MSD: Mask-Guided and Semantic-Guided Diffusion-Based Framework for Stone Surface Defect Detection

Longtao Chen*
Engineering College, Huaqiao University, Quanzhou, Fujian 362021, China
No.269 Chenghua North Rd. Quanzhou, Fujian, China
longtaochen@hqu.edu.cn

Jinjie Zheng
Engineering College, Huaqiao University, Quanzhou, Fujian 362021, China
No.269 Chenghua North Rd. Quanzhou, Fujian, China

Fenglei Xu
Suzhou University of Science and Technology, Suzhou, Jiangsu 215000, China
No.99 Xuefu Road. Suzhou, Jiangsu, China

Jing Lou
Changzhou Vocational Institute of Mechatronic Technology, Changzhou, Jiangsu 213000, China
26 Mingxin Middle Road, Wujin District, Changzhou, Jiangsu, China

Huanqiang Zeng
Engineering College, Huaqiao University, Quanzhou, Fujian 362021, China
Quanzhou Digital Institute, Quanzhou, Fujian 362021, China
No.269 Chenghua North Rd. Quanzhou, Fujian, China

Abstract

Few-Shot Anomaly Detection (FSAD) has a wide range of applications in industrial anomaly monitoring. However, when confronted with the specific challenge of Stone Surface Defect Detection, traditional FSAD methods exhibit limitations, particularly in reconstruction fidelity and segmentation accuracy. Moreover, the lack of high-quality datasets on stone surface defects poses a significant challenge for research in this area. To address these issues, we propose the Mask-Guided and Semantic-Guided Diffusion-based framework (MSD) for stone surface defect detection and introduce Stone Surface Defect Dataset (StoneDD), a few-shot dataset designed specifically for vision-based defect detection and segmentation. Our framework integrates a pixel-space with feature-space and latent-space, further enhanced by a Mask-Guided Knowledge Distillation network (MGKD) and a Semantic-Guided Enhancement Network (SGEN). The MGKD network concentrates on anomalous regions, improving the accuracy of image reconstruction, while the SGEN network skillfully reconstructs these areas while preserving semantic accuracy. By utilizing multi-scale feature information, it ensures precise localization and classification, substantially reducing the likelihood of false detections. Extensive experimental results show that MSD achieves superior performance, underlining its potential for future deployment in industrial quality inspection applications.

Keywords: ***Few-shot anomaly detection, Diffusion-based framework, Stone surface defect detection***

1. Introduction

Stone panels are widely utilized in various settings, including architectural cladding and interior decoration. However, many stone manufacturers still rely on inefficient manual inspection methods, making stone surface defect detection a persistent challenge. Compared with traditional vision detection, stone surface defect detection has the following characteristics: 1) Due to the intricate textures and complex defects inherent in the stone surface, accurately identifying surface anomalies poses a significant challenge. 2) Theoretically, the variety of stone surface imperfections

*Corresponding author

tends to be infinite. 3) The limited number of erroneous stone layouts and the scarcity of surface defects made it challenging to collect the dataset.

Currently, most stone manufacturers continue to rely on manual inspection for detecting surface defects in stone. As an alternative to this inefficient and labor-intensive approach, numerous model-driven conventional methods have been proposed for more effective detection. Several studies [6, 30] have explored synthesis-based and embedding-based methods for detecting surface anomalies. However, these methods encounter substantial hurdles in stone detection, including the management of multiple negative sample categories and the long-tail phenomenon [32]. Diffusion-based methods [28, 10, 31] can largely mitigate these issues due to their capacity to reconstruct images and model complex distributions effectively.

Diffusion-based methods can effectively complete the task of stone surface anomaly detection. A Diffusion-based framework for multi-class Anomaly Detection (DiAD) [10] utilizes a diffusion model to synthesize anomalies and fine-tunes a pre-trained feature extractor for multi-class anomaly detection. Similarly, Denoising Diffusion Anomaly Detection (DDAD) [19] employs score-based diffusion models to generate normal samples and enhances domain transfer with pre-trained feature extractors. However, in stone surface defect detection, diffusion-based methods may misclassify stone background textures as defects during denoising, resulting in higher false detection rates and suboptimal reconstruction outcomes just as shown in Fig. 1. To tackle the mentioned challenges, we introduce a Mask-Guided and Semantic-Guided Diffusion-based framework(MSD) for stone surface defect detection.

Conducting research in the area of stone surface defect detection is challenging due to the lack of publicly available datasets. To address this gap, we introduce a novel dataset called StoneDD, specifically designed for stone surface defect detection. StoneDD presents a range of potential challenges in this field and is the first publicly available dataset of its kind, as illustrated in Fig. 2. This dataset encompasses the following key properties:

- **Defect types:** spot, stain, bubble, gap, chromatic-aberration, and oil-paper (paper sticks to the stone).
- **Data volume:** 1,290 images with corresponding masks.
- **Multi-resolution:** images available in both 900x900 and 256x256 resolutions to support comprehensive experimentation
- **Diversity:** features richly textured stone slabs with subtle variations in local color, presenting realistic challenges in distinguishing natural textures from actual defects

Our key contributions can be summarized as follows:

- We develop the MGKD network that effectively differentiates between anomalous regions and background areas through weighted allocation, leading to enhanced reconstruction precision.
- We introduce the SGEN network, designed to integrate multi-scale features and improve both anomaly detection and localization accuracy, effectively reducing false detection rates.
- We present a few-shot Stone Surface Defect dataset, and demonstrate the effectiveness of our MSD framework through extensive experimentation, achieving outstanding localization and detection scores of 95.8/67.3 and 99.1/97.6 (AUROC/AP), respectively.

