Machine Learning (2024) 113:4381–4406
https://doi.org/10.1007/s10994-023-06415-9

Composite score for anomaly detection in imbalanced real-world industrial dataset

Arnaud Bougaham[1] · **Mohammed El Adoui**[1] · **Isabelle Linden**[2] · **Benoît Frénay**[1]

Received: 7 June 2022 / Revised: 22 June 2023 / Accepted: 3 October 2023 /
Published online: 20 November 2023

Abstract

In recent years, the industrial sector has evolved towards its fourth revolution. The quality control domain is particularly interested in advanced machine learning for computer vision anomaly detection. Nevertheless, several challenges have to be faced, including imbalanced datasets, the image complexity, and the zero-false-negative (ZFN) constraint to guarantee the high-quality requirement. This paper illustrates a use case for an industrial partner, where Printed Circuit Board Assembly (PCBA) images are first reconstructed with a Vector Quantized Generative Adversarial Network (VQGAN) trained on normal products. Then, several multi-level metrics are extracted on a few normal and abnormal images, highlighting anomalies through reconstruction differences. Finally, a classifier is trained to build a composite anomaly score thanks to the metrics extracted. This three-step approach is performed on the public MVTec-AD datasets and on the partner PCBA dataset, where it achieves a regular accuracy of 94.65% and 87.93% under the ZFN constraint.

Keywords Imbalanced learning · Industry 4.0 · Anomaly detection · High resolution images · Zero false negative · Computer vision · Real-world · PCBA

Editors: Nuno Moniz, Paula Branco, Luís Torgo, Nathalie Japkowicz, Michal Wozniak and Shuo Wang.

✉ Arnaud Bougaham
arnaud.bougaham@unamur.be

Mohammed El Adoui
mohammed.eladoui@unamur.be

Isabelle Linden
isabelle.linden@unamur.be

Benoît Frénay
benoit.frenay@unamur.be

[1] Faculty of Computer Science, NaDI Institute, University of Namur, Rue Grandgagnage 21, 5000 Namur, Belgium

[2] Department of Management Sciences, NaDI Institute, University of Namur, Rempart de la Vierge 8, 5000 Namur, Belgium

1 Introduction

Anomaly detection is a ubiquitous concern for industries ensuring robust manufacturing quality control (Ren et al., 2022; Wang et al., 2016). Operation sites considering this challenge need, for instance, high-resolution images of the product being manufactured so that an anomaly detection method can be executed. Usually, an automatic inspection process is devoted to perform this task. It takes several images of the same product with different view angles and lighting conditions. Then, it asks an operator to confirm or infirm a pseudo-error if a doubt on the product quality is raised (Abd Al Rahman & Mousavi, 2020). Negative samples are anomaly-free images, unlike positive ones where a defect is observed on the product image. Severe test limits are necessary to avoid missed detections (false negatives or type II errors), depending on the quality strategy. This way, the defect is early detected, and the abnormal product is not propagated towards the next production processes (Filz et al., 2020). The drawback of this strategy is a high rate of false alarms (false positives or type I errors) due to the product image complexity.

In practice, it causes fault investigation losses and product retests. Moreover, the human operator gets overflowed by the inspection process, often incorrectly considering negative products as positive. The credibility of this anomaly detection process is therefore reduced. This yields in some human misjudgments, classifying positives, well detected by the inspection process, as false negatives, because of the habit of invalidating the process pseudo-errors. Should this happens, the product is stopped later on the production line, but the repair or scrap costs are greater. If it can be repaired, the time needed to access the defect area or component increases due to all the parts composing the product. If the product has to be scrapped, the processes and the workers time needed to manufacture it is lost, as well as the components placed after this inspection process. Consequently, valuable time is wasted, repair costs are increased, and quality risks are taken (Babic et al., 2021). A key challenge is therefore to reduce the false alarms, keeping the requirement to avoid any missed detection.

