale of pretraining on over 2 billion images has the potential to outscale any discriminatively trained model.

For example, Li et al. (2023) utilizes Stable Diffusion's shared language-image representation to construct a zero-shot classifier. Notably, their Diffusion Classifier outperforms supervised ResNets He et al. (2016) and ViTs Dosovitskiy et al. (2020) in robustness, and in some cases, even accuracy. While this may seem surprising, generative modeling shares strong connections with many self-supervised representation learning methods.

For example, Vincent's seminal work on Denoising Autoencoders (DAEs) (Vincent et al. (2010)) highlighted how training an autoencoder to reconstruct clean inputs from their corrupted versions enables it to learn high-level features that transfer to tasks such as edge detection. Pathak et al. (2016) showed that training a model to inpaint the center of a masked image can achieve impressive results from unlabeled data. He et al. (2022) has achieved state-of-the art in a large amount of visual tasks by pretraining to reconstruct masked image tokens, and then fine-tuning for discriminative tasks. Zimmermann et al. (2021) also shows that seemingly discriminative methods that use contrastive learning (Oord et al. (2018); Chen et al. (2020); He et al. (2020)) actually share a strong connection with generative modeling by learning to invert the generation process.

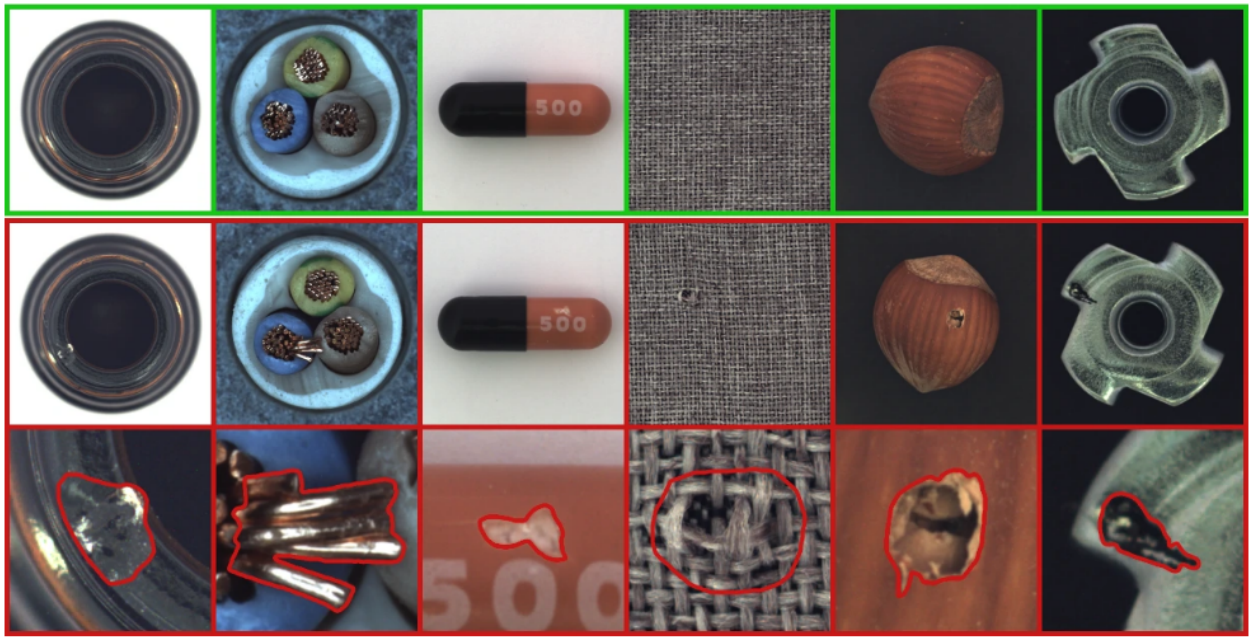

Figure 2: 6 of the 15 MVTec Object Categories With and Without Anomalies (Adapted from Bergmann et al. (2019))

3 Dataset Description

We evaluate on the MVTec Anomaly Detection dataset. The dataset contains fifteen categories of close up high resolution pictures of objects (bottle, cable, capsule, carpet, grid, hazelnut, leather, metal nut, pill, screw, tile, toothbrush, transistor, wood, zipper), with each category being divided into anomaly free images, and images labeled by anomaly type present. For example, the hazelnut object is divided into four anomaly types, crack, cut, hole, and print, along with a collection of anomaly free(good) images. The entire dataset is roughly 6000 images with around $\frac{2}{3}$ of the images being anomaly free. The size of the images for most(11) of the categories is 1024×1024, with the smallest being 700×700 for metal nut. We combine the anomaly and non-anomaly images in each category for our evaluation, in order to provide unlabeled data for classification. We show qualitative samples in Figure 2.