2. Related Work

2.1. Stone Surface Detection

Stone surface detection is an essential task across various industries, including construction, archaeology, and quality control in stone manufacturing. Over the years, numerous methodologies [14, 15, 2, 9, 13, 23, 19]have been proposed to enhance the precision and efficiency of detecting anomalies and features on stone surfaces.

Traditional approaches often relied on manual inspection, a process that is not only time-consuming but also susceptible to human error. With advancements in computer vision and machine learning, automated detection techniques have increasingly taken center stage. For example, P. Kapsalas et al. [14] utilized optical detection techniques to quantify surface decay in stone materials, while J. Lee et al. [15] explored robust and efficient automatic methods for detecting tool defect in polished stone. Additionally, A. Borel et al. [2] discussed optimization methods for wear detection and characterization on stone tool surfaces.

Recently, many studies have begun to use deep learning methods to detect surface defects in stone surface. M. Guerrieri et al. [9] used deep learning with inexpensive detection equipment to identify and measure damage on flexible and rocky road surfaces. H. Kabir et al. [13] employed Mask R-CNN for stone detection and segmentation in underground pipeline inspection robots. M. Smith et al. [23] discussed using machine vision technology for inspecting polished stone materials in manufacturing processes. Additionally, A. Mousakhan et al. [19] utilized a score-based pre-trained diffusion model to generate normal samples and fine-tuned feature extractors for domain transfer, enhancing detection across various stone surfaces.

Despite these advancements, challenges remain due to the variability and complexity of stone surface textures. Future research should focus on developing more robust

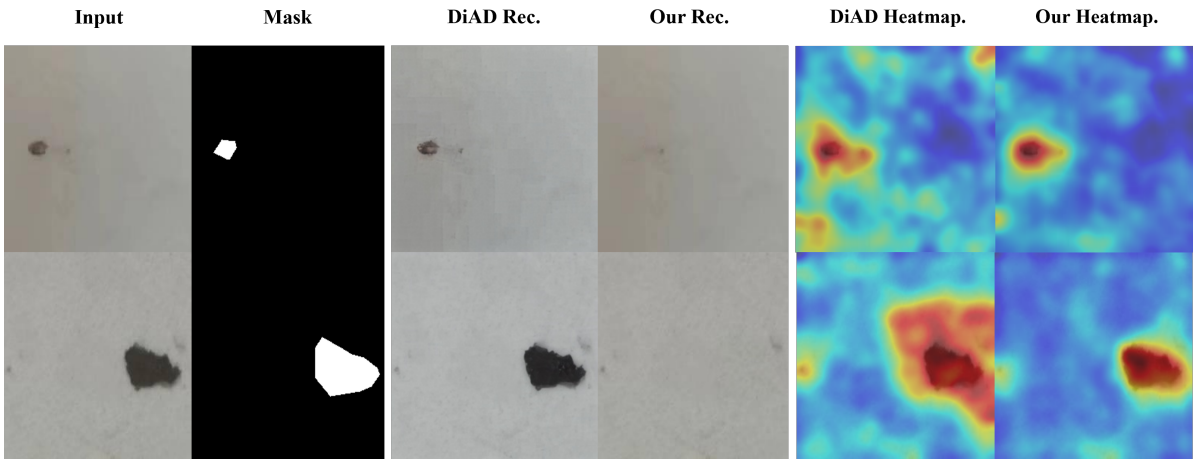

Figure 1: Comparison of results of different methods. (a) DiAD exhibits poor reconstruction quality and insufficient heatmap convergence. (b) Our proposed approach demonstrates superior defect detection capability and improved image reconstruction quality.

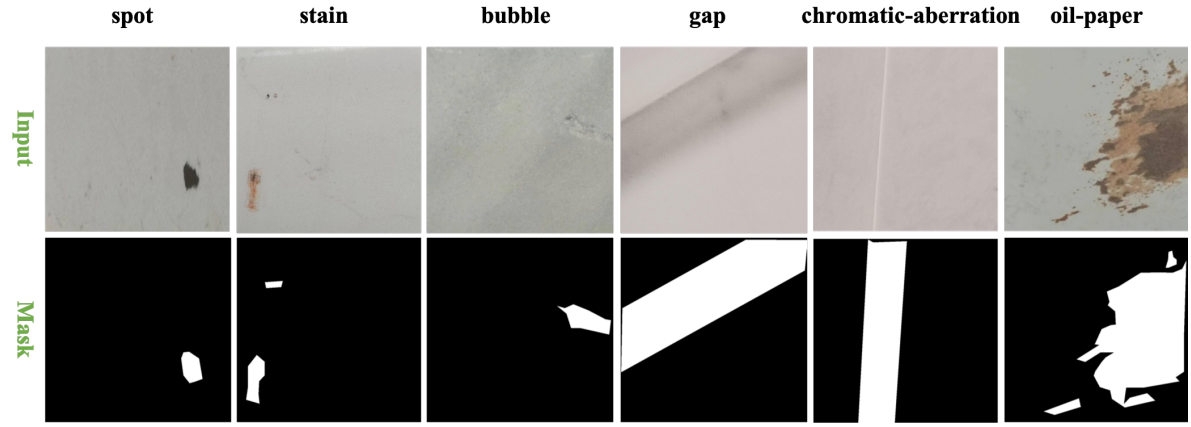

Figure 2: Samples from StoneDD.

and adaptable models, which can use weakly supervised anomaly detection methods to detect various abnormal defects on stone surfaces.

2.2. Anomaly Detection

Traditional stone surface detection methods often struggle with issues such as out-of-distribution detection for unseen samples during training. Anomaly detection techniques, however, can mitigate these challenges. Anomaly detection [8, 16, 24] can be categorized into three primary approaches:

1) **Synthesis-based methods** generate anomalies from normal image samples. During training, both normal and artificially generated abnormal images are fed into the network, helping in anomaly detection and localization. Zavrtanik et al. [30] proposed the Discriminatively Trained Reconstruction Anomaly Embedding Model (DRAEM), an end-to-end network combining a reconstruction component with a discriminative sub-network for synthesizing and generating out-of-distribution anomalies. However, synthesizing all possible anomaly variations remains challenging due to the diverse and unpredictable nature of anomalies in real-world scenarios.

