Inpainting-based Anomaly Detection System with Self-supervised Learning

Chia-Yu Lin* and Yi-Zhen Chen†
* *Department of Computer Science and Information Engineering, National Central University, Taoyuan, Taiwan*
† Department of Computer Science and Engineering, *Yuan Ze University, Taoyuan, Taiwan*
Corresponding Author's E-mail: sallylin0121@ncu.edu.tw

2024 IEEE International Conference on Industry 4.0, Artificial Intelligence, and Communications Technology (IAICT) | 979-8-3503-5346-4/24/$31.00 ©2024 IEEE | DOI: 10.1109/IAICT62357.2024.10617749

***Abstract*—Deep learning for defect detection has become a critical imperative in contemporary electronics manufacturing. We propose an inpainting-based anomaly detection system to identify defects without labeled defects. An image inpainting model, which discerns disparities between the original and restored versions of the defective image, is designed as the core of our methodology. To further address issues related to reconstructing asymmetric images with defects, we incorporate self-supervised learning (SSL) to extract a broader spectrum of features. In experiments, we compare the proposed method to state-of-the-art models based on MVTec open dataset. Our proposed method can achieve a best performance of 97%, and surpasses the SOTA model by a margin of 57%.**

***Index Terms*—Defect detection, image inpainting, self-supervised learning**

I. Introduction

In electronics manufacturing, defect detection is a critical step in ensuring product quality and reliability. Traditional inspection methods rely on manual visual inspection. However, in addition to incurring high labor costs and showing low efficiency, manual inspection can also suffer from inconsistent inspection standards. Another approach involves the use of an AOI machine for defect determination, yet the accuracy of AOI detection is low, and it comes with an extremely high false alarm rate, necessitating a substantial amount of manual re-inspection. Therefore, deep learning becomes the solution to efficient defect detection.

Many supervised and unsupervised defect detection models have been proposed. Supervised classification models extract defect features to distinguish defective from normal data, which demands abundant labeled data for defect learning [1], [2]. As a result, we have chosen an unsupervised learning approach. Among the unsupervised methods for the determination of defects, the commonly used technique is the generative adversarial network (GAN) [3]. These methods leverage the GAN architecture for image defect detection. They mimic the characteristics of the normal class and determine defects by reconstructing them based on differences between normal images and original images [4], [5]. Another notable technique is image inpainting, which involves reconstructing the mask area by repairing the defective region through model training [6]–[8]. The image inpainting model learns to rectify it by referencing surrounding intact areas, thus restoring the image. Moreover, image inpainting not only corrects localized defects but also preserves the overall semantic coherence of the image, even in complex scenes.

Current image inpainting models excel at recovering masked regions and accurately identifying defect classes by contrasting them with the original image. However, predetermined repair zones in numerous defective images often exhibit notable dissimilarity from their adjacent surroundings, featuring asymmetry and incompleteness of the image. This poses a challenge for the inpainting model in effectively repairing such defective areas.

In this paper, we propose an inpainting-based anomaly detection system with self-supervised learning, where the defective part of the image mask is used in conjunction with surrounding points to reconstruct the defective image. In addition, due to the potentially complex and asymmetric nature of the data backgrounds, we incorporate self-supervised learning (SSL) to extract a broader spectrum of features. This amalgamation serves to facilitate a more comprehensive defect repair and, in the following, improves the precision of defect detection. The MVTec anomaly detection open dataset [9] is used to evaluate the performance of the model for verification.

The results of our proposed method, especially in terms of recall, can reach a best performance of 97%. This performance surpasses the ShiftNet [10] model by a margin of 57%.

II. Related Work

Within anomaly detection methods, a prevalent approach involves training Generative Adversarial Networks (GANs) on normal images [4], [5], [11]. Using normal images during training, the model learns to reconstruct normal images. When presented with a defect image, the defective portion is removed. The discrepancy between the input defect and the reconstructed image is then used to identify the defect.

