International Journal of Computing and Digital Systems
ISSN (2210-142X)
Int. J. Com. Dig. Sys. #, No.# (Mon-20..)

http://dx.doi.org/10.12785/ijcds/XXXXXX

A Generative Encoder-Decoder Model for Automated Quality Control Inspections

Khedidja Mekhilef [1], Fayçal Abbas[2] and Mounir Hemam[1]

[1]*University Abbes Laghrour - Khenchela, ICOSI Laboratory, BP 1252 El Houria, 40004, Algeria*
[2]*University Abbes Laghrour - Khenchela, LESIA Laboratory, BP 1252 El Houria, 40004, Algeria*

Received ## Mon. 20##, Revised ## Mon. 20##, Accepted ## Mon. 20##, Published ## Mon. 20##

Abstract: This paper introduces a novel generative model based on an encoder-decoder architecture for defect detection within Industry 4.0 frameworks, focusing on the escalating need for automated quality control in manufacturing settings. Precision and efficiency, crucial in such environments, are significantly enhanced by our approach. At the core of our methodology is the strategic incorporation of random Gaussian noise early in the image processing sequence. This deliberate interference disrupts the model's ability to reconstruct images of defective parts, thereby enhancing both the accuracy and robustness of defect detection.
The model further integrates skip connections during the decoding phase, with a special emphasis on the first two connections. These are augmented with multi-head attention mechanisms and spatial reduction techniques, followed by targeted convolutions. This intricate configuration helps preserve vital local features while filtering out superfluous data, facilitating precise image reconstruction and effectively addressing the often problematic issue of locality loss during the upsampling process. Moreover, our model excels in maintaining contextual integrity and capturing multi-scale features, which is crucial for detailed defect detection. Each block of the architecture connects to a scaled version of the original image, allowing for nuanced feature analysis. Extensive testing and validation on real-world datasets have proven the model's high efficiency and accuracy in identifying defects, marking a significant advancement in automated quality control systems.

Keywords: Anomaly Detection, Vision Transformer, Quality Control, Industry 4.0

1. INTRODUCTION

Advanced technologies such as robotics, artificial intelligence, machine vision, big data, cloud computing, and machine learning have revolutionized manufacturing. They have given rise to what is known as Industry 4.0, which has played a major role in the application of automated visual inspection. This has helped avoid many problems that can be caused by human inspection by using artificial vision techniques, such as cameras and capture devices, to record images and transfer them to a machine to check product quality [1][2][3][4].

The quality of industrial products is defined by their compliance with established standards. Any defect impacting product quality means it has not met the required standards, leading to potential issues such as safety risks, breakdowns, material damage, or even injuries. These incidents can result in financial losses for companies and a negative reputation. This is why defect detection is fundamental in product quality control. Defect detection is the process of identifying anomalies that occur during production, such as contamination, scratches, cracks, color changes, etc. Computer vision is one of the most widely adopted fields for this task. It involves capturing images of the product, with and without defects, and then letting the model operate until it can distinguish between the two. This produces meticulous results. Since defects can vary in different ways, annotating all types of defects becomes impossible due to the time required.

This has prompted researchers to focus on unsupervised learning. Some have explored methods based on feature integration. The fundamental idea of this approach is to generate, during training, a significant vector space to represent normal data. During the testing phase, results are compared to this vector to classify whether it is a defect or not. Conversely, other researchers have opted for reconstruction approaches. The main idea behind reconstruction is to train the model exclusively on images without defects. Although this approach creates divergences when processing images containing defects, these differences effectively reveal the presence of defects during the testing phase. Most work has

E-mail address:khedidja.mekhilef@univ-khenchela.dz,abbas_faycal@univ-khenchela.dz

adopted the reconstruction approach based on convolutional layers, incorporating architectures such as Autoencoders Networks [5] [6], GAN (Generative Adversarial Networks) [7] [8], and Variational Autoencoders [9] [10] [11]. However, the major drawback of these convolutional layers lies in their excessive focus on locality, thereby limiting their explicit modeling of long-term dependencies. This limitation results in often imperfect reconstruction, even for non-defective images during the testing phase, thus compromising accurate defect detection.

With the emergence of the Vision Transformer architecture [12], inspired by the Natural Language Processing (NLP) model [13]known for efficiently modeling long-term dependencies, several research studies have been encouraged to adopt this architecture in reconstruction based methods for defect detection. Some have even substituted the autoencoder's encoder with a transformer [14] [15], while others have explored using it to create a self-attention-based autoencoder for feature reconstruction [16]. However, despite successful modeling of the global context by this architecture, its use has sometimes led to a lack of locality [17]. Mathian et al. [18] aimed to combine locality and globality by using an autoencoder composed of a sequence of a convolutional layer followed by a self-attention mechanism. However, this raises concerns about the quality of the extracted locality, as the exclusive use of convolutional layers may require a well-defined sequence for efficient extraction. Considering the inherent visual complexity of images, characterized by intricate patterns and details, accurate reconstruction of images or features requires considering both global and local information. Nevertheless, a challenge persists in the context of reconstruction-based methods during the testing phase, where the presence of defective images can lead to the reconstruction of defects, thereby complicating the precise detection and localization of anomalies.