2) **Embedding-based methods** map the original image's three-dimensional information into a high-dimensional feature space [21]. Several approaches [11, 25, 26, 17] utilize networks pre-trained on ImageNet [7] for feature extraction. Deng et al. [6] proposed the Reverse Distillation paradigm For Anomaly Detection (RD4AD), using Wide Residual Networks (WideResNet50) as a teacher model for feature extraction and a reverse network as a student model to compute anomaly scores based on cosine similarity. Defard et al. [5] introduced Patch Distribution Modeling (PaDiM), leveraging pretrained CNNs for patch embedding and mul-

tivariate Gaussian distributions to probabilistically represent the normal class.

You et al. [29] developed the Unified model for multi-class Anomaly Detection (UniAD), which enhances reconstruction networks with layer-wise query decoders, neighbor-masked attention modules, and feature jittering. However, discrepancies between industrial images and ImageNet's data distribution may limit these features' applicability for industrial anomaly detection [3, 22, 27].

3) **Diffusion-based methods** train models on anomaly-free data to identify patterns in normal data. The diffusion model [28] has garnered significant attention due to its impressive reconstruction capabilities. It has shown exceptional performance in various tasks, including image generation [31], video generation [12], object detection [4], and image segmentation [1]. Blattmann et al. [20] proposed High-Resolution Image Synthesis with Latent Diffusion Models (LDM), which introduces conditioning via cross-attention to control the generation process. However, accurately preserving the original semantic content in reconstructed images remains a challenge. Wyatt et al. [28] introduced Anomaly Detection with Denoising Diffusion Probabilistic Models (AnoDDPM), the first application of a diffusion model for medical anomaly detection. He et al. [10] developed a Diffusion-based framework for multi-class Anomaly Detection (DiAD), utilizing a diffusion model to generate synthetic anomalies while fine-tuning a pre-trained feature extractor for improved detection across various anomaly classes.

However, the reconstruction results of these methods often exhibit limitations, such as susceptibility to background noise, which can lead to normal features being misclassified as anomalies. To overcome these issues, this paper proposes a mask-guided and semantic-guided diffusion-based framework for stone surface anomaly detection, improving the quality of reconstructed images and enhancing the pixel-level localization of anomalies.

3. Method

During training (S_1, $\{P_1, P_2, P_3\}$, S_2, S_3). S_1: Encoding the input image x as the latent-space representation; $\{P_1, P_2, P_3\}$: $\{P_1$: Forward diffusion, adding noise to the latent-space representation; P_2: Enter the latent-space representation into SGEN for a more definitive location of the defect edge P_3: The latent-space representation is overlaid with masks and then into MGKD for precise defect ranges$\}$; S_2: Integrating Mask-Guided and Semantic-Guided representation to facilitate the reverse denoising process and obtain reconstructed representation; S_3: Decoding the reconstructed representation as the reconstructed image $\hat{x_0}$ (i.e. flawless image/ Repaired image).

During testing (S_4). x and $\hat{x_0}$ are inputted into the same pre-trained feature extraction network to obtain feature maps $\{f_1, f_2, f_3\}$ of different scales, and calculate their anomaly scores S by cosine similarity.

How it work. When the input image passes through the latent space on the right, a defect-free reconstructed image is generated. Both the input and the reconstructed images are then passed to the left to compute the cosine distance. The greater the pixel-level difference, the more pronounced the corresponding anomaly in the heatmap. In simple terms, by comparing the differences between the two images, the regions of discrepancy are displayed in red. Therefore, the accuracy of the heatmap largely depends on the quality and level of detail in the reconstructed image. To address this issue, we employed semantic and mask guidance to optimize the reconstruction process, resulting in more precise reconstructions and achieving heatmap convergence.

The proposed MSD pipeline, depicted in Fig. 3, comprises three components: 1) Feature Space, 2) Pixel Space, 3) Latent Space.

3.1. Feature Space

For defect localization and detection, we utilize the same pre-trained ResNet50 feature extraction network Ψ to extract features from both the input image x and the reconstructed image $\hat{x_0}$. We then calculate the anomaly map on different scale feature maps M_n using cosine similarity:

$$\mathcal{M}^n\left(x_0, \hat{x_0}\right) = 1 - \frac{\left(\Psi^n\left(x_0, \hat{x_0}\right)\right)^T \cdot \Psi^n\left(x_0, \hat{x_0}\right)}{\left\|\Psi^n\left(x_0, \hat{x_0}\right)\right\| \left\|\Psi^n\left(x_0, \hat{x_0}\right)\right\|}, \quad (1)$$

Where n denotes the n-th feature layer f_n, and the anomaly score S for an input-pair of anomaly localization is:

$$\mathcal{S} = \sum_{n \in N} \sigma_n \mathcal{M}^n\left(x_0, \hat{x_0}\right), \quad (2)$$

Where σ_n denotes the upsampling factor to maintain the original dimension of the pixel space image, and N indicates the number of feature layers used during inference.

3.2. Pixel Space

The pixel space autoencoder $\{E, D\}$: an encoder, which maps the input image to the latent-space, and a decoder, which reconstructs the latent-space representation back into an image. These autoencoder module is tasked with learning low-dimensional representations of data. It achieves this by receiving the input image and encoding them as the latent-space representation within the latent-space.

