Detecting Anomalies in Textured Images Using Modified Transformer Masked Autoencoder

Afshin Dini[a] and Esa Rahtu[b]
Unit of Computing Sciences, Tampere University, Finland
{firstname.lastname}@tuni.fi

Keywords: Anomaly Detection, Anomaly Localization, Masked Autoencoders.

Abstract: We present a new method for detecting and locating anomalies in textured-type images using transformer-based autoencoders. In this approach, a rectangular patch of an image is masked by setting its value to gray and then fetched into a pre-trained autoencoder with several blocks of transformer encoders and decoders in order to reconstruct the unknown part. It is shown that the pre-trained model is not able to reconstruct the defective parts properly when they are inside the masked patch. In this regard, the combination of the Structural Similarity Index Measure and absolute error between the reconstructed image and the original one can be used to define a new anomaly map to find and locate anomalies. In the experiment with the textured images of the MVTec dataset, we discover that not only can this approach find anomalous samples properly, but also the anomaly map itself can specify the exact locations of defects correctly at the same time. Moreover, not only is our method computationally efficient, as it utilizes a pre-trained model and does not require any training, but also it has a better performance compared to previous autoencoders and other reconstruction-based methods. Due to these reasons, one can use this method as a base approach to find and locate irregularities in real-world applications.

SCITEPRESS SCIENCE AND TECHNOLOGY PUBLICATIONS

1 INTRODUCTION

Irregularity in vision applications pertains to an image or region of an image that deviates significantly from the typical behaviors exhibited by the majority of samples (Yang et al., 2021). Recent research in the field of computer vision focuses on the detection and localization of visual defects, which involve identifying dissimilar samples and determining the exact area of anomalous data (Pang et al., 2021). The research endeavors are particularly relevant in many real-world applications such as manufacturing and quality control (Rippel and Merhof, 2023), video surveillance (Duong et al., 2023), and medical diagnosis and healthcare (Fernando et al., 2021).

Finding defects in manufacturing products, specifically in the surface (Haselmann et al., 2018; Tsai and Jen, 2021; Liu et al., 2021) and textured (Bergmann et al., 2021; Zhou et al., 2021) images, is an interesting application among the others on which we mainly focus in this work; However, due to some intrinsic complexities of anomalies such as rarity, unknownness, unpredictability, and variety of defects, (Pang et al., 2021) developing a high-performance,

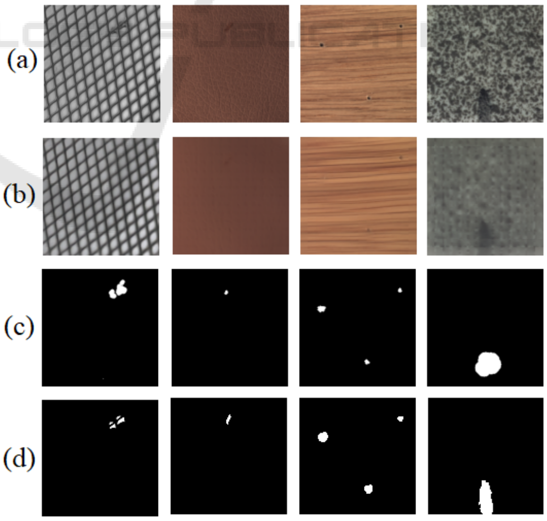

Figure 1: Detecting anomalies in textured-type images of the MVTec dataset with the proposed method. (a) Original images, (b) Reconstructed images, (c) Anomaly maps representing anomalous samples and defective areas, (d) Ground truth.

[a] https://orcid.org/0000-0001-6234-3322
[b] https://orcid.org/0000-0001-8767-0864

Dini, A. and Rahtu, E.
Detecting Anomalies in Textured Images Using Modified Transformer Masked Autoencoder.
DOI: 10.5220/0012416400003660
Paper published under CC license (CC BY-NC-ND 4.0)
In *Proceedings of the 19th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2024) - Volume 2: VISAPP*, pages 191-200
ISBN: 978-989-758-679-8; ISSN: 2184-4321
Proceedings Copyright © 2024 by SCITEPRESS – Science and Technology Publications, Lda.

and generalized method is a challenging task. In other words, anomalies rarely occur in real applications, as a result of which labeling samples and preparing a labeled dataset for training a supervised model is a cumbersome effort in most cases (Pang et al., 2021).

Irregularities are also unknown before they happen, and they can appear in various shapes, sizes, and colors (Chandola et al., 2009). Moreover, generalization is a critical issue in detecting defects as methods trained on one domain may struggle to generalize to new, unseen data, mainly when significant differences exist between the training and testing distributions.

To deal with the abovementioned complexities, many semi-supervised approaches have been developed that only utilize normal samples for training purposes to avoid the difficulties of collecting labeled anomalous samples. Convolutional Autoencoder (CAE) (Tsai and Jen, 2021) and its extensions (Schneider et al., 2022), such as Variational Autoencoder (VAE) (An and Cho, 2015) and Adversarial Autoencoder (AAE) (Beggel et al., 2020), are some of the most common architectures for detecting anomalies; However, their performance is limited as they are not able to find subtle defects in some applications.

On the other hand, self-supervised methods such as TPSAD (Dini and Rahtu, 2022) and CutPaste (Li et al., 2021) demonstrate better performance than autoencoders as they try to simulate anomalies in different pretext tasks and use them alongside normal ones to train more powerful models; However, these models are not generalized enough in some applications (Dini and Rahtu, 2022) as the simulation procedure may not be able to create various types of irregularities to deal with the defect unknownness challenge. A large number of normal and simulated samples are also required for training purposes, causing the training phase to be computationally heavy in these methods (Dini and Rahtu, 2023).