4 Proposed Solution and Experimental Results

4.1 Generative Diffusion Models

We begin with a brief introduction to diffusion models. Forward diffusion models gradually add noise to data, which transforms clear images into pure noise over several steps. Reverse diffusion models do the reverse, by removing this noise which recovers the images from random noise. Latent diffusion models add to this approach by operating in a lower-dimensional latent space created using a variational autoencoder which tries to emphasize semantic features in the images. This additionally reduces complexity allowing for faster training and inference.

4.2 Zero-Shot Diffusion Anomaly Detection Method

Our approach to zero-shot classification uses a diffusion based generative model. We divide our approach into two stages: an error estimation through a diffusion process, and anomaly classification via a normalized cut.

In the first stage we use stable diffusion as our latent diffusion model. We first pre-process every image in the batch set by standardizing the image sizes to fit into the latent diffusion models framework to ensure consistency. Then each image, x, is input into the latent diffusion model, encoding it into a latent representation using a variational autoencoder(VAE). We let $z = E(x)$ be the latent representation. For each timestep t, randomly sampled gaussian noise is added in latent space, $\tilde{z}_t = \sqrt{\alpha_t} z + \sqrt{1 - \alpha_t}\epsilon_t$ where α_t is the respective scaling value from the noise schedule, and ϵ_t represents the randomly sampled gaussian noise. The diffusion model then predicts the noise to remove from this latent representation, $\hat{\epsilon}_t = f(\tilde{z}_t, t, \emptyset)$. We parametrize f with Stable Diffusion (Rombach et al. (2021)); however, any diffusion model can be used in practice. We condition f on $\emptyset$, the null condition. Then, our noise estimation error is $\mathbb{E} = \mathcal{L}(\epsilon_t, \hat{\epsilon}_t)$, where $\mathcal{L}$ is mean squared error. By taking the average noise estimation error $\mathbb{E}$ over all time-steps, we see the average deviation of the predicted noise from the added gaussian noise.

In the second stage, after attaining all the average noise estimation errors, we focus on the clusterings of the errors to differentiate between anomaly and non anomaly images. The latent diffusion model when an anomaly was present, treats these anomaly perturbations in the images as noise, creating larger noise estimation errors, and causing anomalies to cluster further from the average noise estimation errors of good images. We employ the normalized cut approach (Shi & Malik (2000)), which "cuts" data into a fixed number of clusters. Normalized cut is generally used in complex scenarios where classical methods such as k-means fail. This cut separates the average noise estimation errors into two clusters, and we make the assumption the larger cluster will contain the good images, and the smaller cluster as those of anomalies.

We run all experiments on four RTX A6000 GPUs. We use the libraries pytorch, huggingface, numpy, scitkit-learn, and matplotlib (for figures). We source all model weights from HuggingFace.

4.3 BASELINES

One-Class Support Vector Machine (OC-SVM) is a widely used anomaly detection method that learns a decision boundary to distinguish normal images from potential outliers. Given a set of training images, we first resize all images to 64×64 and flatten them into one-dimensional feature vectors. OC-SVM then maps these vectors into a high-dimensional feature space using a kernel function and optimizes a hypersphere that encloses the majority of the data, effectively modeling the distribution of normal images. At test time, new images are classified as normal if they fall within this learned boundary and anomalous otherwise. We use OC-SVM as a baseline for anomaly detection, training it exclusively on normal samples to establish a threshold for novelty detection. The effectiveness of this approach depends on the choice of kernel and hyperparameters, which we tune using cross-validation on a held-out subset of normal data.

	Ours (Normalized Cut)		**One-Class SVM**	
Object	ACC	REC	ACC	REC
bottle	56.85%	53.97%	74.66%	42.86%
cable	63.37%	34.78%	66.04%	25.00%
capsule	60.97%	2.75%	51.00%	25.69%
carpet	62.97%	58.43%	70.53%	21.35%
grid	48.83%	42.11%	75.15%	22.81%
hazelnut	63.07%	80.00%	67.07%	28.57%
leather	56.91%	57.61%	67.21%	15.22%
metal_nut	44.18%	33.33%	69.25%	30.11%
screw	68.33%	68.07%	63.75%	15.97%
pill	40.32%	12.77%	67.05%	21.28%
tile	50.14%	47.62%	70.61%	16.67%
toothbrush	54.90%	33.33%	69.61%	36.67%
transistor	53.99%	47.50%	75.08%	30.00%
wood	58.59%	40.00%	73.01%	6.67%
zipper	70.59%	72.27%	62.92%	26.05%
Mean	56.93%	**45.64%**	68.20%	24.33%