3.3. Latent Space

The latent-space primarily encompasses three components: latent diffusion model (LDM), Mask-Guided Knowledge Distillation network (MGKD), and Semantic-Guided

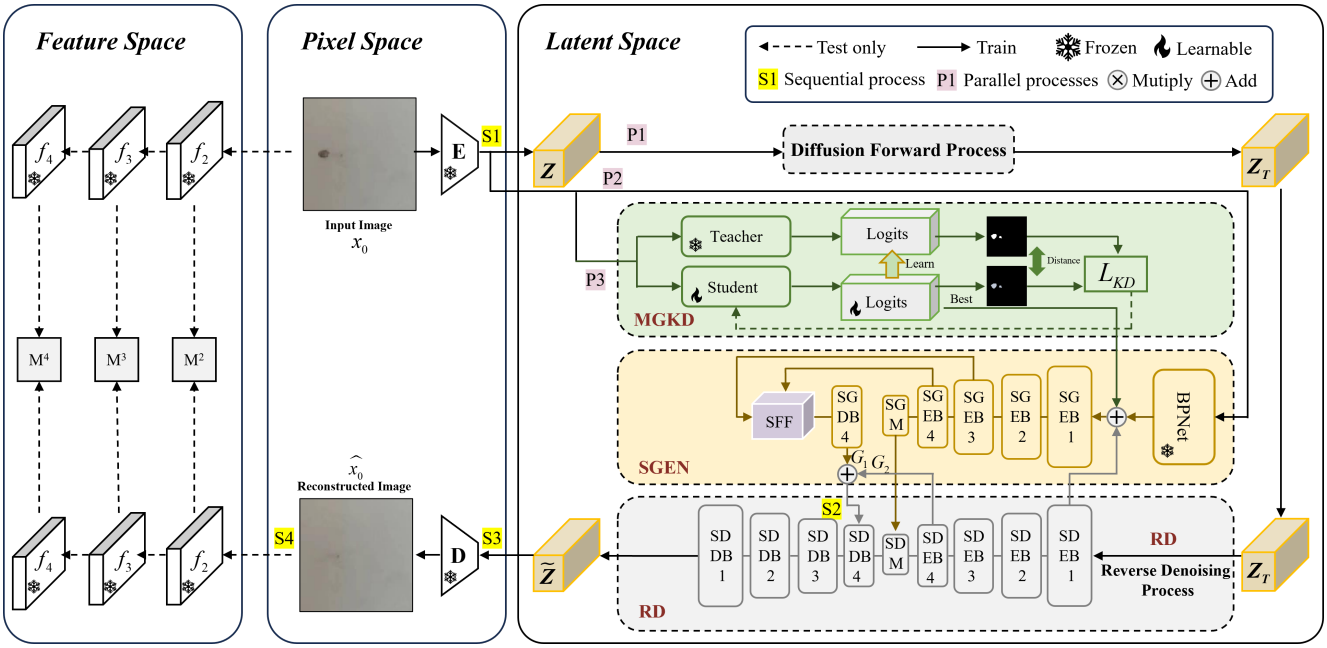

Figure 3: Framework of the proposed MSD. During training (S_1, $\{P_1, P_2, P_3\}$, S_2, S_3), the input image x_0 is processed in parallel by $\{P_1, P_2, P_3\}$, Then reverse denoising Process RD to get the reconstructed image $\hat{x}_0$. During testing (S_4), x_0 and $\hat{x}_0$ are inputted into the feature space to generate feature maps and calculate anomaly scores S.

Enhancement Network (SGEN). We will elaborate on each of these three modules below:

3.3.1 Latent Diffusion Model

Latent Diffusion Model (LDM) focuses on the low-dimensional latent space with conditioning mechanisms. The network compresses images using an encoder, conducts diffusion and denoising operations in the latent representation space, and subsequently reconstructs the images back to the original pixel space using a decoder. The training optimization objective is:

$$\mathcal{L}_{LDM} = \mathbb{E}_{z_0,t,c,\epsilon\sim\mathcal{N}(0,1)}\left[\|\epsilon - \epsilon_\theta\left(Z_t, t, c\right)\|_2^2\right], \quad (3)$$

where c represents the conditioning mechanisms which can consist of multimodal types such as text or image, connected to the model through a cross-attention mechanism. Z_t represents the full-noise representation,

The LDM [28] comprises two processes: diffusion forward process and reverse denoising process, just as Fig. 4.

In the diffusion forward process, a noisy sample x_t is generated through a Markov chain that gradually introduces Gaussian-distributed noise to an initial data sample x_0 as Eq. (4).

$$x_t = \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon_t, \epsilon_t \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \quad (4)$$

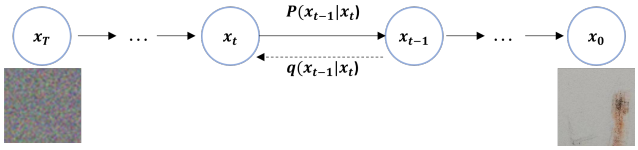

Figure 4: Denoising diffusion probabilistic model.

Where $\alpha_t = 1 - \beta_t$, $\bar{\alpha}_t = \prod_{i=1}^{T}\alpha_i = \prod_{i=1}^{T}(1-\beta_i)$, and β_i represents the noise schedule regulating the amount of noise added at each timestep.

In the reverse denoising process, x_t is sampled from Equation 1, and x_{t-1} is reconstructed using x_t and the model prediction $\epsilon_\theta(x_t, t)$ as Eq. (5).

$$x_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(x_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta\left(x_t, t\right)\right) + \sigma_t z, \quad (5)$$

Where $z \sim \mathcal{N}(0, I)$, σ_t is a fixed constant related to the variance schedule, and θ represents the learnable parameters.