Other semi-supervised methods such VT-ADL (Mishra et al., 2021) aim to enhance their detection performance by integrating vision transformers (Dosovitskiy et al., 2021) into their architecture as they can capture semantic and detailed information of samples (Dini and Rahtu, 2023). However, training these models demands significant computational resources and extensive training datasets due to the significant number of trainable parameters involved.

It is shown in (Kumar et al., 2022) that many foundational models such as ViT (Dosovitskiy et al., 2020), and EfficientNet (Tan and Le, 2019), trained with large datasets such as ImageNet (Deng et al., 2009) to solve a specific problem, can also be used for other purposes such as anomaly detection as they provide detailed representations of data.

By using these foundational models, we present a new semi-supervised approach to tackle some of the limitations mentioned above. Our primary goal is to develop a simple architecture that can be used as a basic and general method to detect and locate anomalies in textured-type data, which is computationally efficient and has better performance than previous similar approaches. In this regard, we use a pre-trained masked autoencoder (He et al., 2022), containing several blocks of transformer encoders and decoders, as the backbone model of our method, which is able to reconstruct a partly masked image properly. To detect defects in each sample, we shift the masked patch through the image and fetch it to the model to reconstruct the unknown part consecutively and then combine the results to rebuild the whole image.

To detect and locate anomalies, a new anomaly map is defined based on the reconstruction error. For this purpose, the Structural Similarity Index Measure (SSIM) (Hassan and Bhagvati, 2012) and the absolute error (L1) are calculated between the reconstructed and original images and then combined together to form the error map. The final anomaly map is produced by applying a Morphology filter (Soille and Soille, 2004) to remove salty noises in the error map. The anomalies can be detected by applying a threshold on the average value of anomaly maps, while the anomaly maps themselves present the locations of irregularities in the defective samples.

It is shown in section 3 that our method is computationally optimized as it eliminates time-consuming tasks typically associated with the training phase, such as tuning the parameters, augmenting samples, and simulating defects. Moreover, our method shows better performance compared to previous similar methods in detecting anomalies in textured datasets, according to which one can use it as the base method for anomaly detection purposes. We have evaluated our method in textured images of the MVTec dataset (Bergmann et al., 2021) and compared the results to similar state-of-the-art methods in section 4.

2 RELATED WORK

There has been significant research emphasis on developing semi-supervised approaches in recent years, as they can deal with a few complexities of detecting anomalies, such as the rarity of defects and the limitation of labeling anomalous samples for the training phase (Mohammadi et al., 2021). Most of these methods can be categorized as reconstruction-based methods, one-class detectors, and self-supervised approaches (Pang et al., 2021) which we mainly focus on reconstruction-based ones due to their simplicity and performance on texture-type images.

Reconstruction-based techniques (Liu et al., 2023) are the most common and easy-to-use methods that attempt to detect defects by reconstructing images from latent space. Since these models are optimized only with normal samples in the training phase, they can only reconstruct normal patterns properly from the latent space, while the unknown irregularities will be missed in the reconstructed images, as a result of which anomalies can be identified from the reconstruction errors. Depending on the reconstruction methodology, these methods vary from Autoencoders (AEs) (Liu et al., 2023) with different architectures and loss functions to various types of Generative Adversarial Networks (GANs) (Li and Li, 2023; Di Mattia et al., 2019), transformer models and inpainting methods (Liu et al., 2023).

Convolutional Autoencoder (CAE) (Tsai and Jen, 2021) is the simplest type of reconstruction-based method utilizing convolutional encoders and decoders to detect anomalies. In these approaches, the encoder and decoder are trained with normal images in such a way that the encoder extracts the latent space of normal samples while the decoder is responsible for reconstructing the normal patterns from the related latent space. With the help of an appropriate anomaly metric, defined based on the reconstruction error, irregularities can be detected properly in these approaches. Depending on the variety of loss functions and anomaly metrics, many types of convolutional autoencoders have been developed as methods like (Bergmann et al., 2018) use SSIM, ITAE (Huang et al., 2019) and (Chung et al., 2020) use a combination of L2 and SSIM, MAMA Net (Chen et al., 2021) makes use of multi-scale SSIM, and CW-SSIM (Bionda et al., 2022) takes advantages of complex-wavelet SSIM to improve its performance, especially on the textured images.

Variational Autoencoders (VAEs) (An and Cho, 2015), such as VAE-grad (Dehaene et al., 2020) and FAVAE (Dehaene and Eline, 2020) are other variants of autoencoders that attempt to improve the detection performance by employing the probabilistic manners of data samples in the training phase. In these approaches, it is assumed that the latent space follows a certain probability distribution, typically a multivariate Gaussian one, according to which a regularization error is defined in addition to the reconstruction error to encourage the latent space to match the related distribution. The combination of regularization and reconstruction errors not only makes the training process efficient but also improves the detection performance (An and Cho, 2015).

Autoencoders are still the most interesting methods in the field of anomaly detection, and many researchers have focused on developing them in recent years due to their simplicity; however, they have some limitations as they are not able to detect detailed irregularities properly as the reconstruction errors of small defects will be small in such a way that they may be neglected in the testing phase (Pang et al., 2021). Moreover, fine-tuning an autoencoder, which is powerful enough to reconstruct only normal images and not anomalies, is cumbersome in some applications as deep architectures with large bottlenecks may mistakenly reconstruct anomalies which will reduce the true negative detection rate, while shallow models with small bottlenecks cannot reconstruct even normal samples as a result of which the false positive detection rate will be significant.