This work is carried out with an industrial partner specialized in Printed Circuit Board Assembly (PCBA) for the automotive industry. These PCBAs are devoted to provide automatic car speed boxes after being sealed in a case, forming an Automatic Transmission Electronic Control Unit (ATECU). The production line is composed of 3 distinct blocks: the placement and soldering of the electronic components on the blank circuit, the connector and case assembly steps, and the final product test stage. All the processes are placed inline to manufacture the product step by step, with quality inspections alongside.

It is commonly agreed that the earlier a defect is detected, the earlier it can be contained (Jia et al., 2004). Based on this statement, the scope of this work is focused on the optical inspection, the first visual test. Its main role is to take images of 100% of the products to estimate the quality at the very first stages of the production line, where electronic components are placed and soldered (Crispin & Rankov, 2007). This critical process ensures an early reaction if needed, making it possible to countermeasure the eventual issues and avoid propagating the defect downstream. A telemetry-based image processing algorithm is currently in charge of comparing the image being treated with a golden sample image, to guarantee the anomaly detection.

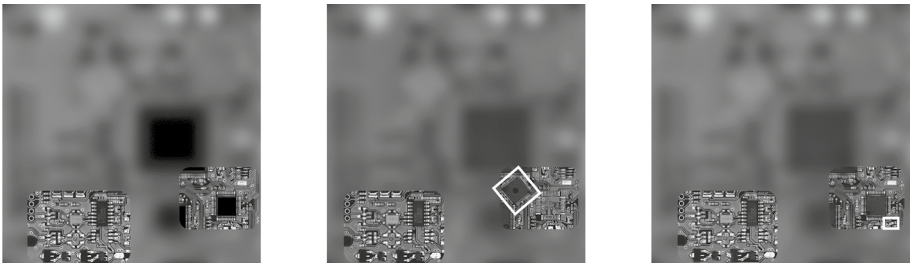

Fig. 1 The left image presents a normal PCBA. On the other ones, the white frames surround a large anomaly (middle image) and a small one (right image). Some parts of the images have been anonymized (material under intellectual property)

Some image examples are shown in Fig. 1[1] (presented in Bougaham et al. 2021), where one can figure out the details in terms of objects, reflections, or texture. The severe test limits imply a large number of false positives raised by the algorithm, which slows down the throughput, and requires tedious labor for human operators to make the final judgment (Vergara-Villegas et al., 2014). A standard acceptable operator misjudgment rate in these conditions is 2% of the production (Down et al., 2010) (true positives detected by the inspection process but incorrectly judged as negatives by the human operator), which is far higher than the actual rate of our industrial partner. However, this low human variability means that multiple efforts and training costs are spent on reducing as possible these false negatives. The initial claim is an average time loss of $\pm$8s/PCBA due to the algorithm false-positive rate, and around 83 parts per million abnormal PCBAs misjudged as normal by the operator, due to process credibility reduction.

Recently, deep learning methods have attracted much interest in the context of Industry 4.0, as they can help alleviate the problem of type I and type II errors (Xia et al., 2022). Thanks to a vast number of images, such an advanced method can be performed to state whether a product is standard or not. Therefore, these techniques can supplement or even replace traditional anomaly detection systems (Wang et al., 2018). However, due to the imbalanced nature of the datasets at hand, it is challenging to design a regular binary classifier. Indeed, the extreme rarity of anomalies yields many more images with normal products and a few images with anomalies. This lack of minority-class images leads industries to deal with representation learning techniques, suitable for extracting an insightful feature representation of the majority class. In a first step, a model learns the normal data distribution, and, in a second step, this model reconstructs a new input image under test, based on this normal representation model. Finally, a distance is computed between the input and the reconstructed images. This anomaly score states how different both images are, and a threshold defines the normality of the input image, under the assumption that the model will recover the eventual abnormal set of pixels (Li & Li, 2022).

The reconstruction quality is a difficult task for complex images, but the difference between normal and abnormal variability is another substantial difficulty. For images with many details, as is the case for the PCBAs, a macro view does not reveal high variability. Nevertheless, in detail, the PCBAs are showing many differences (Bougaham et al., 2021). Therefore, one of our challenges is to distinguish between a small defect and a

[1] Some parts of the images have been blurred to guarantee the intellectual property of our industrial partner. The arguments described also apply to the hidden parts, where information can be extrapolated.