3.3.2 Mask-Guided Knowledge Distillation network

Mask-Guided Knowledge Distillation network (MGKD) focus on more precise and larger possible areas of defects. Specifically, the latent-space representation and grid are overlaid by dot product to obtain the mask representation,

then the mask teacher model and student model are passed in synchronously, the mask weights are used as a benchmark to correct the student model, simultaneously the student model is allowed to learn from the teacher through L_{KD}, as Eq. (1). The gap between them is continuously reduced in n rounds of iteration, and finally the best weights optimize the Semantic-Guided network, which assists in the reverse denoising process so that it is bound to be effective in the flawed region search. The network consists of a pre-trained teacher encoder *E*, sourced from sam-vit-base Automatic-Mask-Generation, and a trainable student decoder *D*. During training, the student decoder learns to mimic the teacher encoder's behavior by minimizing the similarity loss L_{KD}. The best logits are generated based on this learning process, which are then used to further optimize the SGEN model.

$$\mathcal{L}_{\text{MSE}} = \text{Dist}(z_S, z_T) == \frac{1}{n}\sum_{i=1}^{n}\left(z_T^{(i)} - z_S^{(i)}\right)^2 \tag{6}$$

Where Dist$(\cdot)$ denotes the Euclidean distance and n denotes the number of categories, Z_T and Z_s denote the outputs of the mask teacher model and student model;

$$\mathcal{L}_{\text{KL}} = \frac{1}{n}\sum_{i=1}^{n}\sigma(z_T^{(i)}/\tau)\cdot\log\left(\frac{\sigma(z_T^{(i)}/\tau)}{\sigma(z_S^{(i)}/\tau)}\right) \tag{7}$$

where L_{KL} denotes the Kullback-Leibler scattering loss function, and σ denotes the Softmax function, τ denotes the temperature parameter, which is used to adjust the degree of "softening" of Softmax;

$$\mathcal{L}_{\text{KD}} = \alpha\cdot\mathcal{L}_{\text{MSE}} + \beta\cdot\mathcal{L}_{\text{KL}} \tag{8}$$

where L_{KD} denotes the overall loss function of the knowledge distillation network, α and β denote the weighting coefficients, which can be manually adjusted to achieve the optimal distillation effect.

3.3.3 Semantic-Guided Enhancement Network

Semantic-Guided Enhancement Network (SGEN) focuses on accurately locating the edges and contours of defects. Specifically, the pre-trained BPNet first extracts multi-level features and integrates them to obtain semantic representations. Through a series of semantic modules that progressively refine the defective edges, the Spatial-aware Feature Fusion (SFF) block then combines the fused features with the original features, which collaboratively optimizes the reverse denoising process and this will surely be able to find the defective edges for subsequent heat map visualization.

BPNet combines bottom-up and top-down modules to enhance low-resolution feature maps by fusing them with high-resolution ones, thereby enriching semantic information. SGEN reconstructs anomaly areas while retaining the original image's semantics, leveraging multi-scale features for precise localization and classification. The model focuses on identifying anomalies while minimizing attention to irrelevant background elements.

SGEN is designed to tackle the specific challenges faced by LDMs in multi-class defect detection tasks. To address the limitations associated with LDMs, particularly in effectively reconstructing anomalies while retaining the semantic information of the input image, we introduce the SGEN as a solution to enhance performance in multi-class scenarios.

Given an input image $x_0 \in \mathbb{R}^{3\times H\times W}$ in pixel space, the pre-trained encoder $\mathcal{E}$ encodes x_0 into a latent-space representation $z \in \mathbb{R}^{c\times h\times w}$, where $z = \mathcal{E}(x_0)$. Now, the forward diffusion process can be characterized as follows: $z = \mathcal{E}(x_0)$. Now, the forward diffusion process can be characterized as follows:

$$z_t = \sqrt{\bar{\alpha}_t}z_0 + \sqrt{1-\bar{\alpha}_t}\epsilon_t, \epsilon_t \sim \mathcal{N}(\mathbf{0}, \mathbf{I}). \tag{9}$$

The perturbed representation z_T and input X are simultaneously fed into the RD and SGEN network, respectively. After T steps of the reverse denoising process, the final variable $\hat{z}$ is restored to the reconstructed image $\hat{X}$ from the pre-trained decoder $\mathcal{D}$ giving $\hat{X} = \mathcal{D}(\hat{z})$. The training objective of MSD is:

$$\mathcal{L}_{MSD} = \mathbb{E}_{z_0,t,c_i,\epsilon\sim\mathcal{N}(0,1)}\left[\|\epsilon - \epsilon_\theta(z_t, t, c_i)\|_2^2\right] + \lambda L_{\text{KD}}. \tag{10}$$

4. Experiment

4.1. Stone Defect Dataset

In the field of stone surface defect detection, to the best of our knowledge, we are the first team to propose the few-shot Stone Defect Dataset (StoneDD). We hope that the StoneDD will provide effective support for subsequent researchers to advance the stone surface defect detection technology, thus helping the stone industry to move away from traditional manual inspection methods to automated and intelligent defect detection. We aim to provide a small-scale muti-resolution Stone Defect Dataset (StoneDD) with comprehensive annotations that can expose the challenges of stone surface defect detection and segmentation. The specific process of dataset construction is shown in the Fig. 5.

Raw Data Collection. We manually photographed and collected original stone images and selectively collected complete, high-resolution images of the stone surface. Because these images contained many types of defects that were difficult to recognize, we manually selected the six most obvious and distinguishable types of defects, resulting

Table 1: Comparison with other detection methods on StoneDD dataset.

