ADDiT: Anomaly Detection with Diffusion Transformers

June 1, 2024

Participant Information

- **Name(s):** Jirui Wu, Yujie Wu, Hongyi Zhao
- **Affiliation(s):** Xidian University
- **Contact Information:** zhaohongyii@163.com, yu1187442403@163.com, jiruiwu123@gmail.com

- **Track:** Track I - Adapt & Detect

Abstract

Reconstruction-based methods for Anomaly Detection (AD) have attained notable accomplishments. In recent years, several approaches have focused on using diffusion models to assess the performance of reconstruction-based methods for anomaly detection tasks. These methods typically extend the fundamental Denoising Diffusion Probabilistic Models (DDPM) in order to boost performance. This research seeks to explore the possibilities of transformer-based diffusion models in the field of anomaly detection. Our proposal entails the utilization of the Diffusion Transformer (DiT) model to construct an anomaly detection structure. This framework improves the fusion of produced and original images by conducting diffusion and sampling in the latent feature space, resulting in enhanced anomaly detection performance and resilience of the model. Diffusion models possess inherent characteristics that require sampling from distributions with additional noise. As a result, these models exhibit natural resistance to disturbances. This attribute renders them especially well-suited for managing diverse abnormalities in varied situations. To assess the efficacy of our proposed framework, we perform comprehensive experiments using the MVTec AD dataset, which is a widely accepted benchmark for anomaly identification. The experimental findings indicate that the suggested model attains a performance AUROC score of 92%, thereby showcasing its superiority in accurately and reliably detecting abnormalities. This study offers a viable avenue for future research by utilizing transformer-based diffusion models to improve anomaly detection tasks.

1 Introduction

Background

Anomaly detection (AD) pertains to the identification and segmentation of data that deviates from the norm. In industrial contexts, AD is extensively utilized to identify surface anomalies on products, garnering significant attention in recent years.(RPZ+22; DSLA21; YZW+21; BHK24; MBT23).The predominant unsupervised anomaly detection methodologies can be categorized into two groups: representation-based(RPZ+22; DSLA21; YZW+21) and reconstruction-based methods (BHK24; YCS+22; LCM+23; LSYP21). Representation-based methods depend on features extracted from pre-trained neural networks to define a similarity metric for nominal samples, employing a nearest-neighbor strategy to address the problem. Generative model-based methods, which do not require additional data, are versatile across various scenarios. Typically, these methods(CYLL18; YCS+22; GLL+19; SWL24) employ autoencoder-based networks (AEs). The underlying assumption is that after the encoder compresses the input image into a low-dimensional representation, the decoder reconstructs the anomalous regions as normal. Nevertheless, the AE-based approach has inherent limitations: (i) it may lead to an

invariant reconstruction of abnormal regions because the compressed low-dimensional representation still contains anomalous information, resulting in false negative detections; (ii) AEs might produce a coarse reconstruction of normal regions due to limited restoration capability, thereby introducing numerous false positives, particularly in datasets with complex structures or textures. Recently, diffusion models(HJA20; RBL+22) have emerged as prominent deep generative models. This study revisits the reconstruction-based anomaly detection framework, leveraging the potential of diffusion models to achieve remarkable reconstruction of anomalous images.

Challenge Description

The main challenge posed by the track we participate in is to ensure the robustness of the model in real-world scenarios when facing interference from a variety of external factors, such as camera angles, lighting conditions, noise, and other factors.

2 Methodology

2.1 Model Design

This research examines using the Diffusion Transformer (DiT) model(PX23) for anomaly detection tasks. The diffusion approach improves the integration of prospective abnormal images with original images by introducing noise and resampling. In addition, we carry out forward diffusion and reverse denoising processes within the latent space, which enhances the model's capacity for reconstruction and makes it more resistant to intricate disturbances.

The structure of our model comprises two primary aspects: a reconstruction network based on DiT and a detection network performing at the feature level. To minimize the potential of the model reproducing anomalies, we utilize a substantial amount of normal training data to train the reconstruction network. While sampling test pictures, we do not begin with Gaussian noise. Instead, we sample from an intermediate step of the diffusion process to ensure image reconstruction accuracy.Regarding detection, we notice that relying just on image-level detection produces unsatisfactory outcomes. Thus, we develop a feature extraction network employing EfficientNet(TL19) and implement a multi-level fusion technique to generate anomaly score maps, thereby accomplishing the detection and segmentation of anomalies.