In this paper, to fully leverage the complementarity of local and global features, an encoder-decoder architecture is proposed. The initial layers of the encoder capture texture features, while the final layers focus more on semantic features. To restore the image from the extracted features, the decoder applies a set of upsampling and convolution operations. However, during the decoding phase where upsampling operations are performed, there can be a loss of locality. To enable the decoder to fully utilize this information for precise image reconstruction, inspired by U-Net [19], it is integrated with skip connections, the first two of which are combined with multi-head attention, followed by spatial reduction inspired by [20], and then convolution aimed at retaining only specific local features and eliminating those that are not necessary. To maintain the integrity of contextual information on one hand and capture features at different spatial scales on the other, each block is associated with an equivalent representation of the original image, but at a reduced scale. To prevent the problem of reconstructing the defect and hinder the reconstruction of the defective part from random Gaussian noise, the latter is added at the beginning of the image. In addition to this, to enrich the dataset dedicated to defect detection and localization, a new class of data is created.

2. Related Work

A. Methods Based on Reconstruction

Since the database contains only non-defective images, some research has explored the effectiveness of CNNs in reconstruction methods. Bergmann et al.[5] introduced structural similarity as a metric, replacing the simple pixel difference (L2) in their approach. Yang et al.[21] introduced the concept of multi-sequence by combining model blocks at different scales. Zavrtanik et al. [22] proposed image inpainting, masking specific portions in the images to prompt the model to reconstruct the defective parts as if they were non-defective. Zhou et al.[23] Based their approach on the difference between the structural information of the original image and the reconstructed image to detect defects. Li et al.[24]introduced the concept of superpixels to divide the image into regions, then masked these regions randomly to prevent the reconstruction of defects in the test portion. Hou et al.[6] introduced the concept of multi-scale block-wise memory in autoencoders to maximize the difference between the reconstruction of defective and non-defective images.

Other works have explored the potential use of transformers to enhance data representation. Lee et al.[15] introduced the transformer as an encoder for the CNN autoencoder. De et al.[25] applied masking to hide information, focusing particularly on the masking of patches inside blocks. You et al.[26] introduced the transformer into a method based on feature reconstruction. Mishra et al.[14] introduced a Gaussian mixture density network to model the distribution of representative vectors generated by the encoder of the Vision Transformer in the context of defect detection and localization.

B. Feature Integration based Methods

To improve the performance of unsupervised methods, some approaches strive to incorporate the idea of a representative vector or vector space.[27][28][29] Create a hypersphere space during training by minimizing the distance between normal points and the center of the hypersphere. During the test phase, if the distance is no longer close or identical, the instance is considered defective. [30][31] opt for the use of normalized vectors through distribution estimation methods. In the test part, if the distance between the normal and observed distribution is higher, the instance is considered defective.[32] follows an approach where the teacher is considered as a reference vector. During training, a student model tries to adapt to this teacher. During the test phase, if the student fails to mimic the teacher, the instance is considered defective.

3. Method

A. Feature Extraction

Pre-trained CNNs are recognized as being among the most effective models for producing discriminative features