Metrics	Non-Diffusion Methods				Diffusion-based Methods		
	PaDiM 21'ICPR	DRAEM 21'ICCV	RD4AD 22'CVPR	UniAD 22'NeurIPS	AnoDDPM 22'CVPR	DiAD 24'AAAI	**MSD Ours**
AUROCseg	88.7	91.5	93.7	94.8	89.3	93.8	**95.8**
APseg	-	18.3	23.9	43.1	47.1	47.0	**56.8**
F1maxseg	-	17.3	26.3	49.5	11.2	**55.1**	49.7
AUROCcls	86.9	60.9	79.1	96.5	67.9	98.7	**99.1**
APcls	-	84.6	95.2	97.9	78.1	**99.5**	97.6
F1maxcls	-	88.7	90.3	95.7	88.7	99.2	**99.2**

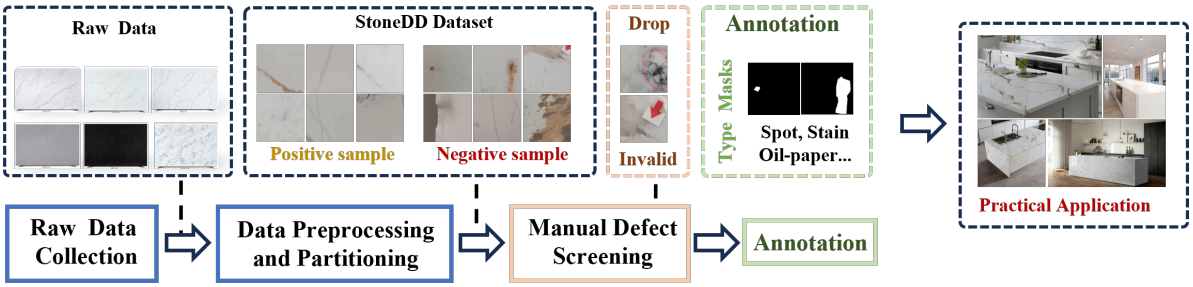

Figure 5: The generation pipeline of StoneDD, which includes raw data collection, data preprocessing and partitioning, manual defect screening, and annotation.

Table 2: Ablation studies of different feature extraction backbones.

Metrics	VGG19	ResNet50	WideResNet101
AUROCseg	92.5	95.8	**96.4**
AUROCcls	91.3	**99.1**	95.6

Table 3: Ablation studies of different feature layers.

f1	f2	f3	f4	f5	AUROCseg	AUROCcls
✓	✓	✓	✓	✓	94.1	93.9
	✓	✓	✓		95.4	97.0
		✓	✓	✓	**95.8**	97.2
✓	✓	✓			**95.8**	**99.1**

in 1,045 original images (defect types: spot, stain, bubble, gap, chromatic-aberration, and oil-paper).

Data Preprocessing and Partitioning. We preprocess the collected raw data through cropping, segmentation, and normalization. The dataset is then organized into positive samples (without defects) and negative samples (with defects) using the COCO format to facilitate subsequent experiments.

Manual Defect Screening. We perform manual screening, primarily removing stone images that contain arrow labels and artificially drawn lines. Subsequently, we select high-quality images to ensure both the quality and quantity of The StoneDD .

Annotation. We use the Labelme tool for annotation, which consists of two main steps: first, labeling the defective regions and generating masks; and second, categorizing each defect and recording its type. This detailed annotation process is crucial for subsequent model training and evaluation. As shown on the right side of Fig. 5, the thoroughness of this approach ensures high-quality annotations that significantly contribute to the model's performance.

The StoneDD dataset supports a range of downstream tasks, such as target detection and instance segmentation. It is characterized by small sample sizes, a wide variety of defects, and similar background textures, making it particularly suited for few-shot anomaly detection.

4.2. Evaluation Metrics

Building on prior research, this study employs AUROC, AP, and F1max metrics to assess both defect localization and detection tasks. Specifically, 'seg' refers to pixel-level defect localization, while 'cls' denotes image-level defect detection.

Among these metrics, AUROC is the most representative. However, since this study focuses on detecting de-

fect edges for localization, we place greater emphasis on the segmentation metric (i.e., AUROCseg is considered the most important and representative metric), subsequent experiments will primarily be evaluated based on the AUROC metric.

4.3. Implementation Details

This design adopts ResNet50 as the feature extraction network and chooses $n \in \{2, 3, 4\}$ as the feature layers used in calculating the anomaly localization. We train on a single NVIDIA Ampere A40 GPU for 1000 epochs on 48GB, with a batch size of 12. The learning rate is set to 1×10^{-5} for the Adam optimizer [18].

4.4. Comparison with Other Anomaly Detection Methods

4.4.1 Quantitative experiment

We employ recent state-of-the-art methods as benchmark comparisons. As illustrated in Table 1, our proposed MSD method achieves significantly superior performance in both pixel-level defect segmentation and image-level localization.

4.4.2 Qualitative experiment

To better illustrate the experimental effects, we present heatmaps for transparent visualization. This allows for a clearer observation of our refined experimental outcomes, demonstrating that our method is more suitable for stone surface defect detection, as shown in Fig. 6 and 7. Where Rec. represents reconstructed images, GT represents ground truth for the anomaly location, and Loc. represents heatmap images.

4.5. Ablation Studies

4.5.1 Effect of pre-trained feature extractors

Table 2 shows the use of different pre-trained backbone networks for quantitative comparison of feature extraction networks. We use ResNet50 as the pre-trained feature extraction network.

4.5.2 Effect of different feature layers used in anomaly score calculating

The specific data is presented in Table 3. After obtaining the feature maps at various degrees, we extract features at five different scales using a pre-trained backbone. The anomaly score is then calculated by determining the cosine similarity between the feature maps of different layers.

4.5.3 The architecture design of MSD

The method of this design achieves 99.1/97.6/99.2 and pixel-level accuracy, respectively. The AUROC/AP/F1max index of 95.8/56.8/49.7. We conducted ablation experiments on the architecture design of MSD to verify the effectiveness of SGEN and MGKD networks. The specific data is shown in Table 4.

Table 4: Ablation studies on the design of MSD.