In methods such as those proposed by [6]–[8], image inpainting is used for defect detection. This involves restoring the defective image to a state that resembles a normal image. Subsequently, the area surrounding the mask is utilized to repair the masked region. This transformation turns the defect into a normal image, resulting in a discernible difference from the original image. Thus, enabling the identification of defects and normal images.

Recent years have witnessed notable advances in image painting, as demonstrated by works like [7], [10], [12] that extend the U-Net architecture in various ways. Moreover, the contextual attention of transformer-based architectures has been introduced by [13], [14]. The research carried out by [8], [15]–[19] extends the application of transformers to image inpainting. These approaches focused on inpainting specific mask regions, employing adversarial training to

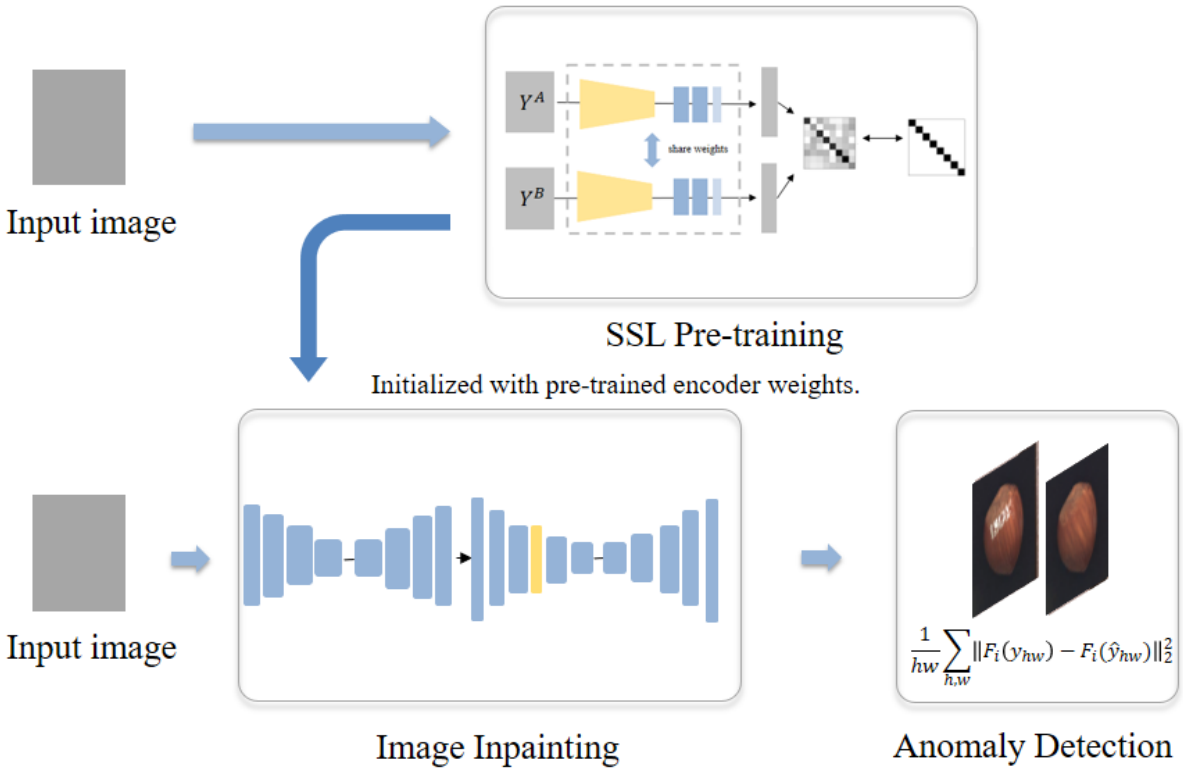

Fig. 1. The overview of the proposed framework.

enhance inpainting network performance. However, existing image-inpainting techniques still face challenges in preserving the global semantic structure and capturing intricate texture details in complex images.