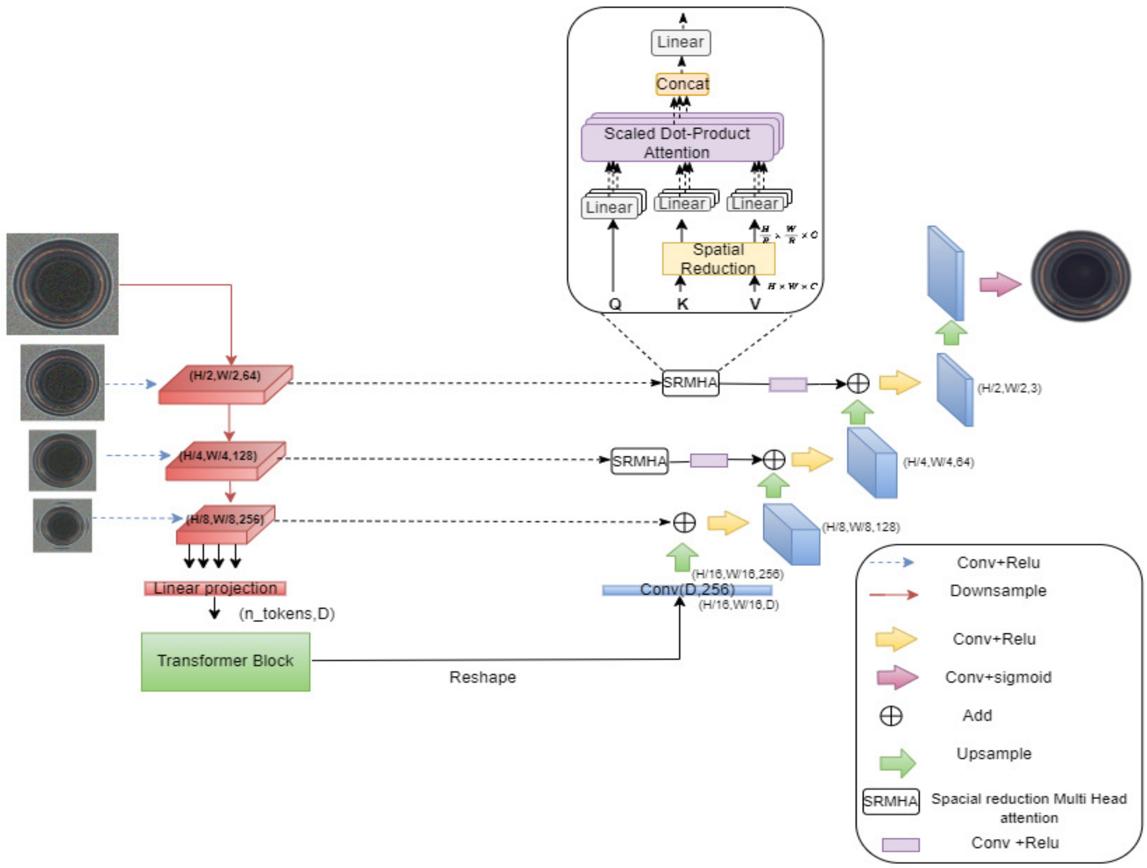

Figure 1. The overall architecture of our model.

that have a significant impact in tasks of defect detection and localization[21], as shown in Figure 1. To enable the model to model local information, the first three blocks of the vgg19 Network pre-trained on the ImageNet dataset are used, the first and second are designed for the extraction of texture features, while the third acts as an intermediary between texture and semantic information.

$f \in \mathbb{R}^{H \times W \times C}$ The feature map of the last block such that C and $H \times W$ indicate the channel and the spatial dimension of the feature map, respectively. As the vision transformer process unfolds, initially the feature map is divided into a set of tokens $N = \frac{HW}{P^2}$ where $P \times P$ represents the resolution of each token, then these tokens are linearly projected into latent vectors of size D combined with position encoding to restore the information to its position before they are introduced to the transformer block to model global information as the permutation is invariant. Regarding The transformer block, it follows the structure of the classic architecture that appears in Figure 2, the encoding passes through a multi-head attention in the first sub-block and a forward propagation in the second sub-block, and normalization and residual connection in both sub-blocks.

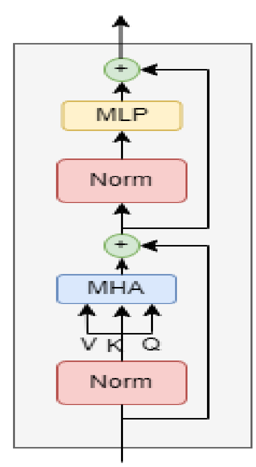

Figure 2. Block transformer components

B. Image Reconstruction

In order to reconstruct the image and decipher hidden features, we begin by reshaping the dimensions of the output of the hybrid encoder,changing from $N \times D$ to $n_1 \times n_2 \times D$, where $n_1 = \frac{H}{P}$ and $n_2 = \frac{W}{P}$. This is followed by the application of a convolutional layer to restore the original dimensions, and then a series of upsampling operations with a magnification factor of 2 to enhance spatial resolution, and *conv3x3* for extracting more complex features. A Relu activation is also applied. In addition, a sigmoid function is used on the final results to normalize the values between 0 and 1. The structure of this proposed model, designed according to an encoder-decoder scheme, naturally allows for the integration of 'skip' type connections between the encoder and decoder. These connections are crucial for effectively associating high-resolution local information with low-resolution global information.

As the encoder becomes more complex, the processed information becomes more global and elaborate, which can however lead to a loss of local details during the decoding process. This particularly affects the reconstruction of objects with variable structures and complex patterns. To address this problem, multi-head attention combined with spatial reduction is integrated at the level of the first and second residual connection. This approach allows for a significant improvement in the ability to weigh and integrate texture information across the entire image, thereby enhancing the representation of relevant features in the overall context of the scene. This is followed by a *conv3x3* and a Relu activation function to accentuate local details.

The multi-head attention mechanism (MHA), designed to identify distant interdependencies, operates as follows: the linear projections of keys (K), queries (Q), and values (V), all of which have the same dimension, are distributed across multiple heads. In each head, a multiplication is performed between the keys and queries, after which a softmax function is applied to the result of this multiplication. The resulting output is then adjusted by multiplying it with the corresponding values. This process can be expressed in the following way.

$$Attention(Q, K, V) = Softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{1}$$

The first and second blocks create a high-resolution feature map, whose integration into MHA increases computational load and memory usage. The implementation of Spatial Reduction for multi-head attention involves adjusting the dimension of the keys and values via a spatial reduction R, before proceeding to the attention operation 1.

$$K_{reduced} = Reshape\left(\frac{HW}{R^2}, C \cdot R^2\right) W\left(C \cdot R^2, C\right) \tag{2}$$

$$V_{reduced} = K_{reduced} = Norm\left(K_{reduced}\right) \tag{3}$$

W refers to a linear projection designed to preserve channel dimensions, while Norm refers to the normalization layer. To enrich the information represented by the three blocks of the pre-trained CNN and to address issues related to resolution reduction and capturing features at various spatial scales, additional features are introduced for each block. These features are generated from a sequence that includes a 2×2 average pooling operation, a 2×2 convolution, followed by activation using the Relu function on the original image. Careful restoration of visual data can sometimes lead to the reappearance of defects during the testing phase. To avoid this, a proactive approach has been implemented: the deliberate introduction of a random Gaussian disturbance in the input image. This disturbance is designed to mask certain information while preserving the overall quality of the reconstruction. The formula used for integrating this noise is as follows:

$$X_{noisy} = X + \lambda \quad where, \quad \lambda \sim N\left(0, \sigma^2\right) \tag{4}$$

As X represents the input image and σ is the maximum standard deviation of the Gaussian noise added to the input image.