Table 1: Comparison of our proposed method and One-Class SVM performance metrics on the MVTec Dataset. Metrics abbreviations: ACC = Overall Accuracy, REC = Test Recall (Anomaly)

4.4 ZERO SHOT ANOMALY DETECTION

Using the noise estimation error, we performed Normalized Cut clustering on our data. We also used One-Class SVM and compared our results to these in terms of accuracy and recall. The mean accuracy of the Normalized Cut was 56.93%, was slightly lower than the accuracy of the One-Class SVM, 68.2%. This suggests that in terms of simply clustering the objects into the groups of defective or non-defective, the One-Class SVM model performed marginally better.

However, in the process of anomaly detection, what truly matters is appropriately identifying anomalies. For our purposes, it does not matter if a non-defective product is classified as defective, nor does it matter if the overall classification is high. We are merely interested in catching all anomalies, because in the application context, we need to make sure that all defective products are being caught (meaning we must maximize the true positive rate). This means that the recall value needs to be higher in our model than in the One-Class SVM model in order for us to conclude that our model is superior. We calculated the mean recall for both of the classes and found that the mean recall for Normalized Cut was much higher than that of the One-Class SVM model (45.64 compared to 24.33, a nearly 20% difference). Given how much higher the average recall was for Normalized Cut based on diffusion compared to One-Class SVM, we can say with some confidence that the diffusion model is more suitable in the context of anomaly detection than One-Class SVM.

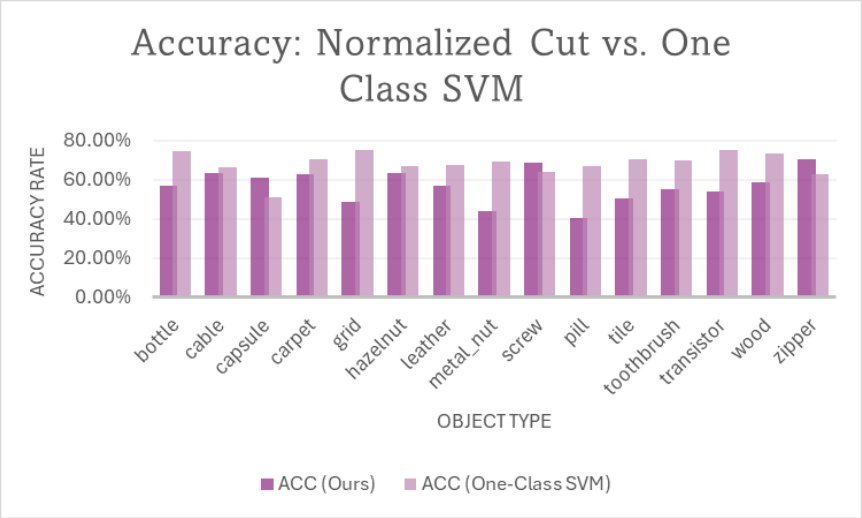

Figure 3: Overall accuracy between our method and One-Class SVM on anomaly classification

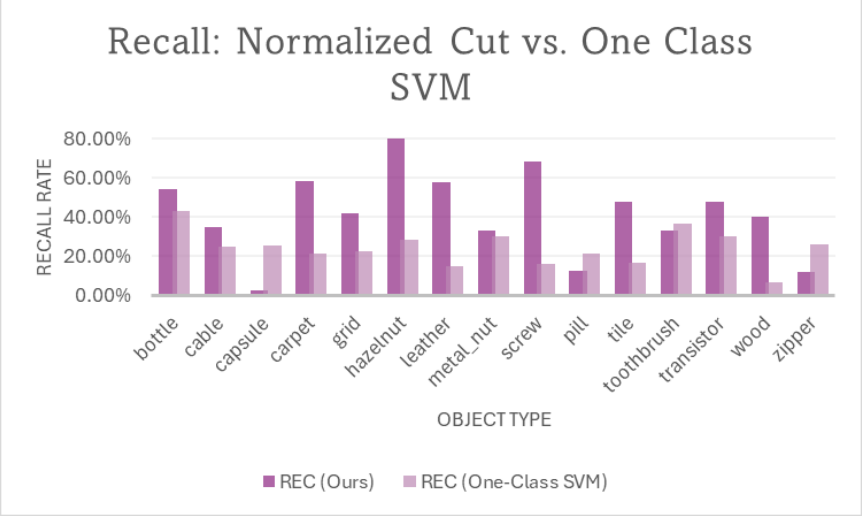

Figure 4: Anomaly recall between the our method and One-Class SVM on anomaly classification