Furthermore, [20] presents the application of self-supervised learning in image classification. Self-supervised learning is a powerful paradigm that utilizes unlabeled data to acquire meaningful feature representations. The study by [21] focused on predicting spatial relationships between patches in an image. This involved dividing the image into patches and selecting a pair of patches within the image. Then, one patch was used to infer the relative spatial positioning of the other patch. Similarly, [22] divides the image into patches, disrupts the order of these patches, and expects the model to recognize the composition distribution of the patches, arranging them in the correct order. Fundamentally, these methods constitute a self-supervised learning approach. The models extract information from the intrinsic characteristics of the image, eliminating the need for external annotations.

Other methods, such as [23] and [24], employed different transformations to maximize cross-correlation, capturing shared information. The integration of self-supervised learning into our system enhances its ability to extract features from vast amounts of unlabeled data. This newfound understanding contributes to improved performance in various tasks, showcasing the versatility and effectiveness of our approach.

III. Inpainting-Based Anomaly Detection System

The proposed inpainting-based anomaly detection system with self-supervised learning is shown in Fig. 1, encompassing image inpainting model, SSL pre-training, and the anomaly detection stage.

A. Image Inpainting Model

The image inpainting model is trained with normal images. The model reconstructs the images with defects to resemble a normal image state. Thus, we can compare the difference between the reconstructed and original images to identify defects. Fig. 2 and Fig. 3 visually elucidate our image inpainting architecture, which includes the rough network and the refinement network.

1) Rough Network: The rough network, employing the U-Net architecture, serves as the initial stage in reconstructing the input image I_{gt}, producing the output image I_p. It gains an advantage from pre-trained encoder weights acquired through SSL pre-training during initialization. The application of self-supervised learning in this context augments the image inpainting model's comprehension of the intrinsic data structure, enhancing its capacity to capture local features effectively.

2) Refinement Network: The refinement network improves the smoothness and coherence of the repaired pixels while improving the correlation between each repaired patch in the designated repair area. Drawing inspiration from Liu et al.'s methodologies [14], we emphasize the significance of coherent semantic attention in image inpainting tasks. This strategy revolves around establishing a coherent semantic attention layer (CSA) within the refinement network. The CSA layer identifies the most analogous pixel for each pixel in the missing region concerning the pixels in the known region, as shown in Fig. 4. The meticulous patch-by-patch repair process utilizes the most similar feature pixels

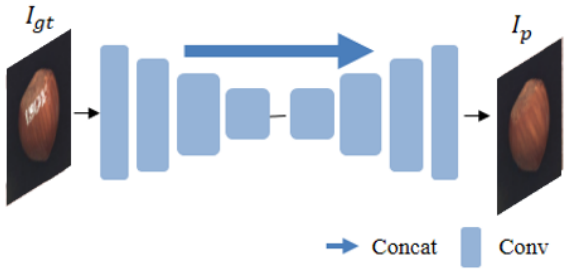

Fig. 2. Rough network architecture of image inpainting.

Fig. 3. Refinement network architecture of image inpainting.

in the surrounding area, ensuring that each repaired patch maintains high relevance to the preceding patch throughout the restoration process. The correlation among the process of evaluating the generated patches is expressed in Equations 1 and 2, which indicates the cross-correlation between each patch m_i in the set M and its corresponding patch $\bar{m}_i$ in set $\bar{M}$, effectively generating a vector of values. Within set M, the patch located in the upper left corner represents the initial value, denoted as m_{i-1}, and the next patch m_i will add the features closest to the cross-correlation of the previous patch, and so on. For each generated patch m_i, we assign it the maximum cross-correlation value D_{max_i} using the most similar patch $\bar{m}_i$. Consequently, the CSA layer plays a pivotal role in fostering meaningful connections and context awareness during inpainting. The outcome is a seamlessly integrated repaired area that blends with the surrounding content, resulting in a visually believable and coherent repaired image.