C. The Loss Function

During the training phase, a loss function was used that combines both the pixel-focused L_2 method and the SSIM [5]. The pixel-focused L_2 method is used to assess the error in the value of each corresponding pixel, while the SSIM is employed to judge the brightness, determined by the average value of all the pixels, the contrast, which refers to how intensities are distributed within the image and are calculated using the standard deviation, and the structural similarity, indicating the correlation between the two images and measured by the divergence in the intensity directions of the two images.

$$SSIM(X, \hat{X}) = \frac{(2\mu_x\mu_{\hat{x}} + C_1) + (2\sigma_{x\hat{x}} + C_2)}{(\mu_x^2 + \mu_{\hat{x}}^2 + C_1) + (\sigma_x^2 + \sigma_{\hat{x}}^2 + C_2)} \tag{5}$$

$$Loss_T = L_2(X, \hat{X}) + SSIM(X, \hat{X}) \tag{6}$$

Where,

- X the original image.
- $\hat{X}$ the reconstructed image.
- μ_x the average sample of the image X.
- $\mu_{\hat{x}}$ the average sample of the image $\hat{X}$.
- σ_x^2 the variance of X.
- $\sigma_{\hat{x}}^2$ the variance of $\hat{X}$.
- $\sigma_{x\hat{x}}$ the covariance of X and $\hat{X}$.
- $C_1 = (K_1L)^2$ and $C_2 = (K_2L)^2$ are two variables to stabilize the division with a weak denominator.
- L the dynamic range of pixel values (typically it's $2^{bitsperpixel} - 1$).
- $K_1 = 0.01$ and$K_2 = 0.03$ by default.

During the test phase and to evaluate the performance of our model, we use the multi-scale gradient magnitude similarity method (MSGMS)[22], a multi-scale extension of GMS[33], to evaluate the similarity of structure and contours, in conjunction with the L2 loss to calculate the pixel-to-pixel difference between the values. This method allows us to estimate the anomaly score between the reconstructed image and the original image. The function for calculating this anomaly score is presented as follows:

$$A_{score} = (1_{H\times W} - MSGMS(X, \hat{X})Conv_f) + L_2(X, \hat{X})Conv_f \quad (7)$$

The anomaly score is obtained by subtracting the anomaly map obtained from $MSGMS$ from $1_{H\times W}$, where $1_{H\times W}$ is a matrix of ones, and then adding the result to the anomaly map obtained from theL_2 loss. The anomaly maps obtained from $MSGMS$ and L_2 have been previously processed by a mean filter convolution of size 21 $Conv_f$.

$A_{score}$is a matrix representing the anomaly score of each pixel. To calculate the score for the entire image, the maximum among all scores is taken into account.

4. Experiments

A. Data Sets

In the context of our study, we used two datasets to assess the effectiveness and accuracy of our model for defect detection.

The first is a dataset consisting of real images of buttons that we created. This set consists of 173 images, divided into two categories: 131 images for training and 42 for testing. Each image in this dataset has dimensions of 704 pixels in width by 708 pixels in height and is presented in RGB color format. These images were captured using a mobile phone camera, which offers a high resolution of 4032x2268 pixels. This capture method guarantees high image quality, essential for detailed and precise analysis. To enhance the effectiveness of the defect detection process, masks were generated for all images showing anomalies. These masks play a crucial role in our study, as they allow precise localization of defects on the images. This method greatly facilitates the evaluation of our model's performance in terms of detection and localization of defects on button images.

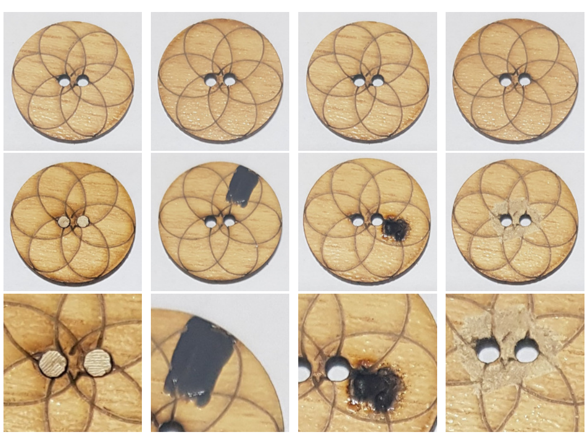

Figure 3. Non-defective samples (first row) and Defective samples (second row) with Defect overview (last row)

The second dataset is MVTec Anomaly Detection (MVTec AD) [34], a diverse and specialized dataset, essential for evaluating the effectiveness of anomaly detection methods in unsupervised machine learning. This dataset consists of fifteen distinct industrial categories, including Five different texture kinds and ten different object categories This variety allows a comprehensive and exhaustive evaluation of anomaly detection models, offering a wide range of possible scenarios and use cases. For each category, MVTec AD provides two distinct sets of images: one for training and another for testing. The training images are carefully selected to present no defects, thus ensuring that the models learn from examples without anomalies. In contrast, the test set consists of both defective and non-defective images. This composition is crucial for testing the models' ability to distinguish anomalies from normal conditions in real environments. A particularly relevant aspect of MVTec AD is the inclusion of annotated masks for each defective image in the test set. These masks provide precise information on the location and nature of defects in the images. Using these masks not only allows for assessing whether a model can detect an anomaly but also measures its accuracy in locating and characterizing specific defects.