Generative adversarial methods, such as GANomaly (Akcay et al., 2019) and AnoGAN (Schlegl et al., 2017), are other types of reconstruction-based methods that typically contain a generator, which reconstructs images from a vector space, and a discriminator, trying to find out whether the created images are similar to the original ones or not. This adversarial process pushes both networks to improve iteratively, resulting in the generator producing increasingly realistic samples. Once the model is trained with normal samples, the discriminator's output can be used as the anomaly metric for detecting defects. When presented with a test image, the discriminator assigns a probability score indicating how well the image matches the distribution of the normal samples, as a result of which images with low scores are considered anomalies. In these techniques, the interaction between the generator and the discriminator allows the model to be trained in such a way that it shows better performance than autoencoders in detecting large and fine-grained irregularities; however, they have stability problems, specifically when the training dataset is small, which limits their usage in some real-world applications (Li and Li, 2023).

Transformer-based methods, such as AnoViT (Lee and Kang, 2022) and VT-ADL (Mishra et al., 2021), are recently developed methods that exhibit an enhanced ability to capture semantic and detailed information of images by utilizing convolutional networks alongside transformers in their architecture, as a result of which they become new reconstruction frameworks for anomaly detection that outperform autoencoders and GAN-based methods. These models have considerably larger sizes compared to convolutional networks, demanding a greater number of samples and computational resources for their train-

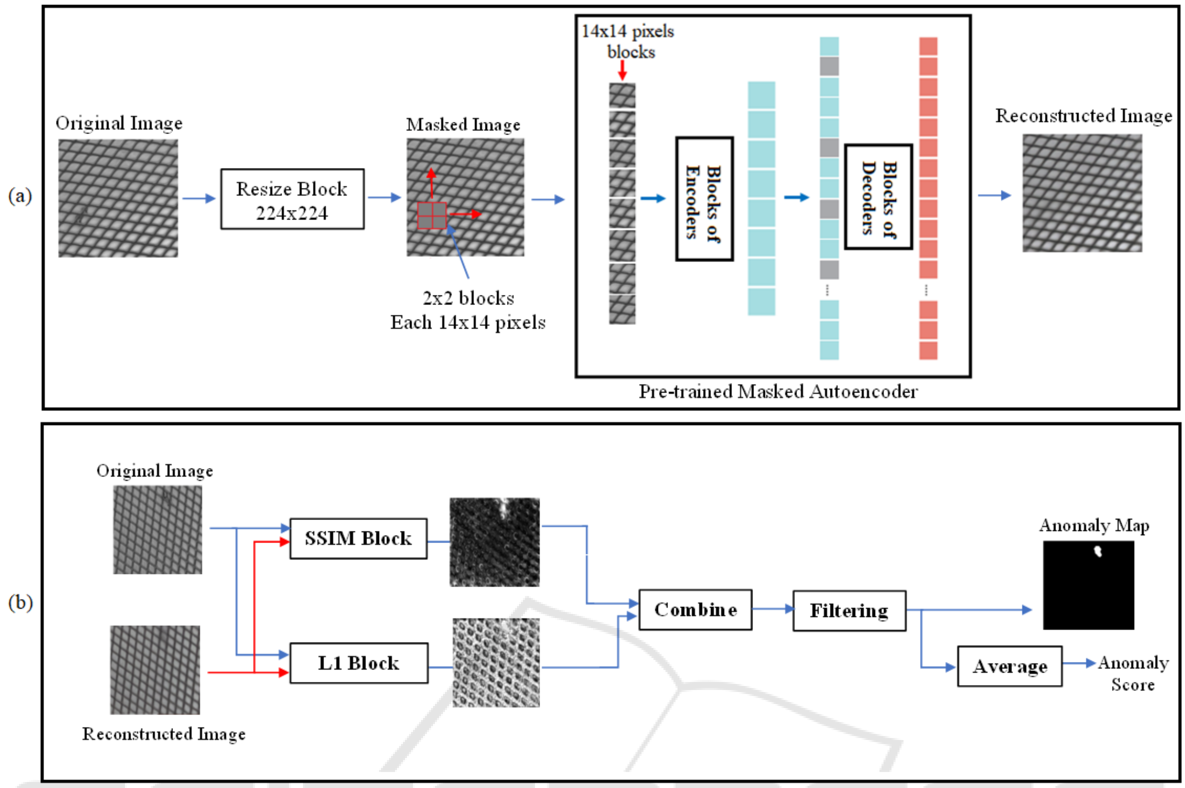

Figure 2: Overview and model architecture of the proposed method for visual anomaly detection in texture-type datatsets. (a) Using a pre-trained masked autoencoder (He et al., 2022) as the backbone model to reconstruct masked images, (b) Creating anomaly maps and scores based on the combination of SSIM and L1 error between reconstruction and original images.

ing which limits their usage considering the fact that many anomaly detection datasets contain only a small number of normal samples for training purposes.

Inpainting methods, such as InTra (Pirnay and Chai, 2022), divide the images into multi-scale patches and attempt to reconstruct the images by integrating the information of larger patches into the related smaller ones. Although these methods exhibit high accuracy in detecting anomalies, they are computationally heavy in the training and testing phase due to the need to divide images into several patches and combine the patches' information. Moreover, selecting the appropriate patch size is an important issue in these approaches and affects their performance significantly, which increases the complexity of these methods and reduces the generalizability (Pirnay and Chai, 2022) as the patch size shall be fine-tuned to fit different applications. These challenges limit the usage of inpainting methods in real-time applications (Dini and Rahtu, 2023).