$$D_{max_i} = \frac{\langle m_i, \bar{m}_i \rangle}{\|m_i\| \|\bar{m}_i\|} \tag{1}$$

$$D_{ad_i} = \frac{\langle m_i, m_{i-1} \rangle}{\|m_i\| \|m_{i-1}\|} \tag{2}$$

3) Loss Functions: We use the reconstruction loss and consistency loss in the image inpainting model. The reconstruction loss, represented as L_r in Equation 3. We opt for the L1 distance, commonly known as the average absolute difference. This loss metric quantifies the pixel-level disparity between the inpainted image and its corresponding ground truth.

$F_v()$ denotes the extraction of high-level features from the original image using the pre-trained VGG-16 model, which has been pre-trained on ImageNet. The consistency loss, denoted as L_c in Equation 4, involves calculating the L2 distance between the features extracted by $F_v()$ and those

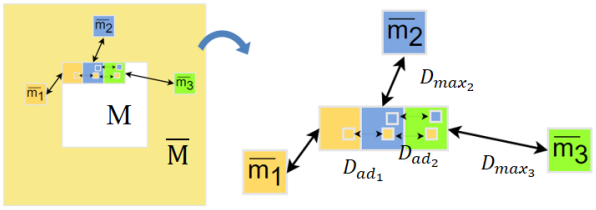

Fig. 4. The coherent semantic attention layer.

obtained from the CSA layer in the encoder (expressed as $CSA_e()$). In addition, it calculates the L2 distance between $F_v()$ and the corresponding layer of the decoder of the CSA layer (expressed as $CSA_d()$). This process contributes to enhancing the similarity to the original image.

$$L_r = |I_p - I_{gt}| + |I_r - I_{gt}| \tag{3}$$

$$L_c = \sum_{y \in M} \left(\|CSA_e(I_p)_y - F_v(I_{gt})_y\|_2^2 + \|CSA_e(I_p)_y - F_v(I_{gt})_y\|_2^2 \right) \tag{4}$$

B. SSL Pre-training Stage

Given the intricate nature and inherent asymmetry within our data, we integrate a self-supervised learning (SSL) approach to empower our model with the ability to adeptly extract and represent nuanced features. We design an initial pre-training phase, drawing inspiration from the Barlow Twins framework proposed by Zbontar et al. [23]. As shown in Fig. 5, input images are subjected to various distortion methods, generating two distorted samples from a single picture. These samples serve as inputs to two identical networks, and the objective function is computed by assessing the cross-correlation matrix between the embedding vectors produced by these networks. By this process, the model can extract intricate feature representations from unlabeled image data.

To align the U-Net encoder framework employed in the rough network, we modify Barlow Twins to U-Net architecture. The features are transformed into embedding vectors through two identical network architectures. Both architectures employ the U-Net encoder, followed by three linear layers with batch normalization after the initial two linear layers, and rectified linear units for activation.

The input image undergoes a random transformation, resulting in two distorted images labeled Y^A and Y^B. Subsequently, two sets of identical network architectures are employed to obtain the embedding vectors Z^A and Z^B. The cross-correlation matrix for these output embedding vectors is then computed.

The cross-correlation matrix C_{ij} between the two embedding vectors Z^A and Z^B is calculated using Equation 6. The loss function, denoted as L_{BT} and defined in Equation 5, aims to align the cross-correlation matrix with the identity matrix. The objective is to maximize the preservation of crucial sample features, enhancing similarities, and discarding unnecessary features.