SD	MGKD	SGEN	AUROCseg	AUROCcls
			93.8	98.7
✓			92.0	95.3
✓	✓		95.7	93.1
	✓	✓	94.6	98.6
✓	✓	✓	**95.8**	**99.1**

5. Conclusion

This paper presents a Mask-Guided and Semantic-Guided Diffusion-based framework (MSD) for Stone Surface Anomaly Detection, addressing the challenges of poor reconstruction quality and high false positive rates in diffusion-based methods. The Mask-Guided Knowledge Distillation Network (MGKD) emphasizes anomalous regions to enhance image reconstruction precision, while the Semantic-Guided Enhancement Network (SGEN) integrates multi-scale features to improve anomaly detection accuracy and reduce false detection rates.

We also introduce StoneDD to include diverse stone surface defects at varying resolutions. Our MSD approach achieves impressive localization and detection AUROC/AP scores of 99.1/97.6 and 95.8/67.3, respectively, on StoneDD. Despite its high performance, there is potential for further improvement in pixel-level localization and detection. Future work will focus on expanding The StoneDD and utilizing larger models to enhance reconstruction performance.

Acknowledgement

This work is supported by Xiamen Natural Science Foundation(Grant No.3502Z202372034), the research startup foundation of Huaqiao university(Grant No.20201XD022, HQJGYB2406) and Quanzhou Science and Technology Projects(Grant No.2023N013).

References

[1] T. Amit, T. Shaharbany, E. Nachmani, and L. Wolf. Segdiff: Image segmentation with diffusion probabilistic models. *arXiv preprint arXiv:2112.00390*, 2021. 4

[2] A. Borel, R. Deltombe, P. Moreau, T. Ingicco, M. Bigerelle, and J. Marteau. Optimization of use-wear detection and characterization on stone tool surfaces. *Scientific Reports*, 11(1):24197, 2021. 2

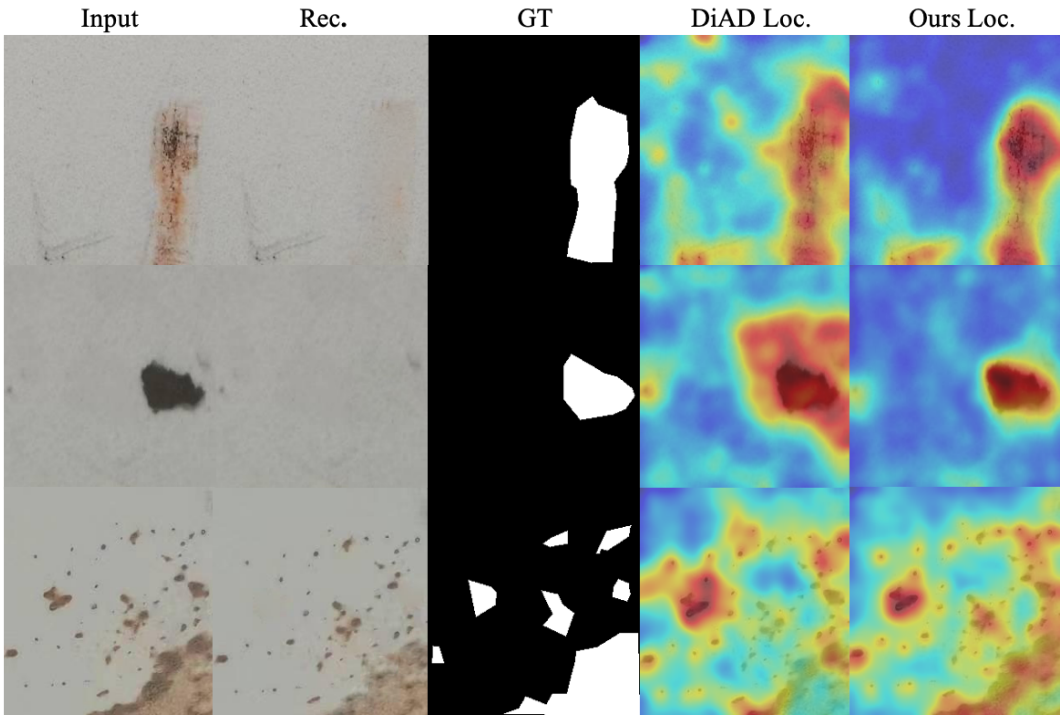

Figure 6: Qualitative comparison results of anomaly localization(256 * 256).