Reconstruction network. We implement the reconstruction network via DiT model, which reformulates the reconstruction process as a noise-to-norm paradigm. First, we obtain a latent representation z of the input image x based on a learned encoder E. Following the patchfiy operation in ViT(), the latent representation z is transformed into a sequence of n tokens by linearly embedding each patch in the input. The number of tokens n created by patchify is determined by the patch size hyperparameter p. In the forward noising process, we gradually add noise to input data z_0 at a random time step t. By applying the reparameterization trick, we can sample:

$$z_t = z_0\sqrt{\overline{\alpha}_t} + \epsilon_t\sqrt{1-\overline{\alpha}_t}, \qquad \epsilon_t \sim N(0, I) \tag{1}$$

The input data z_0 gradually loses its discriminative features and approaches an isotropic Gaussian distribution as the time step increases. During training, DiT replaces the U-Net of the diffusion model with an elaborate Transformer architecture to predict ϵ by minimizing the training objective:

$$L := \mathbb{E}_{\mathcal{E},\epsilon\sim N(0,1),t}\left[||\epsilon - \epsilon_\theta(z_t, t)||_2^2\right] \tag{2}$$

At the end of DiT blocks, our series of image tokens have to be decoded into an output diagonal covariance prediction and an output noise prediction. The form of these two outputs is the same as that of the original latent input. We use a standard linear decoder to do this.

Feature alignment network. We begin by employing a pre-trained feature extractor to compare the test image with its reconstructed counterpart at the feature level. While the feature extractor is pre-trained on the ImageNet dataset, it fails to effectively adapt to the domain-specific nuances required for anomaly detection tasks. To overcome this limitation, we propose an unsupervised domain adaptation method that aligns various extracted layers from two nearly identical images. This alignment strategy allows the network to disregard minor discrepancies introduced during the reconstruction process, thereby improving its capacity to learn and generalize within the specific domain of the task. More

specifically, given a target image y and its reconstructed image $\hat{y}$ sampled by DiT, we can assume that their features are approximate, and therefore, we use the following optimization method to fine-tune the network ϕ:

$$L = L_{Similarity}(y, \hat{y})$$

This approach ensures that the domain adaptation network effectively adapts to the specific characteristics of the domain, enhancing the overall performance of the anomaly detection system.

Anomaly Detection with Reconstruction Our solution relies on a pre-trained EfficientNet model on ImageNet to extract features from the reconstructed and original images. We document the output of each component of EfficientNet and combine all the results. This fusion technique leverages the benefits of both shallow and deep features, resulting in enhanced resilience against an extensive variety of anomalies. Afterward, we measure the similarity between feature maps by calculating the Euclidean distance. This allows us to generate an anomaly score map of the original image size by upsampling. After obtaining the test image feature map f and reconstruction feature map $\hat{f}$ through the aforementioned network, the pixel-level anomaly score s_i is calculated as follows:

$$s_i = ||f_i - \hat{f}_i||^2, \quad i \in \mathcal{N} \tag{3}$$

where i denotes the i-th pixel of the input test image. For image-level classification, we use the maximum distance score s^* among all the pixels to represent the image-level score as:

$$s^* = max\{||f_i - \hat{f}_i||^2 | i \in \mathcal{N}\} \tag{4}$$

2.2 Dataset & Evaluation

Dataset. We assessed our model's performance using MVTec AD dataset, which serves as a real-world AD benchmark. MVTec AD dataset includes 5354 images in 15 categories. 3629 of these images are defect-free and the remaining 1725 have defects. Each category has an average of five different types of defects. The image resolution ranges from 700×700 to 1024×1024.

Evaluation. To quantify the model performance of image-level anomaly detection, we employ the Area Under the Receiver Operator Curve (AUROC) and the F1-score at the optimal threshold (F1-max) as our evaluation metrics, being consistent with the previous works(). Furthermore, we utilize pixel-level AUROC and F1-max to measure the defect localization performance.

2.3 Implementation details

The input image size of MVTec-AD is 256× 256× 3, after being fed into the pre-trained VAE, the feature maps become 32 × 32 × 4, namely, the patch size is 8. We use Adam with weight decay 0 for optimization. Our model is trained for 1400 epochs on 1 GPUs (NVIDIA GeForce RTX 3090 24GB) with batch size 4. The learning rate is 1×10^{-4}. and Our DiT model is trained from scratch.

3 Results

3.1 Performance Metrics

4 Discussion

4.1 Challenges & Solutions

- Challenges: The main challenges in applying the diffusion model to anomaly detection are as follows: The sampling process of diffusion typically starts from standard noise and progressively denoises to restore an image within the training image domain. In other words, the sampling process of diffusion is an image sampled from the distribution of the training image, which has not only correctness but also diversity. However, for anomaly detection tasks, we are more

	Image AUROC	Image F1-max	Pixel AUROC	Pixel F1-max
bottle	99.52	97.67	97.53	67.11
cable	89.06	86.73	94.78	43.04
capsule	80.30	93.04	97.93	38.47
carpet	100	100	98.63	54.88
grid	93.73	94.74	96.6	35.79
hazelnut	99.50	97.87	98.09	53.12
leather	100	100	99.09	40.32
metal_nut	98.68	97.30	95.15	70.36
pill	84.37	92.47	93.98	45.71
screw	84.07	90.48	98.07	28.38
tile	95.85	93.79	89.32	44.60
toobrush	88.33	92.31	98.58	49.30
transistor	96.88	92.50	97.9	69.96
wood	96.67	94.31	93.31	44.86
zipper	76.34	93.28	95.63	39.31
average	92.22	94.43	96.31	48.35

concerned with the information in the specific input image at hand. We focus on the uniqueness of a single image and expect that the reconstructed image by the diffusion model retains most of the information from the input image, including the background and other non-anomalous details. If the reconstructed image does not match the input image in these aspects, it may be mistakenly identified as abnormal.