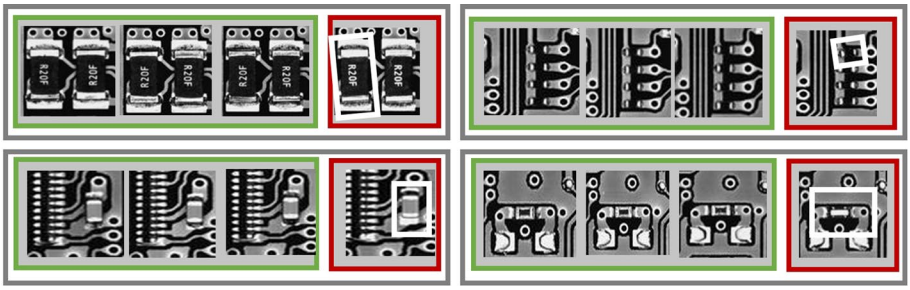

Fig. 2 (Color online) Four blocks of zoomed areas ($\approx X10$) for 4 different PCBA images. For each block, the left green frame (3 first columns) shows normal variations unlike the red frame one (last columns), where a small defect is observed. The challenge is to discriminate normal and abnormal variations. These very small areas represent around 1 cm × 1 cm over the 10 cm × 10 cm surface of the entire product

slight normal variation. Figure 2 shows such normal and abnormal variations in 4 different zoomed areas of the images, where one can appreciate the very small difference in component shift for both cases.

Based on the Vector Quantized Generative Adversarial Network (*VQGAN*) and the Generative Adversarial Network Anomaly Detection Through Intermediate Patches (*GanoDIP*) works (Esser et al., 2021; Bougaham et al., 2021), a methodology called ***VQGanoDIP*** is proposed. It is aimed to tackle the imbalance and complexity dimension of the real-world industrial PCBA dataset, composed of high-resolution images with a fine distinction between normal and abnormal variations. The main contribution is (i) to associate these works to get the best representation possible of the majority class, in addition (ii) to develop techniques (such as weighting normal variations or multi-level distances collection) to localize estimated defect areas, and (iii) to compute a composite anomaly score that characterizes them through a binary classifier. The objective is to reduce the false-positive rate while enforcing the zero-false-negative (ZFN) rate requirement.

First, a literature review on industrial anomaly detection and image synthesis is proposed in Sect. 2. Afterward, Sect. 3 details the *VQGanoDIP* methodology, stating how the three-step (reconstruction, metrics extraction, and normal/abnormal classification) can achieve the objectives. Finally, the method performance is qualitatively and quantitatively reported and discussed in Sect. 4, on multiple datasets.

2 Related works

Anomaly detection in computer vision presents a significant interest within multiple domains such as biomedical (Schlegl et al., 2017), industrial (Bougaham et al., 2021) and security (Kiran et al., 2018; Abdallah et al., 2016). Nowadays, several studies are dedicated to summarize these methodologies. The scope of our literature research is focused on the studies that perform anomaly detection on images through Generative Adversarial Networks (GANs). In this context, Xia et al. (2022) reviewed various existing methods applying deep learning algorithms including GANs for anomaly detection. According to this study, these methods often depend on large training samples. Therefore, data imbalance is one of the main application limitations. They state that GANs are one of the best

solutions proposed in the literature to deal with it, by learning representation features of the majority class, in an unsupervised manner.