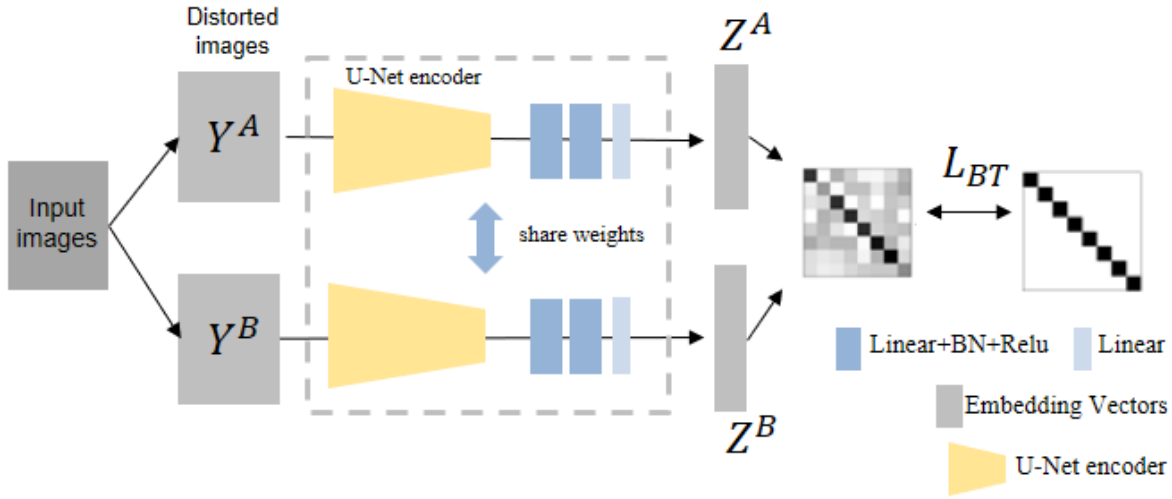

Fig. 5. The concept of the pre-training architecture.

The U-Net model in image inpainting is initialized with pre-trained encoder weights obtained during the SSL pre-training stage. This involves transferring weights from the entire encoder network to initialize the U-Net architecture during the image repair stage, providing a foundation of pre-trained weights.

Through this self-supervised learning approach, our model gains a profound understanding of the inherent data structure, thereby improving its capability to capture local features with increased precision.

$$L_{BT} := \sum_i (1 - C_{ii})^2 + \lambda \sum_i \sum_{j \neq i} C_{ij}^2 \tag{5}$$

$$C_{ij} := \frac{\sum Z_i^A Z_j^B}{\sqrt{\sum (Z_i^A)^2} \sqrt{\sum (Z_j^B)^2}} \tag{6}$$

C. Anomaly Detection

The image inpainting process involves repairing missing or damaged segments within the image. As a consequence of defect repair, the resulting image will naturally differ from the original version. Therefore, we employ a method to evaluate the defect and normal data within the input image by calculating dissimilarity metrics between the original and the reconstructed image. Subsequently, on the basis of this comparison, we identify the defect class.

In our detection formula (specified in Equation 7), we integrate the perceptual loss function [25]. This function considers the perceptual quality of the image, in order to better align with human visual perception.

To quantify the dissimilarity between two images, we utilize a pre-trained network. Specifically, we employ the pre-trained SqueezeNet [26] to capture features from the seventh layer of the network based on the provided images. The formulation, denoted as F_s, seeks to minimize the $L2$ distance in the feature space between the original input image Y_{hw} and the reconstructed image $\hat{Y_{hw}}$. Here, h and w represent the height and width of the image, respectively.

$$Detect = \frac{1}{hw} \sum_{h,w} \| F_s\left(y_{hw}\right) - F_s\left(\hat{y_{hw}}\right) \|_2^2 \tag{7}$$

IV. Experiments

To thoroughly evaluate the efficacy of our proposed model and gauge its proficiency in the classification task, we employed metrics that included the recall, the true negative rate (TNR) and computed the area under the curve (AUC).

A. Data Description

We conducted our experiments using the MVTecAD dataset [9], an open dataset designed for anomaly detection. This dataset encompasses five textures and ten object categories in various domains. Our selection of three classes: Transistor, Hazelnut, and Cable informed by the location of data defects and the background complexity within the dataset.