B. Implementation Details

At the beginning of the process, Before feature extraction begins, images are first scaled to 224 pixels. Then, the parameters of the transformer block head and the multi-head attention for the second level and the first level of the skip connection are set to 4, 2, and 1, respectively. Furthermore, the encoding parameters of the transformer block size (D) and the multi-head attention for the second and first skip connections are fixed at 512, 128, and 64. Lastly, the spatial reduction rate for the first skip connection is set to 8, and for the second skip connection, it is fixed at 4. Dropout with a value of 0.25 is applied in both the MLP and the attention blocks of the transformer. The model is run with the Adam optimizer with a learning rate equal to 0.0001. The dataset is divided into 80% for processing and 20% for validation with batch sizes of 8. The model was trained for 2000 epochs, with an early stopping mechanism activated from epoch 800. This mechanism ends the training if the validation loss shows no improvement for 300 consecutive epochs. It is important to note that the training loss and validation loss remain very close in value, with minimal difference. This suggests that the model generalizes well to the validation data and shows no significant signs of overfitting. Regarding the noise rate, each class is run and evaluated individually and independently of other categories. We ran each class with a different noise rate to select the optimal rate offering the best performance. The noise rates chosen for each class are as follows: 0.1 for 'bottle', 0.25 for 'cable', 0.2 for 'capsule', 0.2 for 'hazelnut', 0.3 for 'metal nut', 0.09 for 'pill', 0.2 for 'screw', 0.3 for 'toothbrush', 0.3 for 'transistor', 0.1 for 'zipper', 0.4 for 'carpet', 0.16 for 'grid', 0.1 for 'leather', 0.11 for 'tile', 0.09 for 'wood', and 0.09 for the new class 'constructed button'.The network was implemented in PyTorch.

TABLE I. Comparison of pixel-level detection on the MVTec AD dataset.

Class	AnoGAN[7]	SMAI[24]	KDAD[32]	HaloAE[18]	FCDD[29]	OUR
Bottle	86	86	96.3	91.9	**97**	94.1
Cable	78	**92**	82.4	87.6	90	87.1
Capsule	84	93	95.9	**97.8**	93	**97.8**
Hazelnut	87	97	94.6	**97.8**	95	97.7
Metal nut	76	92	86.4	85.2	**94**	90.6
Pill	87	92	89.6	91.5	81	**98.6**
Screw	80	96	96.0	**99.0**	86	97.9
Toothbrush	90	96	96.1	92.9	94	**99.1**
Transistor	80	85	76.5	87.5	**88**	87.9
Zipper	78	90	93.9	96.0	92	**96.4**
$Mean_{obj}$	82.6	91.9	90.7	92.7	91	**94.7**
Carpet	54	88	95.6	89.4	**96**	85.6
Grid	58	97	91.8	83.1	91	**97.6**
Leather	64	86	98.1	98.5	98	**99.4**
Tile	50	62	82.8	78.5	91	**95**
Wood	62	80	84.8	**91.1**	88	84
$Mean_{tex}$	57.6	82.6	90.6	88.1	**92.8**	92.3
Mean	74	89	90.7	91.2	92	**93.9**

TABLE II. Comparison of image-level results on the MVTec AD dataset.

Class	Ganomaly[8]	AnoVit [15]	KDAD[32]	DAAD [6]	OUR
Bottle	89.2	83	**99.4**	97.6	**99.4**
Cable	75.5	74	**89.2**	84.4	79
Capsule	73.2	73	80.5	76.7	**82.7**
Hazelnut	78.5	88	**98.4**	92.1	96.6
Metal nut	70.0	**86**	73.6	75.8	82.6
Pill	74.3	72	82.7	90.0	**90.8**
Screw	74.6	**100**	83.3	98.7	89.8
Toothbrush	65.3	74	92.2	99.2	**99.7**
Transistor	79.2	83	85.6	87.6	**95.4**
Zipper	74.5	73	93.2	85.9	**98**
$Mean_{obj}$	75.4	80.6	87.8	88.8	**91.4**
Carpet	69.9	50	79.3	**86.6**	57.2
Grid	70.8	52	78.0	**95.7**	94.7
Leather	84.2	85	95.1	86.2	**100**
Tile	79.4	89	91.6	88.2	**99.4**
Wood	83.4	95	94.3	**98.2**	96.2
$Mean_{tex}$	77.5	74.2	87.7	**91**	89.5
Mean	76.2	78	87.7	89.5	**90.8**

C. Results and Discussion

To evaluate the effectiveness of our model in detecting and locating defects, we undertook an extensive comparison of our results with those obtained by current state-of-the-art methods in this field. Our analysis focused on various approaches, including knowledge distillation with KDAD[32], various reconstruction methods such as SMAI[24], AnoGAN[7], and HaloAE[18], as well as one-class classification techniques like FCDD[29], specifically for defect localization. Additionally, we also examined recognized defect detection methods, including Ganomaly[8], KDAD[32], AnoViT[15], and DAAD[6]. This comparative analysis allowed us to position our model on current standards in the field and to evaluate its performance in a quantifiable manner. To measure the effectiveness of these different methods, including ours, we opted for the use of the evaluation matrix of the area under the curve (AUC) of the receiver operating characteristics (ROC). This metric is widely recognized for its ability to provide a reliable and comprehensive evaluation of binary classification model performance, taking into account both the sensitivity and specificity of the model.