It is worth mentioning that some of the above approaches, such as (Dini and Rahtu, 2023) attempt to utilize fundamental architectures, pre-trained on large datasets such as ResNet (He et al., 2016), to extract the representations of data and utilize them within their techniques to detect and locate anomalies. Using pre-trained models allows these approaches to deal with the challenge of small datasets in anomaly detection problems.

Our proposed method, discussed in more detail in section 3, is developed based on autoencoders and transformers in such a way that it can address the aforementioned limitations. Moreover, the performance of our method is compared with various types of convolutional and variational autoencoders, GAN-based methods, and transformer-based methods in section 4 to show that it can be used as a basic and general method to detect defects in textured images.

3 METHOD

3.1 Overview

We propose a new reconstruction-based method for detecting anomalies in texture-type images that utilizes a pre-trained masked autoencoder (He et al., 2022) as the backbone model, containing several transformer encoders and decoders. This pre-trained model is able to properly reconstruct masked images, where a rectangular patch of them is replaced with gray values in accordance with the unmasked areas. The rectangular patch will be slid through the image step by step, and in each step, the masked image will be fetched into the model to create the unknown area and finally reconstruct the whole image. Since masked defects cannot be reconstructed properly in this iterative procedure, they will not appear in the final images, as a result of which the combination of the absolute error and the structural similarity index measure between the reconstructed and original images is an appropriate metric to detect anomalies.

The overview and model architecture of the proposed method is described in detail in Fig. 2. Selecting a pre-trained transformer-based architecture, combining appropriate metrics to create an anomaly map, and detecting anomalies in addition to the masking procedure are some important properties of our method that will be explored in detail in this section.

3.2 Model Architecture

Selecting an appropriate model architecture in the reconstruction-based defect detection methods, specifically autoencoders, is an important task and can affect the performance of the method significantly (Liu et al., 2023) as an appropriate model should be able to capture semantic and detailed information to recreate only non-anomalous areas and not defective ones (Dini and Rahtu, 2023). Generally, in deep autoencoders with large bottlenecks, the model is too powerful that reconstructs normality and abnormality at the same time, leading to the fact that anomalous samples cannot be detected due to small reconstruction errors in the anomalous areas. Shallower models, on the other hand, cannot reconstruct even normal images properly, as a result of which normal samples will be mistakenly considered as anomalies. A small training dataset makes this issue even more challenging, as large models require a huge number of samples for the training phase.

To deal with this problem, we utilize a transformer-based autoencoder (He et al., 2022), pre-trained on ImageNet (Deng et al., 2009), which contains 32 blocks of encoders and 8 blocks of decoders. Each test image is resized to a 224 × 224 image and divided into several non-overlapping patches of 14 × 14 pixels, and then a few patches are selected and masked randomly. It is good to mention that resizing images to 224 × 224 is a common transformation in the anomaly detection approaches, as irregularities will still be visible in the resized images. The encoders attempt to embed unmasked patches and learn their structures, while the decoders aim to reconstruct the masked parts based on the unmasked areas, Fig. 2(a). We utilize this property for detecting anomalies and show in section 4 that this model can replace anomalous patches that are masked intentionally with the normal ones, based on the structure of the known patches, as a result of which anomalies can be detected in the reconstruction errors.

This method can be used as a basic approach to detect defects in textured images as it has an impressive capability to find subtle and large irregularities and is computationally efficient due to the fact that it skips the difficulties of the training phase, such as data augmentation, architecture configuration, and tuning parameter. Defining an appropriate anomaly map and masking procedure are the two important issues by the proposed method that are discussed in the next sections.

3.3 Masking Procedure

Selecting an appropriate mask size during the testing phase depends on the size of probable irregularities as well as the patch size used in the encoder and decoder blocks.

First of all, multiple blocks of patches with 14 × 14 pixels can be masked. Moreover, since the proposed method attempts to replace the masked areas with the patterns of unmasked regions, it is important to select large masks that can hide small and large anomalies as they slide through the image but not too large to miss the normal patterns. On the other hand, larger masks increase the speed of the testing phase compared to the smaller ones, which is important in real-time applications.

Taking into account the above context, we attempt to represent the patching and masking procedures with math equations as it gives a better understanding of the whole process. Considering X_T as the resized test dataset, a test sample, $x^t \in X_T$ with the size of 224 × 224 pixels, can be divided into a set of patches, P^t, with the size of 14 × 14 pixels, as:

$$P^t = \{ \begin{bmatrix} x^t_{14i,14j} & \cdots & x^t_{14i,14j+13} \\ \vdots & \ddots & \vdots \\ x^t_{14i+13,14j} & \cdots & x^t_{14i+13,14j+13} \end{bmatrix} , (i,j) \in [0,15] \times [0,15] \} \quad (1)$$

where $x^t_{m,n}$ is the pixel value of the related test sample x^t at location (m,n).

For the masking procedure, we experimentally discover that selecting 2 × 2 blocks of patches as masks is an appropriate option compared to other possible sizes, which not only is suitable for detecting various types of large and small defects but also speeds up the testing phase compared to smaller mask areas. Similar to Eq. 1, a set of masks, M^t, is defined for each test sample in Eq. 2:

$$M^t = \{ \begin{bmatrix} p^t_{2l,2k} & \cdots & p^t_{2l,2k+1} \\ \vdots & \ddots & \vdots \\ p^t_{2l+1,2k} & \cdots & p^t_{2l+1,2k+1} \end{bmatrix} , (l,k) \in [0,7] \times [0,7] \} \tag{2}$$

where $p^t_{m,n}$ represents the pixel values of the related 14×14 patch at location (m,n) for the test sample x^t.