- Solutions: To address the challenges mentioned above, we modify the reverse process of the diffusion model. First, when a test image is input, we add noise to it to obtain X_t, and then denoise it step by step from X_t to obtain the reconstructed image, where $t << T$. This approach helps X_t retain most of the original image information. Simultaneously, by adding some noise, the anomalous parts are partially blurred. Since the diffusion model is trained on normal images, during the reverse diffusion process, the anomalous parts can be reconstructed as normal, which is the desired outcome. Finally, by comparing the reconstructed image with the input image, anomalies can be successfully identified, thus completing the anomaly detection task.

4.2 Model Robustness & Adaptability

- Robustness:The diffusion model is unique in that it adds noise during the forward process and denoises during the reverse process, giving it strong robustness to random disturbances. During the testing phase, random disturbances are added to the test data to simulate domain shifts, aligning with the denoising scenario of the diffusion model. Therefore, DiT can better adapt to uncertainties caused by random disturbances.

- Adaptability: Different from the original diffusion model, we do not start from standardized noise in the inference process, but first add noise to the test data, and then de-noise the noisy data step by step to obtain a reconstructed image, which makes the model have the adaptability to different input data, and retain its background or unique image information as much as possible.

4.3 Future Work

In the future, we will add more conditional guidance mechanisms on the basis of DiT model to control the relationship between the generated image and the input image. At the same time, DiT will be used in One for all scenarios, that is, only one model is trained for all categories, which will face more problems and challenges.

5 Conclusion

In this paper, We explore the applicability of diffusion models based on Transformer architecture for anomaly detection. The experimental results show that there is still room to improve the performance of applying DiT directly on AD tasks. Therefore, we introduce a feature adaptor for further fine-tuning of the sampled representations and finally reach a more satisfactory experimental result.

References

[BHK24] Kilian Batzner, Lars Heckler, and Rebecca König. Efficientad: Accurate visual anomaly detection at millisecond-level latencies. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 128–138, 2024.

[CYLL18] Zhaomin Chen, Chai Kiat Yeo, Bu Sung Lee, and Chiew Tong Lau. Autoencoder-based network anomaly detection. In *2018 Wireless telecommunications symposium (WTS)*, pages 1–5. IEEE, 2018.

[DSLA21] Thomas Defard, Aleksandr Setkov, Angelique Loesch, and Romaric Audigier. Padim: a patch distribution modeling framework for anomaly detection and localization. In *International Conference on Pattern Recognition*, pages 475–489. Springer, 2021.

[GLL+19] Dong Gong, Lingqiao Liu, Vuong Le, Budhaditya Saha, Moussa Reda Mansour, Svetha Venkatesh, and Anton van den Hengel. Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1705–1714, 2019.

[HJA20] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[LCM+23] Wenrui Liu, Hong Chang, Bingpeng Ma, Shiguang Shan, and Xilin Chen. Diversity-measurable anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12147–12156, 2023.

[LSYP21] Chun-Liang Li, Kihyuk Sohn, Jinsung Yoon, and Tomas Pfister. Cutpaste: Self-supervised learning for anomaly detection and localization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9664–9674, 2021.

[MBT23] Arian Mousakhan, Thomas Brox, and Jawad Tayyub. Anomaly detection with conditioned denoising diffusion models. *arXiv preprint arXiv:2305.15956*, 2023.

[PX23] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4195–4205, 2023.

[RBL+22] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

[RPZ+22] Karsten Roth, Latha Pemula, Joaquin Zepeda, Bernhard Schölkopf, Thomas Brox, and Peter Gehler. Towards total recall in industrial anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14318–14328, 2022.

[SWL24] Zhongju Sun, Jian Wang, and Yakun Li. Ramfae: a novel unsupervised visual anomaly detection method based on autoencoder. *International Journal of Machine Learning and Cybernetics*, 15(2):355–369, 2024.

[TL19] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019.

[YCS+22] Zhiyuan You, Lei Cui, Yujun Shen, Kai Yang, Xin Lu, Yu Zheng, and Xinyi Le. A unified model for multi-class anomaly detection. *Advances in Neural Information Processing Systems*, 35:4571–4584, 2022.

[YZW+21] Jiawei Yu, Ye Zheng, Xiang Wang, Wei Li, Yushuang Wu, Rui Zhao, and Liwei Wu. Fastflow: Unsupervised anomaly detection and localization via 2d normalizing flows. *arXiv preprint arXiv:2111.07677*, 2021.