B. Experiments

We compare the proposed model with Skip-GANomaly [5], ShiftNet [10], and CSA [14]. The result is shown in Table I. It's noteworthy that our model excels, particularly in transistors and cables, which are marked by intricate data backgrounds. It shows a superior recall compared to other models, reaching an impressive 0.97 in the cable category. This performance underscores the efficacy and supremacy of our proposed model in effectively addressing the challenges inherent in this dataset.

Fig. 6 presents the results derived from the MVTecAD open dataset. The first column displays the original image, followed by the outputs of SkipGANomaly, ShifNet, CSA, and lastly, our proposed model. It's noteworthy that while SkipGanomaly often reconstructs the defect, other image inpainting models adeptly eliminate the defect. Notably, our model stands out by showcasing the most thorough restoration, especially evident in transistor data.

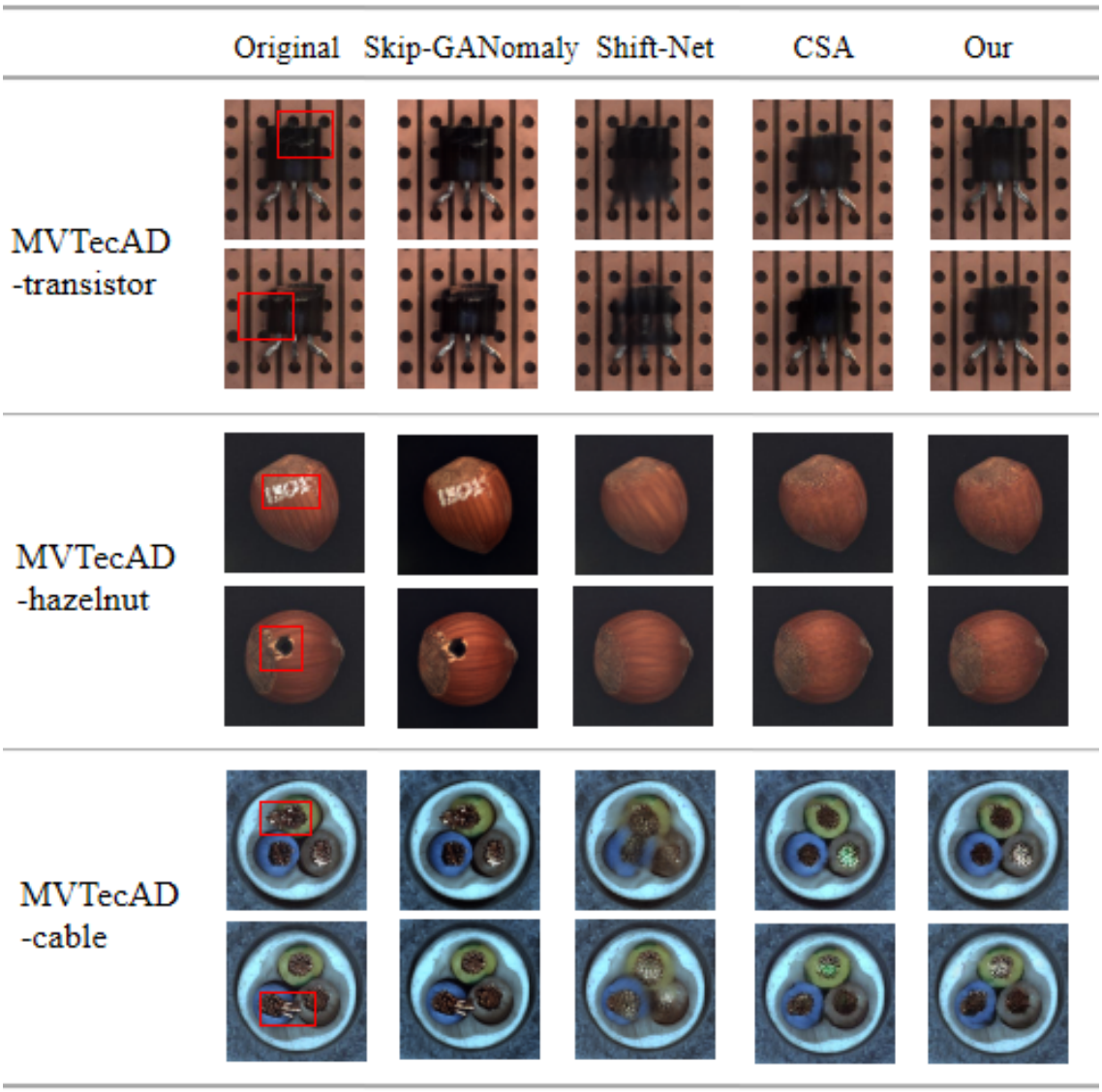

Fig. 6. Displayed here are sample results from the experimental evaluation of MVTecAD image inpainting.

TABLE I
RESULTS OF MVTECAD.

Data	*Model*	*Recall*	*TNR*	*AUC*
Transistor	Skip-GANomaly	0.65	0.58	0.65
	Shiftnet	0.55	0.59	0.63
	CSA	0.77	**0.63**	0.77
	Our	**0.85**	0.60	**0.82**
Hazelnut	Skip-GANomaly	0.45	0.58	0.50
	Shiftnet	**0.96**	**0.6**	**0.93**
	CSA	0.88	0.58	0.88
	Our	0.89	0.58	0.87
Cable	Skip-GANomaly	0.63	**0.60**	0.62
	Shiftnet	0.40	0.57	0.51
	CSA	0.95	0.58	0.92
	Our	**0.97**	0.59	**0.95**

V. CONCLUSION

We present a novel inpainting-based anomaly detection system with self-supervised learning. By seamlessly integrating self-supervised learning techniques into the image inpainting model, we elevate the model's capacity to extract pertinent image features during the inpainting process, thereby enhancing the overall inpainting outcomes. In experiments, we