Besides deep learning and GANs for image anomaly detection, several researchers focus on more traditional methods such as machine learning and classical image processing (Hasoon et al., 2021; Erfani et al., 2016; Sridhar et al., 2022; Matteoli et al., 2010). However, due to the complexity of the anomalies to be detected and the high dimensionality of our images, these methods are not suitable for our specific application. Recently, a research conducted by Liu et al. (2021) introduced an autoencoder technique for the detection of manufacturing errors in aluminum surfaces through image analysis. This work presents a challenge as it aims to detect anomalies in an unsupervised manner. To overcome this challenge, the authors proposed a dual prototype loss strategy, which prompts the encoder to produce feature vectors that are consistent with their own prototypes. The root mean square error (RMSE) between feature vectors was utilized as an anomaly indicator to evaluate the effectiveness of the approach. Unlike the autoencoder approach, the use of a discriminator in a GAN allows for direct detection of anomalies in the high-dimensional space of the input data. As a result, GANs offer greater flexibility and the capability to detect more complex anomalies compared to autoencoders.Consequently, GANs are more flexible and can be used to detect more complex anomalies in the industrial data compared to autoencoders. Several works have proven the efficiency of GAN for the detection of anomalies (Akcay et al., 2019; Akçay et al., 2019; Schlegl et al., 2019; Bougaham et al., 2021). In this context, Akçay et al. (2019) introduced an approach using unsupervised anomaly detection within a GAN training scheme. This approach is based on an autoencoder with skip-connections, terminated by a GAN discriminator that provides effective training for the normal class. However, they suggested to apply their method on higher resolution images as future work. This identified limitation is indeed an obstacle when small defects detection is a strong requirement, which is our case. Schlegl et al. (2017, 2019) explored the encoded latent vector thanks to a GAN generator learning the data distribution. First, the authors trained a generator and a discriminator using images without anomaly. Then, the pre-trained generator and discriminator are frozen, and a latent vector mapping is performed. Despite the high performance reported, its computational complexity remains expensive. In addition, the authors limited their experiments to low-resolution images and applied them to a unique type of images (retina optical coherence tomography scan), unlike the approach we introduced where the genericity dimension is considered.

In our previous work (Bougaham et al., 2021), we proposed to use intermediate patches for the inference step after a Wasserstein GAN (WGAN) training procedure. Our objective was to make anomaly detection possible on real-world industrial Printed Circuit Board Assembly (PCBA) images. This approach showed that our previous technique can be used to support current industrial image processing algorithms and avoid wasting time for industries using manual techniques. However, real-world implementation is still challenging, due to the high diversity of defects possible in a PCBA, particularly the very small ones, undetectable below the megapixel resolution. Van Den Oord et al. (2017) incorporated the concept of vector quantization (*VQ-VAE*) in order to learn a discrete latent representation. Following their methodology, the model is able to generate expressive images and speech data. Still the image resolution considered in this work is not sufficient to be considered for our high-resolution constraint. Razavi et al. (2019) improved the Vector Quantized Variational AutoEncoder (*VQ-VAE2*) models for large-scale image generation. They enhanced the auto-regressive priors used in their architecture to produce synthetic samples of higher coherence. One of their main contributions was to keep the encoder-decoder architecture simple

and lightweight. Regardless of the performance demonstrated by the VAE architectures introduced by these two studies, Esser et al. (2021) showed that the *VQ-VAE* methods produced reconstructions yielding blurred details, being an issue to reconstruct our PCBA images with sufficient fidelity. They addressed this limitation by synthesizing realistic detailed high-resolution images with a Vector Quantized Generative Adversarial Network (*VQGAN*). Their approach, based on *VQ-VAE*, consisted of representing the images as a composition of coherent and rich details, adding a GAN discriminator to improve the images realness, and considering a perceptual loss.

In addition to the anomaly detection and image synthesis problems, our business specificity requires to guarantee that no defect can be missed by the algorithm. Some studies consider the classification as an optimization or a cost-sensitive problem (Sangalli et al., 2021; Krawczyk et al., 2014), in order to prioritize the false positive misclassifications instead of the false negative ones. Although these works consider the constraint in an end-to-end manner, the missed detections are minimized without any guarantee on their absence. Roth et al. (2021) proposed to adjust a 100% recall threshold for predictions of a patch-features encoding anomaly detection method, for industrial public dataset images. Their method exploits a threshold that guarantees no false negatives, but operates at low resolution, being an issue for our PCBA anomaly detection task.