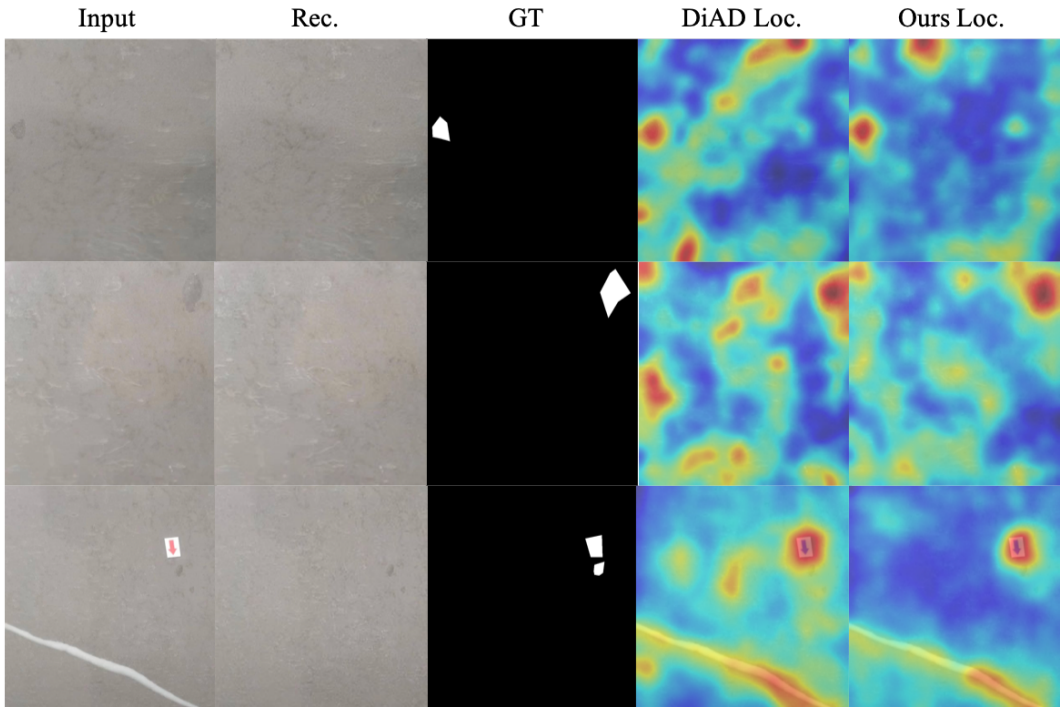

Figure 7: Qualitative comparison results of anomaly localization(900 * 900).

[3] R. Chen, G. Xie, J. Liu, J. Wang, Z. Luo, J. Wang, and F. Zheng. Easynet: An easy network for 3d industrial anomaly detection. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 7038–7046, 2023.

4

[4] S. Chen, P. Sun, Y. Song, and P. Luo. Diffusiondet: Diffusion model for object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19830–19843, 2023. 4

[5] T. Defard, A. Setkov, A. Loesch, and R. Audigier. Padim: a patch distribution modeling framework for anomaly detection and localization. In *International Conference on Pattern Recognition*, pages 475–489. Springer, 2021. 3

[6] H. Deng and X. Li. Anomaly detection via reverse distillation from one-class embedding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9737–9746, 2022. 2, 3

[7] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 3

[8] C. Ding, G. Pang, and C. Shen. Catching both gray and black swans: Open-set supervised anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7388–7398, 2022. 3

[9] M. Guerrieri and G. Parla. Flexible and stone pavements distress detection and measurement by deep learning and low-cost detection devices. *Engineering Failure Analysis*, 141:106714, 2022. 2

[10] H. He, J. Zhang, H. Chen, X. Chen, Z. Li, X. Chen, Y. Wang, C. Wang, and L. Xie. A diffusion-based framework for multi-class anomaly detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 8472–8480, 2024. 2, 4

[11] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 3

[12] J. Ho, W. Chan, C. Saharia, J. Whang, R. Gao, A. Gritsenko, D. P. Kingma, B. Poole, M. Norouzi, D. J. Fleet, et al. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022. 4

[13] H. Kabir and H.-S. Lee. Mask r-cnn-based stone detection and segmentation for underground pipeline exploration robots. *Applied Sciences*, 14(9):3752, 2024. 2

[14] P. Kapsalas, P. Maravelaki-Kalaitzaki, M. Zervakis, E. Delegou, and A. Moropoulou. Optical inspection for quantification of decay on stone surfaces. *NDT & E International*, 40(1):2–11, 2007. 2

[15] J. Lee, M. L. Smith, L. N. Smith, and P. S. Midha. Robust and efficient automated detection of tooling defects in polished stone. *Computers in industry*, 56(8-9):787–801, 2005. 2

[16] C.-L. Li, K. Sohn, J. Yoon, and T. 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. 3

[17] Y. Liang, J. Zhang, S. Zhao, R. Wu, Y. Liu, and S. Pan. Omni-frequency channel-selection representations for unsupervised anomaly detection. *IEEE Transactions on Image Processing*, 2023. 3

[18] I. Loshchilov and F. Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 8

[19] A. Mousakhan, T. Brox, and J. Tayyub. Anomaly detection with conditioned denoising diffusion models. *arXiv preprint arXiv:2305.15956*, 2023. 2

[20] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. 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. 4

[21] K. Roth, L. Pemula, J. Zepeda, B. Schölkopf, T. Brox, and P. 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. 3

[22] K. Roth, L. Pemula, J. Zepeda, B. Schölkopf, T. Brox, and P. 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. 4

[23] M. Smith and L. Smith. Machine vision inspection for polished stone manufacture. *Key Engineering Materials*, 250:131–137, 2003. 2

[24] D. S. Tan, Y.-C. Chen, T. P.-C. Chen, and W.-C. Chen. Trustmae: A noise-resilient defect classification framework using memory-augmented auto-encoders with trust regions. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 276–285, 2021. 3

[25] M. Tan and Q. Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019. 3

[26] X. Tao, X. Gong, X. Zhang, S. Yan, and C. Adak. Deep learning for unsupervised anomaly localization in industrial images: A survey. *IEEE Transactions on Instrumentation and Measurement*, 71:1–21, 2022. 3

[27] Y. Wang, J. Peng, J. Zhang, R. Yi, Y. Wang, and C. Wang. Multimodal industrial anomaly detection via hybrid fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8032–8041, 2023. 4

[28] J. Wyatt, A. Leach, S. M. Schmon, and C. G. Willcocks. Anoddpm: Anomaly detection with denoising diffusion probabilistic models using simplex noise. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 650–656, 2022. 2, 4, 5

[29] Z. You, L. Cui, Y. Shen, K. Yang, X. Lu, Y. Zheng, and X. Le. A unified model for multi-class anomaly detection. *Advances in Neural Information Processing Systems*, 35:4571–4584, 2022. 4

[30] V. Zavrtanik, M. Kristan, and D. Skočaj. Draem-a discriminatively trained reconstruction embedding for surface anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8330–8339, 2021. 2, 3

[31] L. Zhang, A. Rao, and M. Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023. 2, 4

[32] B. Zhou, Q. Cui, X.-S. Wei, and Z.-M. Chen. Bbn: Bilateral-branch network with cumulative learning for long-tailed vi-

sual recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9719–9728, 2020. 2