leverage open datasets to validate our model, and the proposed method surpasses the performance of other models. In particular, our models consistently achieve a recall exceeding 0.85, indicative of their robust and steady anomaly detection capabilities. Moreover, in the cable data, our approach attains an impressive recall of 0.97.

ACKNOWLEDGEMENTS

This work is jointly sponsored by National Science and Technology Council (NSTC) under the project NSTC 112-2622-8-A49-021 and NSTC 110-2222-E-008-008-MY3.

REFERENCES

[1] Chia-Yu Lin, Yan-Hung Chou, and Yun-Chiao Cheng, "A deep learning-based general defect detection framework for automated optical inspection," in *IEEE International Conference on Industry 4.0, Artificial Intelligence, and Communications Technology (IAICT)*, 2023, pp. 332–337.

[2] Shi-Qi Ye, Chen-Sheng Xue, Cheng-Yuan Jian, Yi-Zhen Chen, Jia-Jiun Gung, and Chia-Yu Lin, "A deep learning-based generic solder defect detection system," in *IEEE International Conference on Consumer Electronics-Taiwan*, 2022, pp. 99–100.

[3] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio, "Generative adversarial nets," *Advances in neural information processing systems*, vol. 27, 2014.

[4] Samet Akcay, Amir Atapour-Abarghouei, and Toby P Breckon, "Ganomaly: Semi-supervised anomaly detection via adversarial training," in *Asian Conference on Computer Vision*. Springer, 2019, pp. 622–637.

[5] Samet Akçay, Amir Atapour-Abarghouei, and Toby P Breckon, "Skip-ganomaly: Skip connected and adversarially trained encoder-decoder anomaly detection," in *IEEE International Joint Conference on Neural Networks (IJCNN)*, 2019, pp. 1–8.

[6] Matthias Haselmann, Dieter P Gruber, and Paul Tabatabai, "Anomaly detection using deep learning based image completion," in *IEEE international conference on machine learning and applications (ICMLA)*, 2018, pp. 1237–1242.