Considering the above studies using quantized autoencoder with GAN methodologies and showing promising results within the field of anomaly detection, we propose a new approach exploiting adversarial quantized auto-encoders to reconstruct an input image, collect metrics from this reconstruction, train a binary classifier on this metrics dataset, for high-resolution and challenging real-world images. Such a method aims to discriminate between anomalies that are not necessarily clear and patterned compared to the normal variation, and to guarantee that no missed detection is possible (frequent requirement for the industrial or medical applications).

3 VQGanoDIP: VQGAN anomaly detection through intermediate patches

This section details the proposed *VQGanoDIP* (*VQGAN* + *GanoDIP*) methodology, designed to localize and quantify abnormal areas in the PCBA and the MVTEC-AD 1024×1024 images (Bergmann et al., 2019), a set of different high-resolution industrial images composed of products with and without anomalies. The first step is based on *VQGAN* (Esser et al., 2021), in particular, the reconstruction part. Eventual anomalies are expected to be recovered on a set of images of the two classes, thanks to the model previously trained on normal the majority class, spotting the differences between the input and the reconstruction. Based on these differences, a highlighting technique inspired by our previous work (Bougaham et al., 2021) is performed, where a patching method localizes anomalies in a reduced set of areas. Then, the differences are quantified with several metrics, image-wise, patch-wise, and pixel-wise. Finally, these collected metrics on normal and abnormal images are used to train a binary classifier model that compute a composite anomaly score qualifying the product quality. To determine the product quality, a threshold is adjusted to avoid missing any true positive.

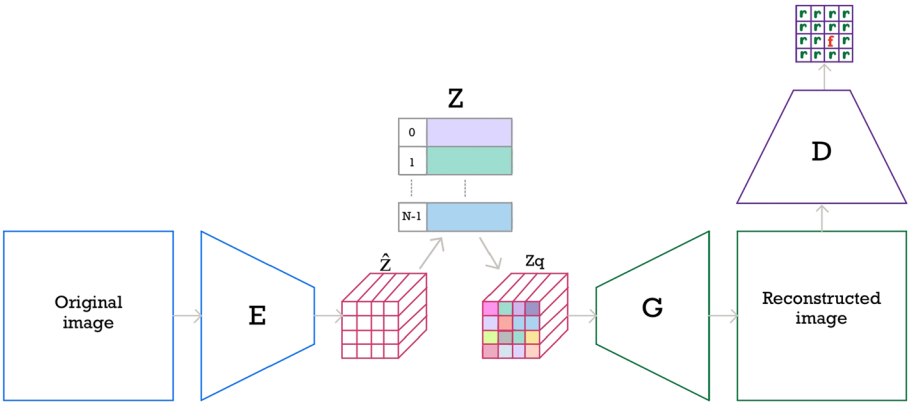

Fig. 3 (Color online) Figure inspired from Esser et al. (2021) presenting the training strategy of *VQGAN*, which is used as the first step of the proposed *VQGanoDIP* methodology

3.1 First step: image reconstruction and anomaly localization

The first step of the *VQGanoDIP* methodology is the image reconstruction and the most abnormal sets of pixels localization.

3.1.1 VQGAN reconstruction

The *VQGAN* method has been selected for its ability to efficiently learn a data representation and synthesize small details in an information-rich image. The specificity of our PCBA dataset lies in the fact that the images are similar in a global manner, with no significant variation in the component placement, the circuit color, or the solder pads. However, it offers a lot of small local variations due to the placement and solder process windows. The CNNs inductive bias that encourages the local interactions coming from this method allows dealing with these small variations and can follow the data distribution locally and globally. Moreover, its "context-rich vocabulary learning of the image constituents" (Esser et al., 2021) reduces the practical computational resources and allows generation in the megapixel regime, which makes it possible to work in a high-resolution space and thus capture very small defects.