Table I illustrates the performance achieved using the area under the ROC curve (AUC) for receiver operating characteristics at the pixel level. By comparing our model with some leading models in MVTec AD, one using the transformer (HaloAE) and the other convolutions, our model stands out markedly in 6 categories, with a lead ranging from 0.4% to 7.1%. This remarkable superiority is also observed in the overall average of all object categories, where our model exceeds other methods by 2%. It is worth noting that these categories represent 67% of the total data. Taking into account the overall average for all data categories, the performance of our method exceeds those of other compared methods by 1.9%.

Table II presents a comparison of image-level detection results on the MVTec AD dataset, revealing that our model surpasses other models in eight categories. In addition, the average of all categories for our model is higher than the total average of other approaches, Despite a negative impact observed in the 'carpet' category, these results highlight our model's ability to effectively detect defects.

Table III which presents the results obtained with the AUC of the ROC features in comparing our model with the state-of-the-art KDAD on the constructed data class, evaluated the results of the latter since the class we created was not included in the original article. The results reveal that our model beats the KDAD score in the button category by 6.8%. Thus, combining the results of Tables 1 and 2, our model outperforms KDAD in 13 of the 16 categories at the pixel level and offers almost the best results at the image level. It should be noted that KDAD employs the Teacher-Student mechanism, where the Teacher is pre-trained on ImageNet, while our model is based solely on the initial pre-trained layers.

TABLE III. Comparison results for the constructed class 'Button'.

Model	pixel-level	image-level
OUR	**97.4**	**99.7**
KDAD[32]	90.6	99.5

Figures 4, 5, and 6 represent the evaluation of the model's performance in terms of the visual localization of defects. In Figure 4, we focus on the constructed data class, while Figures 5 and 6 are concerned with the MVTEC AD dataset. Each line in the figures presents six columns: the first and fourth columns show the input image, the second and fifth columns display the segmentation mask, and the third and sixth columns represent the anomaly scores, where the red color indicates a high anomaly score. These representations demonstrate the proposed model's ability to localize defects, whether their size is small, medium, or large. It is notable that the classes where the model excels in terms of score also demonstrate excellent visual localization of defects, as in the classes of toothbrush, leather, capsule, and screw, even for very small size compared to other classes such as carpet, and wood, the model manages to provide accurate localization of anomalies, thus demonstrating the robustness of its approach.

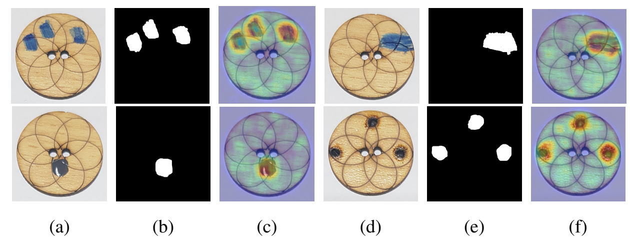

Figure 4. Qualitative results of our model on the 'Button' class. Rows a and d: Input images, b and e: Ground truth, c and f: Anomaly maps.

5. Conclusion

In this work, we have developed an innovative architecture combining convolutional neural networks (CNNs) and transformers to leverage their respective strengths in extracting local and global features. Our encoder-decoder architecture is distinguished by the integration of CNN blocks, pre-trained to capture fine and local details, in the early layers of the encoder. This approach is complemented by the use of transformers in the final layers to capture and integrate information on a broader scale. The skip connections from the encoder to the decoder, especially the first two which are reinforced by multi-head attention and spatial reduction followed by a convolutional operation, play a crucial role in effectively weighting the features relevant to the decoding task. Moreover, the introduction of random Gaussian noise upstream of the image contributes to preventing the reconstruction of defects, which is a significant step towards model robustness. Establishing a specific data class represents a notable advancement in our research methodology.

Our future work will mainly focus on improving the model's performance in classes that currently have a negative impact on our results. This approach will involve a thorough analysis of these specific categories to identify challenges and obstacles that hinder their effective processing. We will consider integrating new deep-learning techniques and optimizing the architecture to refine the model's ability to handle more complex or unconventional cases. Additionally, we will explore the effectiveness of different types of noise or regularization techniques to further enhance the model's ability to generalize and avoid overfitting, particularly in scenarios where data is limited or highly specific. The ultimate goal of these future works will be to improve the robustness and accuracy of the model, making it more effective and adaptable to various practical applications.