Interestingly, while the mean recall was much higher for Normalized Cut than One-Class SVM, on the individual object level this was not always the case. Both the pill and capsule had higher recall for One-Class SVM. Given the intricacy of pills and capsules and their generally smooth textures, this might be indicative of a potential weakness of the model in handling more minute defects. This possibility is further supported by the fact that more textured objects, like the carpet, hazelnut, and zipper, had much higher recall with Normalized Cut than One-Class SVM.

4.5 Analysis of Clustering Methods

As shown in Table 2, we evaluated three clustering methods, Kmeans, Normalized Cut, and the Gaussian Mixture Model (GMM) on the mean noise estimation error distributions.

Normalized Cut is a clustering method where first we construct a similarity matrix where each image is compared to every other image. Now we have a fully connected graph where the edge weights are defined as the distances between the noise estimation errors of the images. Normalized cut then creates a division

	Ours (Normalized Cut)		KMeans		GMM	
Object	ACC	REC	ACC	REC	ACC	REC
bottle	56.85%	53.97%	63.70%	53.97%	83.22%	33.33%
cable	63.37%	34.78%	63.37%	34.78%	73.53%	21.74%
capsule	60.97%	2.75%	61.25%	5.50%	62.68%	0.00%
carpet	62.97%	58.43%	62.97%	56.18%	65.49%	51.69%
grid	48.83%	42.11%	52.92%	57.89%	52.92%	57.89%
hazelnut	63.07%	80.00%	63.07%	80.00%	69.46%	78.57%
leather	56.91%	57.61%	46.61%	33.70%	49.32%	30.43%
metal_nut	44.18%	33.33%	46.57%	33.33%	46.27%	24.73%
screw	68.33%	68.07%	68.33%	68.07%	66.25%	44.54%
pill	40.32%	12.77%	40.32%	12.77%	39.86%	9.93%
tile	50.14%	47.62%	51.87%	47.62%	56.20%	44.05%
toothbrush	54.90%	33.33%	54.90%	33.33%	54.90%	36.67%
transistor	53.99%	47.50%	58.47%	32.50%	77.64%	20.00%
wood	58.59%	40.00%	58.59%	40.00%	61.35%	33.33%
zipper	70.59%	72.27%	69.57%	61.34%	72.38%	36.13%
Mean	56.93%	**45.64%**	57.50%	43.40%	62.10%	34.87%

Table 2: Performance metrics for each object. Metrics abbreviations: ACC = Overall Accuracy, REC = Test Recall (Anomaly)

boundary between the two clusters minimizes the sum of the weights of the edges it cuts, while maximizing the connectedness of the split clusters. We score similar accuracy and a higher recall on the data, showing stronger anomaly detection performance.

Kmeans works by attempting to minimize the within-cluster sum of squares using Euclidean distance as a metric. It achieved moderate overall accuracy and recall, however it had inconsistent performance in distinguishing between anomalies in some of the objects in the dataset.

Lastly we evaluated using the Gaussian Mixture Model, which models the error distribution as a combination of two Gaussian distributions. This method resulted in slightly higher accuracy but failed to achieve the same recall performance.

Overall we chose to use Normalized Cut as it offered the highest anomaly recall, which we deemed the most import metric in assessing anomaly classification models.

4.6 Evaluating test time compute

As the number of diffusion timesteps increased, both accuracy and recall improved significantly before stabilizing around 50–100 timesteps. Early on, the model does not have enough denoising iterations to accurately estimate the noise component for each image, causing it to be unable to distinguish between anomaly and anomaly free images. As more timesteps are added, the diffusion process created a better estimate of the latent representation, and the noise estimation errors for regions containing anomalies became more pronounced. Having more reverse diffusion steps allowed for the model to better remove noise from images, improving the estimated noise gradient. This increased sensitivity to deviations in image space led to higher recall and increased overall accuracy. However, the marginal benefit of additional timesteps diminished after 50-100 timesteps due to the model already reaching a precise enough score estimate, causing performance to stagnate.