VQGAN is a vector quantized autoencoder model augmented with a GAN. On top of this method, image synthesis is achievable thanks to a transformer architecture (out of our scope since the objective is only the reconstruction). Figure 3 shows the framework of the *VQGAN* reconstruction model.

The architecture is composed of 3 stages. First, the central part of the framework is an autoencoder. The encoder E learns a mapping function to transform the high-dimensional original image into a low-dimensional latent representation. Afterward, a reconstructed high-dimensional image is decoded from the latent representation, thanks to the latent-image space mapping, which is the decoder (or generator G).

The second stage is the vector quantized (VQ) part of this autoencoder. It adds the advantage of transforming the learned latent representation $\hat{z}$ into a quantized representation z_q, instead of a continuous one, yielding many possible values to be decoded

(difficult to learn). Therefore the model is able to focus on a restricted number of latent vectors, which significantly helps model convergence and avoids mode collapse (identified in our previous work (Bougaham et al., 2021), ignoring the latent vector due to the decoder performance). The vector quantization mechanism is based on an embedding matrix of a discrete number of vectors to learn, resulting in a codebook Z. Its purpose is to provide vectors as close as possible to the images constituents, represented in the overall latent representations of the autoencoder. The encoded representation of the input image is therefore replaced by the nearest neighbor from the spatial collection of vectors learned, and is then decoded to a reconstructed image.

Finally, the reconstructed image is fed to the discriminator D of a patchGAN (Isola et al., 2017) (similar to a regular GAN but qualifying $N \times N$ patches, instead of the entire image through a single scalar). The discriminator objective is to collect a patch-wise reconstruction loss, giving information regarding its realness, for the training procedure. This way, the decoder part of the *VQGAN* architecture takes the role of the generator part of the patchGAN, and the discriminator competes with it, stimulating the autoencoder and the codebook to provide realistic images, by receiving both the original and the reconstructed images. Its benefit is to provide images with high quality, instead of blurred ones that the *VQ-VAE* suffer from Esser et al. (2021).

Overall, the entire architecture is composed of 4 neural networks with a total number of 94,075,587 parameters, 28,372,480 for the encoder, 1,573,888 for the quantizer part (with the codebook), 40,472,193 for the decoder (generator), and 23,657,026 for the discriminator.

The end-to-end training procedure is guided by a combination of the reconstruction loss, the vector quantization loss, and the GAN loss. The reconstruction loss is a pixel-wise mean square error to capture detailed information, and a perceptual loss to capture semantic one: the original and reconstructed images are fed into a pre-trained VGG-16 network, and their last layer feature vectors MSE is computed. The global loss to optimize is defined as

$$\mathcal{Q}^* = \underset{E,G,Z}{argmin}\ \underset{D}{max}\ \underset{\mathrm{x}\sim p(x)}{\mathbb{E}} \left[\mathcal{L}_{VQ}(E,G,Z) + \lambda \mathcal{L}_{GAN}(\{E,G,Z\},D)\right] \tag{1}$$

for an encoder E, a decoder or generator G, a codebook Z and a discriminator D, where

$$\begin{cases} \mathcal{L}_{GAN}(\{E,G,Z\},D) &= \left[\log D(x) + \log\left(1 - D\left(G\left(Z\left(E(x)\right)\right)\right)\right)\right], \\ \mathcal{L}_{VQ}(E,G,Z) &= \left\|x - G\left(Z\left(E(x)\right)\right)\right\|^2 + \|sg[E(x)] - z_q\|_2^2 + \|sg[z_q] - E(x)\|_2^2, \\ \lambda &= \frac{\nabla_{GL}[\mathcal{L}_{rec}(\{E,G,Z\})]}{\nabla_{GL}[\mathcal{L}_{GAN}(\{E,G,Z\},D)]+\delta}, \end{cases}$$

with sg being the stop-gradient operation, z_q being the quantized representation, $\mathcal{L}_{rec}(\{E,G,Z\})$ being the perceptual loss capturing the differences between x and $\mathrm{G}\left(\mathrm{Z}\left(\mathrm{E}(\mathrm{x})\right)\right)$ with the pre-trained VGG-16 network, $\nabla_{GL}[.]$ the gradient of its inputs w.r.t the last L layer of the decoder and $\delta = 10^{-6}$ a small constant added for numerical stability (Esser et al., 2021).