3.4 Anomaly Map and Score

To identify abnormal samples, it is necessary to assign a distinct value, commonly referred to as an anomaly score, to each test sample. By setting an appropriate threshold on these scores, we can identify defective samples. Similarly, an anomaly map can be created by assigning anomaly scores to pixels within an image. This map helps in identifying the exact locations of irregularities presented in the associated abnormal samples.

We find out that defining an anomaly score based on a precise anomaly map detects irregularities better than previous methods (Pang et al., 2021) utilizing a single vector for anomaly detection. In this regard, the SSIM map and L1 error between the reconstructed and original images are combined together, filtered by a Morphology filter, and thresholded to create the final anomaly map for each test image. The average value of the anomaly map is considered an anomaly score for detecting anomalies, while the anomaly map itself represents the location of abnormalities 2(b). In section 4, we present an evaluation of our method and provide a detailed analysis of the obtained results.

4 EXPERIMENT

4.1 Dataset and Metric

Following the common procedure in anomaly detection research, we assess the effectiveness of our approach on the textured images of the MVTec dataset (Bergmann et al., 2021). This dataset consists of high-resolution images, ranging from 840 × 840 to 1024 × 1024 pixels, obtained from various real-world industrial applications. With its diverse range of image types, colors, and textures, this dataset serves as an appropriate choice for evaluating the performance and generalizability of our proposed method on texture-type images.

Table 1: Summary of MVTec (Bergmann et al., 2021) textured sub-datasets.

Category	Grid	Leather	Wood	Carpet	Tile
Traning Samples	264	245	247	280	230
Normal Test Samples	21	32	19	29	33
Defective Test Samples	57	95	60	89	84
Defective Test Samples	57	95	60	89	84
Defective Groups	5	5	5	5	5
Image Size	1024	1024	1024	1024	840
Type	Gray	Color	Color	Color	Color

The MVTec textured images are categorized into five groups such as grid, leather, wood, carpet, and tile, as is shown in Tab. 1. Each category consists of a limited number of normal images for training purposes, ranging from 230 to 280, which poses a significant challenge in developing deep models that have a large number of trainable parameters. Moreover, each category includes a small number of normal and anomalous test samples with defects of various sizes, shapes, colors, and types to assess the method's generalizability. The color scheme of these images can differ, with some being in full color while others are in grayscale.

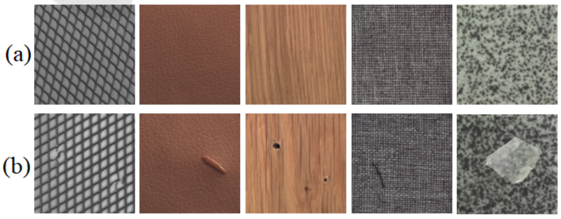

Figure 3: Overview of the MVTec textured-type samples. (a) Normal images, (b) Anomalous images.

It is also worth mentioning that grid, wood, and leather images have regular patterns in their textures in addition to the fact the defective areas are small or middle size, while in the carpet and tile datasets, the texture patterns are irregular, and the anomalous areas are larger than previous datasets specifically in the tile dataset, as is shown in Fig. 3. We discuss these properties of samples in the next section in more detail as

SCITEPRESS SCIENCE AND TECHNOLOGY PUBLICATIONS

Table 2: Comparison of our approach with different types of reconstruction-based methods vary from autoencoders to generative adversarial methods and transformer-based methods, such as CAE (Tsai and Jen, 2021), CAE-SSIM (Bergmann et al., 2018), CAE-MSSSIM (Bionda et al., 2022), CAE-CWSSIM (Bionda et al., 2022), ITAE (Huang et al., 2019), DAAD (Hou et al., 2021), VAE (An and Cho, 2015), VAE-grad (Dehaene et al., 2020), FAVAE (Dehaene and Eline, 2020), GANomaly (Akcay et al., 2019; Tang et al., 2020), AnoGAN (Schlegl et al., 2017; Tang et al., 2020), Skip-GANomaly (Akçay et al., 2019; Tang et al., 2020), VT-ADL (Mishra et al., 2021), AnoViT (Lee and Kang, 2022), and Feature Dictionary (Napoletano et al., 2018). The detection results are presented on texture-type images of the MVTec dataset (Bergmann et al., 2021), using the AUROC metric.

Reconstruction-based Anomaly Detection Methods		Grid	Leather	Wood	Carpet	Tile	Average
Convolutional AEs	CAE	66.2	65.8	57.4	54.3	55.2	59.8
	CAE-SSIM	85.2	80.0	70.6	75.4	65.0	75.2
	CAE-MSSSIM	92.7	69.6	64.0	75.3	67.9	73.9
	CAE-CWSSIM	97.8	97.9	75.3	**94.7**	84.7	90.1
	ITAE	88.3	86.2	92.3	70.6	73.5	82.2
	DAAD	97.5	62.8	95.7	67.1	82.5	81.1
Variational AEs	VAE	88.8	83.4	69.5	58.0	46.5	69.2
	VAE-grad	96.1	92.5	83.8	73.5	65.4	82.3
	FAVAE	97.0	67.5	94.8	67.1	80.5	81.4
Generative Adversarial Networks	GANomaly	74.3	80.8	92.0	82.1	72.0	80.2
	AnoGAN	87.1	45.1	56.7	33.7	40.1	52.6
	Skip-GANomaly	65.7	90.8	91.9	79.5	85.0	82.6
Transformer based ADs	VT-ADL	87.1	72.8	78.1	77.3	79.6	79.0
	AnoViT	52.0	85.0	95.0	50.0	**89.0**	74.2
Others	Feature Dictionary	87.2	81.9	72.0	94.3	85.4	84.2
Masked Transformer AE	Ours	**98.2**	**99.4**	**96.9**	89.9	78.2	**92.5**

SCITEPRESS SCIENCE AND TECHNOLOGY PUBLICATIONS

they can affect the results. The AUROC, which stands for the Area Under the Receiver Operating Characteristic curve, is employed to evaluate and compare the detection performance of our approach with other similar results from previous studies.