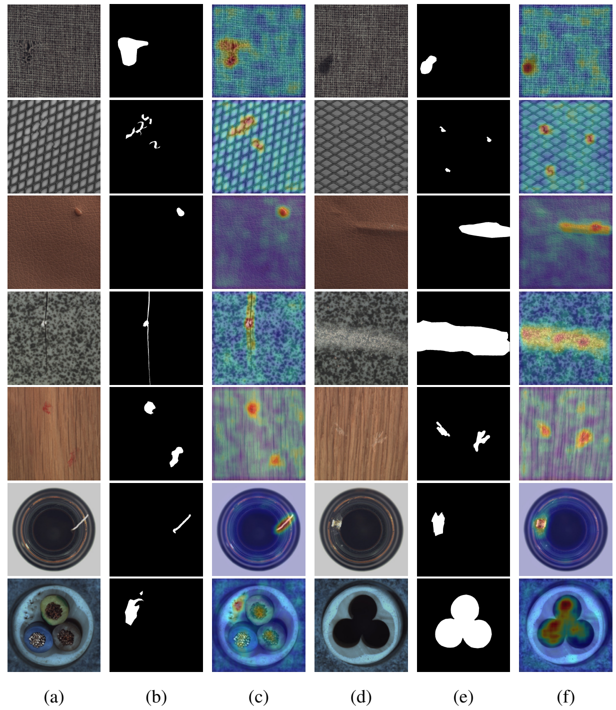

Figure 5. Qualitative results of our model on 7 out of the 15 classes in the MVTEC AD database .Rows a and d: Input images, b and e: Ground truth, c and f: Anomaly maps.

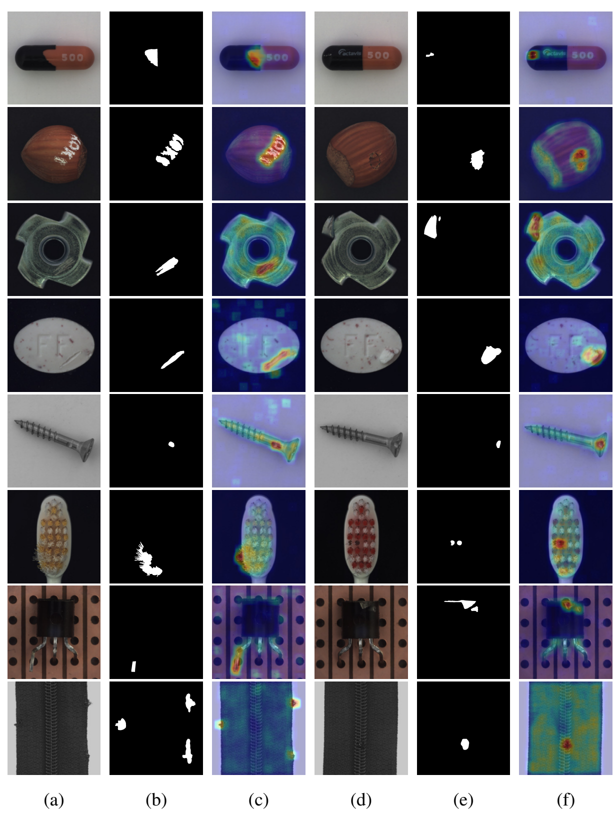

Figure 6. Qualitative results of our model on 8 out of the 15 classes in the MVTEC AD database.Rows a and d: Input images, b, and e: Ground truth, c, and f: Anomaly maps.

References

[1] M. Javaid, A. Haleem, R. P. Singh, S. Rab, and R. Suman, "Exploring impact and features of machine vision for progressive industry 4.0 culture," *Sensors International*, vol. 3. Elsevier, Jan 2022.

[2] F. K. Konstantinidis, S. G. Mouroutsos, and A. Gasteratos, "The role of machine vision in industry 4.0: an automotive manufacturing perspective," in *2021 IEEE International Conference on Imaging Systems and Techniques (IST)*. IEEE, 2021, pp. 1–6.

[3] M. Javaid, A. Haleem, R. P. Singh, S. Rab, and R. Suman, "Artificial intelligence applications for industry 4.0: A literature-based study," *Journal of Industrial Integration and Management*, vol. 7, no. 1, pp. 83–111. World Scientific, Mar 2022.

[4] R. Rai, M. K. Tiwari, D. Ivanov, and A. Dolgui, "Machine learning in manufacturing and industry 4.0 applications," *International Journal of Production Research*, vol. 59, no. 1, pp. 4773–4778. Taylor & Francis, Aug 2021.

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

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

[7] T. Schlegl, P. Seeböck, S. M. Waldstein, U. Schmidt-Erfurth, and G. Langs, "Unsupervised anomaly detection with generative adversarial networks to guide marker discovery," in *International Conference on Information Processing in Medical Imaging*. Springer International Publishing, May 2017, pp. 146–157.

[8] S. Akcay, A. Atapour-Abarghouei, and T. P. Breckon, "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*, vol. 14. Springer International Publishing, 2019, pp. 622–637.

[9] D. Zimmerer, S. A. Kohl, J. Petersen, F. Isensee, and K. H. Maier-Hein, "Context-encoding variational autoencoder for unsupervised anomaly detection," *arXiv preprint arXiv:1812.05941*, 2018.

[10] Y. Lu and P. Xu, "Anomaly detection for skin disease images using variational autoencoder," *arXiv preprint arXiv:1807.01349*, 2018.

[11] A. A. Pol, V. Berger, C. Germain, G. Cerminara, and M. Pierini, "Anomaly detection with conditional variational autoencoders," in *2019 18th IEEE international conference on machine learning and applications (ICMLA)*. IEEE, Dec 2019, pp. 1651–1657.