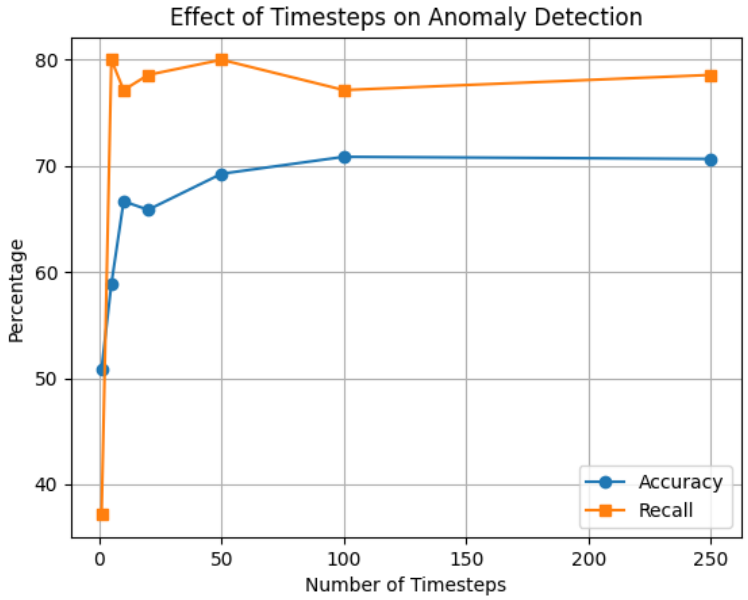

Figure 5: Usage of differing timesteps on the Hazelnut category from MVTec anomaly detection dataset to compare accuracy and recall rates

5 Conclusion and Discussion

We aimed to improve zero-shot anomaly detection using a latent diffusion model. We started with a dataset containing images of various objects, some of which had anomalies and some of which were defect-free. Our model effectively classified images by modeling the noise estimation errors through the latent diffusion process, and performing Normalized Cut to cluster the anomaly free and anomaly present images. We achieved a higher anomaly recall rate than One-Class SVM, showing that zero-shot generative methods can be applied to anomaly categorization.

Our method by removing the need for any labeled data demonstrates the learned representations of generative models and their adaptability in computer vision as a whole. Results of our method have future implications for the quality-control field, and our approach could lead to improved quality control processes. Specialists could potentially reduce the amount of time spent looking for a small amount of defective products to include in training data, thereby preventing overfitting and risk of poor generalizability. More broadly, potentially life-threatening defects can be identified and removed through the implementation of our model.

References

Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. Mvtec ad–a comprehensive real-world dataset for unsupervised anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9592–9600, 2019.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PmLR, 2020.

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

Amil Dravid, Yossi Gandelsman, Alexei A. Efros, and Assaf Shocher. Rosetta neurons: Mining the common units in a model zoo. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023. URL https://openaccess.thecvf.com/content/ICCV2023/papers/Dravid_Rosetta_Neurons_Mining_the_Common_Units_in_a_Model_Zoo_ICCV_2023_paper.pdf.

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

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9729–9738, 2020.

Kaiming He, Xiangyu Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16000–16009, 2022.

Geoffrey E Hinton. To recognize shapes, first learn to generate images. *Progress in brain research*, 165: 535–547, 2007.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, pp. 6840–6851, 2020.

Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. The platonic representation hypothesis. *arXiv preprint arXiv:2405.07987*, 2024. URL https://arxiv.org/abs/2405.07987.

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

Alexander C. Li, Mihir Prabhudesai, Shivam Duggal, Ellis Brown, and Deepak Pathak. Your diffusion model is secretly a zero-shot classifier. *arXiv preprint arXiv:2303.16203*, 2023. URL https://arxiv.org/abs/2303.16203.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

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

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

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in neural information processing systems*, 35:25278–25294, 2022.

Jianbo Shi and Jitendra Malik. Normalized cuts and image segmentation. *IEEE Transactions on pattern analysis and machine intelligence*, 22(8):888–905, 2000.

Jascha Sohl-Dickstein, Eric A Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *Proceedings of the 32nd International Conference on Machine Learning*, pp. 2256–2265. PMLR, 2015.

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

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

Pascal Vincent, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, Pierre-Antoine Manzagol, and Léon Bottou. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of machine learning research*, 11(12), 2010.

Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. Generating videos with scene dynamics. *Advances in neural information processing systems*, 29, 2016.

Roland S. Zimmermann, Yash Sharma, Steffen Schneider, Matthias Bethge, and Wieland Brendel. Contrastive learning inverts the data generating process. 139:12979–12990, 2021. URL `http://proceedings.mlr.press/v139/zimmermann21a.html`.