4.2 Implementation Details

The large model of Masked Autoencoder (He et al., 2022), pre-trained on ImageNet (Deng et al., 2009), is the backbone of our method to reconstruct masked images. In the first step, the images are resized to 224×224 and divided into patches of 14×14 pixels to match the autoencoder structure. For the masking procedure, 2×2 blocks of patches are masked, slid through the image, and fetched into the model to reconstruct the whole image.

To create an anomaly map, the structural similarity matrix between the original and reconstructed images is created with a window size of 7, 9, or 11, depending on the sub-dataset, and combined with the related absolute error. The final anomaly mask is created by applying a Morphology filter with a size of 2 or 4. The average value of pixels of the anomaly map is considered as the anomaly score for each image on which setting a threshold identifies the anomalous samples. The anomaly map itself also represents the locations of irregularities in defective samples.

4.3 Results

To evaluate our method's performance and make a comparison with similar reconstruction-based approaches, we execute our approach on textured sub-datasets and compute the AUROC for each sub-dataset. Additionally, we determine the average AUROC across all categories to give a better comparison. We present the outcome of evaluating our method for detecting anomalies in Tab. 2.

To show that this method can be used as the basic and general method to detect anomalies in texture-type images, we compare our results with various types of reconstruction-based methods with different kinds of loss functions and anomaly scores.

In this respect, the results are compared with convolutional autoencoders with L2 loss function (Tsai and Jen, 2021), SSIM (Bergmann et al., 2018), multi-scale SSIM (Bionda et al., 2022), a combination of L2 and SSIM (Huang et al., 2019), complex-wavelet SSIM function (Bionda et al., 2022) specif-

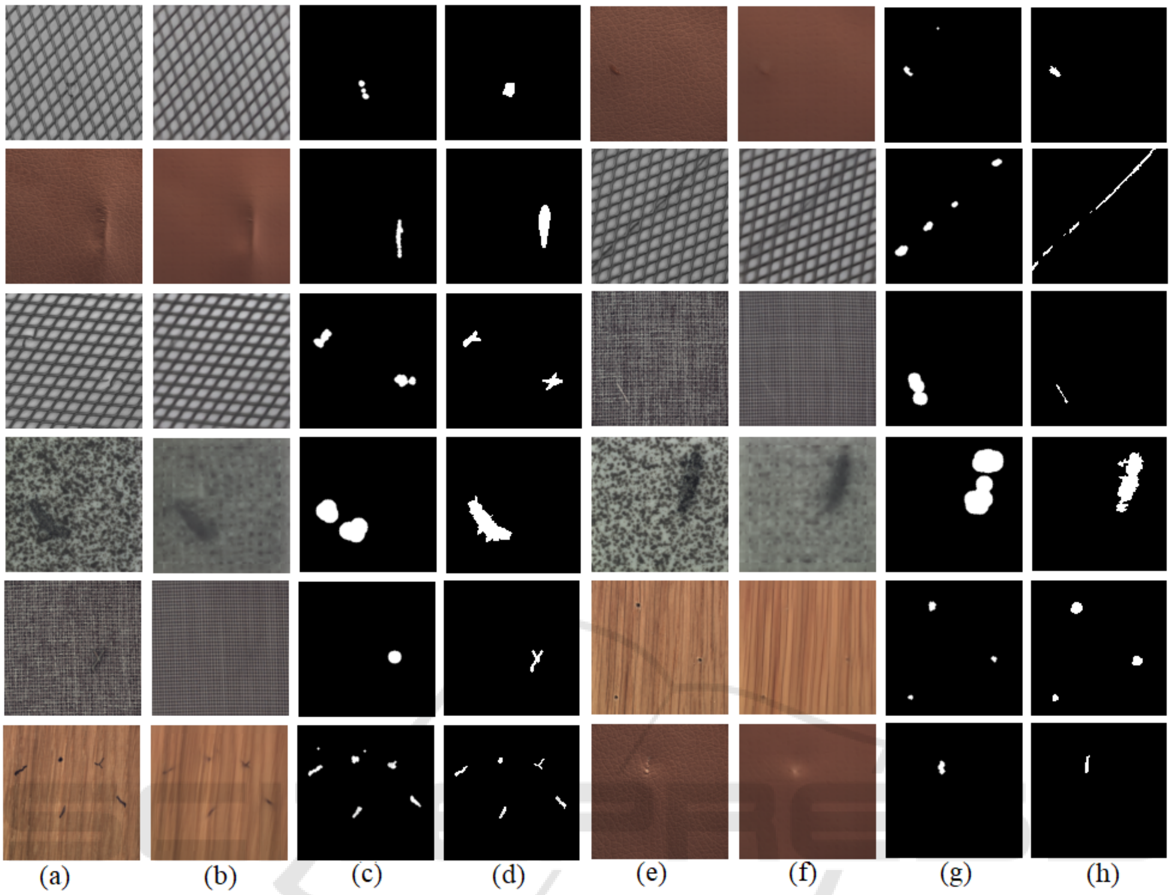

Figure 4: Visualization of defect localization using the proposed method. (a) and (e) Original images, (b) and (f) Reconstructed images, (c) and (g) Anomaly maps, (d) and (h) Ground truth maps.

ically developed for detecting textured defects, and with block-wise autoencoders (Hou et al., 2021). It can be deducted from Tab. 2 that our method not only improves the average AUROC of all categories by more than 2.4 percent but also can detect small and large irregularities from grid, leather, and wood categories better than the recently developed convolutional autoencoders.