[12] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[13] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

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

[15] Y. Lee and K. Pilsung, "Anovit: Unsupervised anomaly detection and localization with vision transformer-based encoder-decoder," *IEEE Access*, vol. 10, pp. 46 717–46 724, May 2022.

[16] Y. Yang, "Self-attention autoencoder for anomaly segmentation," 2021.

[17] H. Wu, B. Xiao, N. Codella, M. Liu, X. Dai, L. Yuan, and L. Zhang, "Cvt: Introducing convolutions to vision transformers," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 22–31.

[18] E. Mathian, H. Liu, L. Fernandez-Cuesta, D. Samaras, M. Foll, and L. Chen, "Haloae: An halonet based local transformer autoencoder for anomaly detection and localization," *arXiv preprint arXiv:2208.03486*, Aug 2022.

[19] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III*, vol. 18. Springer International Publishing, 2015, pp. 234–241.

[20] W. Wang, E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo, and L. Shao, "Pyramid vision transformer: A versatile backbone for dense prediction without convolutions," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 568–578.

[21] J. Yang, Y. Shi, and Z. Qi, "Dfr: Deep feature reconstruction for unsupervised anomaly segmentation," *arXiv preprint arXiv:2012.07122*, Dec 2020.

[22] V. Zavrtanik, M. Kristan, and D. Skočaj, "Reconstruction by inpainting for visual anomaly detection," *Pattern Recognition*, vol. 112, p. 107706, Apr 2021.

[23] K. Zhou, Y. Xiao, J. Yang, J. Cheng, W. Liu, W. Luo, Z. Gu, J. Liu, and S. Gao, "Encoding structure-texture relation with p-net for anomaly detection in retinal images," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XX*, vol. 16. Springer International Publishing, 2020, pp. 360–377.

[24] Z. Li, N. Li, K. Jiang, Z. Ma, X. Wei, X. Hong, and Y. Gong, "Superpixel masking and inpainting for self-supervised anomaly detection," *In Bmvc*, Sep 2020.

[25] A. De Nardin, P. Mishra, G. Luca Foresti, and C. Piciarelli, "Masked transformer for image anomaly localization," *International Journal of Neural Systems*, vol. 32, no. 07, July 2022.

[26] Z. You, K. Yang, W. Luo, C. Lei, Y. Zheng, and X. Le, "Adtr: Anomaly detection transformer with feature reconstruction," in *International Conference on Neural Information Processing*. Springer International Publishing, Nov 2022, pp. 298–310.

[27] J. Yi and Y. Sungroh, "Patch svdd: Patch-level svdd for anomaly detection and segmentation," in *Proceedings of the Asian conference on computer vision*, 2020.

[28] F. V. Massoli, F. Falchi, A. Kantarci, A. Şeymanur, H. K. Ekenel, and G. Amato, "Mocca: Multilayer one-class classification for anomaly detection," *IEEE Transactions on neural networks and learning systems*, vol. 33, pp. 2313–2323, Dec 2021.

[29] P. Liznerski, R. A. Ruff, Lukas Vandermeulen, B. J. Franks, M. Kloft, and K.-R. Müller, "Explainable deep one-class classification," in *In International Conference on Learning Representations*, Jul 2020.

[30] O. Rippel, A. Chavan, C. Lei, and D. Merhof, "Transfer learning gaussian anomaly detection by fine-tuning representations," *arXiv preprint arXiv:2108.04116*, Aug 2021.

[31] O. Rippel, P. Mertens, and D. Merhof, "Modeling the distribution of normal data in pre-trained deep features for anomaly detection," in *2020 25th International Conference on Pattern Recognition (ICPR)*. IEEE, Jan 2021, pp. 6726–6733.

[32] M. Salehi, N. Sadjadi, S. Baselizadeh, M. H. Rohban, and H. R. Rabiee, "Multiresolution knowledge distillation for anomaly detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 14 902–14 912.

[33] W. Xue, L. Zhang, X. Mou, and A. C. Bovik, "Gradient magnitude similarity deviation: A highly efficient perceptual image quality index," *IEEE transactions on image processing*, vol. 23, no. 2, pp. 684–695, Dec 2013.

[34] P. Bergmann, M. Fauser, D. Sattlegger, and C. Steger, "Mvtec ad–a comprehensive real-world dataset for unsupervised anomaly

detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 9592–9600.

Khedidja mekhilef is currently a Ph.D. student of Computer Science from Abbes laghrour khenchela, Algeria. she obtained a master's degree from the University of Constantine 2 - Abdelhamid Mehri, Algeria, in 2021. her research interests focus on image processing,deep learning, and Quality Control.

Fayçal abbas is an associate professor in the department of computer science at the university of khenchela Algeria and a researcher at LESIA Laboratory. He received his Ph.D. in computer Science in image synthesis and artificial life from Biskra University Algeria in 2018. His research interests are realistic rendering, deep learning, and computer vision.

Mounir Hemam is an Associate Professor at Khenchela University (Algeria), member of the ICOSI Laboratory. He has completed PhD degree in Computer sciences from Mentouri University in January 2012. He received his habilitation qualifications from Mohamed Khider University in 2017. He mainly worked on knowledge representation and reasoning; formal knowledge representation for Semantic Web, and Semantic IoT Interoperability, he became more involved in the area of semantic sensor networks. Currently, he is working on image quality assessment and machine learning algorithms to solve issues related to Industry 4.0.