[7] Vitjan Zavrtanik, Matej Kristan, and Danijel Skočaj, "Reconstruction by inpainting for visual anomaly detection," *Pattern Recognition*, vol. 112, pp. 107706, 2021.

[8] Jonathan Pirnay and Keng Chai, "Inpainting transformer for anomaly detection," in *International Conference on Image Analysis and Processing*. Springer, 2022, pp. 394–406.

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

[10] Zhaoyi Yan, Xiaoming Li, Mu Li, Wangmeng Zuo, and Shiguang Shan, "Shift-net: Image inpainting via deep feature rearrangement," in *European conference on computer vision (ECCV)*, 2018, pp. 1–17.

[11] Chia-Yu Lin, Tzu-Ting Chen, Li-Chun Wang, and Hong-Han Shuai, "Health-based fault generative adversarial network for fault diagnosis in machine tools," in *The 4th International Workshop on Artificial Intelligence of Things*, 2020.

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[13] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang, "Generative image inpainting with contextual attention," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5505–5514.

[14] Hongyu Liu, Bin Jiang, Yi Xiao, and Chao Yang, "Coherent semantic attention for image inpainting," in *IEEE/CVF International Conference on Computer Vision*, 2019, pp. 4170–4179.

[15] Ye Deng, Siqi Hui, Sanping Zhou, Deyu Meng, and Jinjun Wang, "Learning contextual transformer network for image inpainting," in *ACM International Conference on Multimedia*, 2021, pp. 2529–2538.

[16] Wenbo Li, Zhe Lin, Kun Zhou, Lu Qi, Yi Wang, and Jiaya Jia, "Mat: Mask-aware transformer for large hole image inpainting," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 10758–10768.

[17] Shuyi He, Qingyong Li, Yang Liu, and Wen Wang, "Semantic segmentation of remote sensing images with self-supervised semantic-aware inpainting," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2022.

[18] Shyam Nandan Rai, Rohit Saluja, Chetan Arora, Vineeth N Balasubramanian, Anbumani Subramanian, and CV Jawahar, "Fluid: Few-shot self-supervised image deraining," in *IEEE/CVF Winter Conference on Applications of Computer Vision*, 2022, pp. 3077–3086.

[19] Daryl LX Fung, Qian Liu, Judah Zammit, Carson Kai-Sang Leung, and Pingzhao Hu, "Self-supervised deep learning model for covid-19 lung ct image segmentation highlighting putative causal relationship among age, underlying disease and covid-19," *Journal of Translational Medicine*, vol. 19, pp. 1–18, 2021.

[20] Da Chen, Yuefeng Chen, Yuhong Li, Feng Mao, Yuan He, and Hui Xue, "Self-supervised learning for few-shot image classification," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 1745–1749.

[21] Carl Doersch, Abhinav Gupta, and Alexei A Efros, "Unsupervised visual representation learning by context prediction," in *IEEE International Conference on Computer Vision*, 2015, pp. 1422–1430.

[22] Mehdi Noroozi and Paolo Favaro, "Unsupervised learning of visual representations by solving jigsaw puzzles," in *European Conference on Computer Vision*. Springer, 2016, pp. 69–84.

[23] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny, "Barlow twins: Self-supervised learning via redundancy reduction," in *International Conference on Machine Learning*. PMLR, 2021, pp. 12310–12320.

[24] Ishan Misra and Laurens van der Maaten, "Self-supervised learning of pretext-invariant representations," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 6707–6717.

[25] Justin Johnson, Alexandre Alahi, and Li Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *Proceedings of European Conference on Computer Vision*, 2016, pp. 694–711.

[26] Forrest N Iandola, Song Han, Matthew W Moskewicz, Khalid Ashraf, William J Dally, and Kurt Keutzer, "Squeezenet: Alexnet-level accuracy with 50x fewer parameters and¡ 0.5 mb model size," *arXiv preprint arXiv:1602.07360*, 2016.