It is obvious that the proposed method gives better results than variational autoencoders such as VAE (An and Cho, 2015), VAE-grad (Dehaene et al., 2020), and FAVAE (Dehaene and Eline, 2020) in all categories and improves the average AUROC by more than 10 percent. It is obvious that the proposed method can appropriately detect subtle anomalies, Fig. 4 first and fifth rows, as well as larger ones, Fig. 4 second and fourth rows, properly as a result of which it aims to solve the problems of autoencoders mentioned in section 2.

By comparing our results with GAN-based reconstruction methods such as GANomaly (Akcay et al., 2019), AnoGAN (Schlegl et al., 2017), and Skip-GANomaly (Akçay et al., 2019), we deduce that not only our method does not have the stability problems of GAN-based approaches, but also it is computationally efficient as it utilizes a pre-trained network instead of training the backbone model from scratch.

It is also clear from Tab. 2 that our method has better performance than transformer-based autoencoders such as VT-ADL (Mishra et al., 2021) and AnoViT (Lee and Kang, 2022) as transformers require a large dataset for training a powerful reconstruction model while our backbone model skips the training phase and is powerful enough to detect anomalies based on the reconstruction error.

By analyzing and comparing the results in Tab. 2 and Fig. 4, one can deduce that the AUROC in the tile dataset is smaller than a few methods, although it is greater than most of the other approaches. This mainly happens due to the fine and irregular patterns of tile images in addition to the existence of large defective areas in these samples. In other words, since anomalies are larger than the masked area, some parts of the anomalies will also be reconstructed, which reduces the reconstruction error in the defective areas. On the other hand, the irregular and fine patterns of tile images cannot be reconstructed properly, which increases the error totally, resulting in a smaller AUROC than a few methods.

It is also important to mention that as the anomalous samples are detected based on the generated anomaly maps, the proposed method can also represent the location of irregularities, as is shown in Fig. 4, although the detected areas of the defects might be a bit different compared to the ground truth area in some cases.

5 CONCLUSION

We have developed a new approach for detecting anomalies in textured images based on a pre-trained masked autoencoder, which is able to replace abnormal masked areas of an image with normal patterns and detect anomalies based on generating an anomaly map from the difference between the reconstructed and original areas in the test samples.

Through the evaluation of our method on the textured images of the MVTec dataset, we demonstrate that the proposed method not only has the superior capability to detect and locate subtle and large anomalies but also is computationally efficient as it bypasses the time-consuming process of training deep models from scratch. Improving the accuracy of detecting anomalies as well as having good stability are other properties of our method compared to similar approaches. These aspects make our method a well-suited candidate for detecting anomalies in real-world applications.

REFERENCES

Akcay, S., Atapour-Abarghouei, A., and Breckon, T. P. (2019). Ganomaly: Semi-supervised anomaly detection via adversarial training. In *Computer Vision–ACCV 2018: 14th Asian Conference on Computer Vision, Perth, Australia, December 2–6, 2018, Revised Selected Papers, Part III 14*, pages 622–637. Springer.

Akçay, S., Atapour-Abarghouei, A., and Breckon, T. P. (2019). Skip-ganomaly: Skip connected and adversarially trained encoder-decoder anomaly detection. In *2019 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE.

An, J. and Cho, S. (2015). Variational autoencoder based anomaly detection using reconstruction probability.

Beggel, L., Pfeiffer, M., and Bischl, B. (2020). Robust anomaly detection in images using adversarial autoencoders. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2019, Würzburg, Germany, September 16–20, 2019, Proceedings, Part I*, pages 206–222. Springer.

Bergmann, P., Batzner, K., Fauser, M., Sattlegger, D., and Steger, C. (2021). The mvtec anomaly detection dataset: a comprehensive real-world dataset for unsupervised anomaly detection. *International Journal of Computer Vision*, 129(4):1038–1059.

Bergmann, P., Löwe, S., Fauser, M., Sattlegger, D., and Steger, C. (2018). Improving unsupervised defect segmentation by applying structural similarity to autoencoders. *arXiv preprint arXiv:1807.02011*.

Bionda, A., Frittoli, L., and Boracchi, G. (2022). Deep autoencoders for anomaly detection in textured images using cw-ssim. In *International Conference on Image Analysis and Processing*, pages 669–680. Springer.

Chandola, V., Banerjee, A., and Kumar, V. (2009). Anomaly detection: A survey. *ACM computing surveys (CSUR)*, 41(3):1–58.

Chen, Y., Zhang, H., Wang, Y., Yang, Y., Zhou, X., and Wu, Q. M. J. (2021). Mama net: Multi-scale attention memory autoencoder network for anomaly detection. *IEEE Transactions on Medical Imaging*, 40(3):1032–1041.

Chung, H., Park, J., Keum, J., Ki, H., and Kang, S. (2020). Unsupervised anomaly detection using style distillation. *IEEE Access*, 8:221494–221502.

Dehaene, D. and Eline, P. (2020). Anomaly localization by modeling perceptual features. *arXiv preprint arXiv:2008.05369*.

Dehaene, D., Frigo, O., Combrexelle, S., and Eline, P. (2020). Iterative energy-based projection on a normal data manifold for anomaly localization. *arXiv preprint arXiv:2002.03734*.

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

Di Mattia, F., Galeone, P., De Simoni, M., and Ghelfi, E. (2019). A survey on gans for anomaly detection. *arXiv preprint arXiv:1906.11632*.

Dini, A. and Rahtu, E. (2022). Tpsad: Learning to detect and localize anomalies with thin plate spline transformation. *2022 26th International Conference on Pattern Recognition (ICPR)*, pages 4744–4750.

Dini, A. and Rahtu, E. (2023). Visual anomaly detection and localization with a patch-wise transformer and convolutional model. In *International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*. SCITEPRESS.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*.

Duong, H.-T., Le, V.-T., and Hoang, V. T. (2023). Deep learning-based anomaly detection in video surveillance: A survey. *Sensors*, 23(11):5024.

Fernando, T., Gammulle, H., Denman, S., Sridharan, S., and Fookes, C. (2021). Deep learning for medical anomaly detection–a survey. *ACM Computing Surveys (CSUR)*, 54(7):1–37.

Haselmann, M., Gruber, D. P., and Tabatabai, P. (2018). Anomaly detection using deep learning based image completion. In *2018 17th IEEE international conference on machine learning and applications (ICMLA)*, pages 1237–1242. IEEE.

Hassan, M. and Bhagvati, C. (2012). Structural similarity measure for color images. *International Journal of Computer Applications*, 43(14):7–12.

He, K., Chen, X., Xie, S., Li, Y., Dollár, P., and Girshick, R. (2022). Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009.

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

Hou, J., Zhang, Y., Zhong, Q., Xie, D., Pu, S., and Zhou, H. (2021). Divide-and-assemble: Learning block-wise memory for unsupervised anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8791–8800.

Huang, C., Cao, J., Ye, F., Li, M., Zhang, Y., and Lu, C. (2019). Inverse-transform autoencoder for anomaly detection. *arXiv preprint arXiv:1911.10676*, 2(4).

Kumar, J. S., Anuar, S., and Hassan, N. H. (2022). Transfer learning based performance comparison of the pretrained deep neural networks. *International Journal of Advanced Computer Science and Applications*, 13(1).

Lee, Y. and Kang, P. (2022). Anovit: Unsupervised anomaly detection and localization with vision transformer-based encoder-decoder. *IEEE Access*, 10:46717–46724.

Li, C.-L., Sohn, K., Yoon, J., and Pfister, T. (2021). 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.

Li, H. and Li, Y. (2023). Anomaly detection methods based on gan: a survey. *Applied Intelligence*, 53(7):8209–8231.

Liu, J., Song, K., Feng, M., Yan, Y., Tu, Z., and Zhu, L. (2021). Semi-supervised anomaly detection with dual prototypes autoencoder for industrial surface inspection. *Optics and Lasers in Engineering*, 136:106324.

Liu, J., Xie, G., Wang, J., Li, S., Wang, C., Zheng, F., and Jin, Y. (2023). Deep industrial image anomaly detection: A survey. *arXiv preprint arXiv:2301.11514*, 2.

Mishra, P., Verk, R., Fornasier, D., Piciarelli, C., and Foresti, G. L. (2021). Vt-adl: A vision transformer network for image anomaly detection and localization. In *2021 IEEE 30th International Symposium on Industrial Electronics (ISIE)*, pages 01–06. IEEE.

Mohammadi, B., Fathy, M., and Sabokrou, M. (2021). Image/video deep anomaly detection: A survey. *arXiv preprint arXiv:2103.01739*.

Napoletano, P., Piccoli, F., and Schettini, R. (2018). Anomaly detection in nanofibrous materials by cnn-based self-similarity. *Sensors*, 18(1):209.

Pang, G., Shen, C., Cao, L., and Hengel, A. V. D. (2021). Deep learning for anomaly detection: A review. *ACM computing surveys (CSUR)*, 54(2):1–38.

Pirnay, J. and Chai, K. (2022). Inpainting transformer for anomaly detection. In *International Conference on Image Analysis and Processing*, pages 394–406. Springer.

Rippel, O. and Merhof, D. (2023). Anomaly detection for automated visual inspection: A review. *Bildverarbeitung in der Automation: Ausgewählte Beiträge des Jahreskolloquiums BVAu 2022*, pages 1–13.

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

Schneider, S., Antensteiner, D., Soukup, D., and Scheutz, M. (2022). Autoencoders-a comparative analysis in the realm of anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1986–1992.

Soille, P. and Soille, P. (2004). Erosion and dilation. *Morphological Image Analysis: Principles and Applications*, pages 63–103.

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

Tang, T.-W., Kuo, W.-H., Lan, J.-H., Ding, C.-F., Hsu, H., and Young, H.-T. (2020). Anomaly detection neural network with dual auto-encoders gan and its industrial inspection applications. *Sensors*, 20(12):3336.

Tsai, D.-M. and Jen, P.-H. (2021). Autoencoder-based anomaly detection for surface defect inspection. *Advanced Engineering Informatics*, 48:101272.

Yang, J., Xu, R., Qi, Z., and Shi, Y. (2021). Visual anomaly detection for images: A survey. *arXiv preprint arXiv:2109.13157*.

Zhou, K., Li, J., Xiao, Y., Yang, J., Cheng, J., Liu, W., Luo, W., Liu, J., and Gao, S. (2021). Memorizing structure-texture correspondence for image anomaly detection. *IEEE Transactions on Neural Networks and Learning Systems*, 33(6):2335–2349.