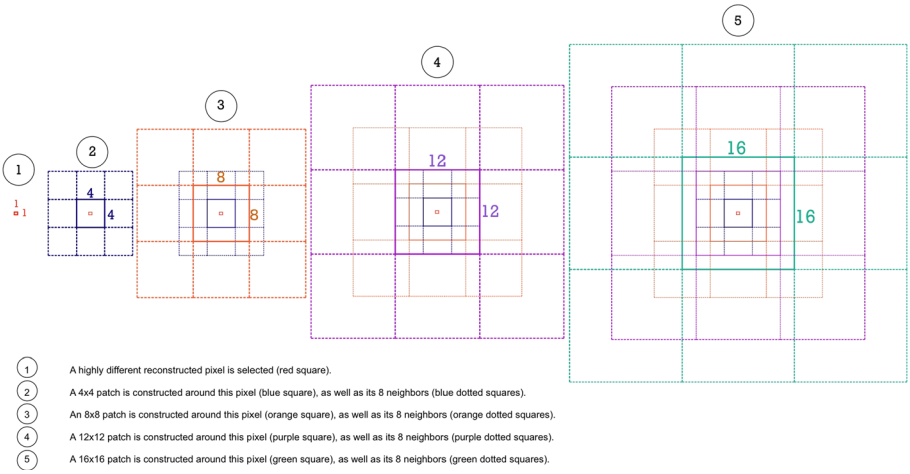

Fig. 4 (Color online) Overview of the zoom-out-and-shift technique. $n \times 9$ patches are created (here $n = 4$; enlarged factor $\alpha = 4$) to focus on the p most different pixels, at different scales

3.1.2 GanoDIP abnormal candidates isolation

Once the reconstruction model is trained, we reconstruct the test set images. The *GanoDIP* inference step we developed in our previous work (Bougaham et al., 2021) is applied by extracting the most different patches between the original and the reconstructed image. In this work, instead of considering only the highest MSE patch-wise, we first keep the p pixels showing the highest absolute differences, then we construct patches by zooming out and shifting around these p pixels. Figure 4 gives an overview of this technique.

The zoom-out-and-shift method is repeated n times, enlarging the patches of α pixels each time, yielding different $s \times s$ patch sizes. The shift step allows constructing 8 more patches for each size in the neighborhood (top-left, top-center, top-right, center-left, center-right, bottom-left, bottom-center, bottom-right), covering an entire estimated abnormal set of pixels. According to the industrial partner experts, this method imitates the natural way of a human visual search, when an anomaly is expected in a small area. It assumes that, if a defect occurs, several close pixels are concerned instead of a single one. As reported in Eckstein (2011), the normal areas are considered as a white noise, and the center-surround mechanism is applied to efficiently search and appreciate the defect.

These $n \times 9$ (center + 8 neighbors) patches repeated on the p most different pixels are finally used to compute the Frechet Inception Distance (Heusel et al., 2017) (FID) between the original and the reconstructed images, considering their feature vector difference (generated with an Inception Resnet-V3 network, pre-trained with ImageNet). This metric helps getting perceptual differences, pertinent with our computer vision task, instead of pure pixel differences, unrelated to the visual similarity between two images.

From these $p \times n \times 9$ FID values, we keep the q highest as the estimated abnormal candidates. At the end, we get q patches with different sizes, containing the highest perceptual difference between the original and the reconstructed image. They will first localize the estimated abnormal areas in the overall image and then be exploited for the second step of the methodology, the metrics collection described in the next subsection.

3.2 Second step: multi-level difference metrics collection

Recent anomaly detection methods (Bougaham et al., 2021; Akçay et al., 2019; Schlegl et al., 2017, 2019) design an anomaly score directly with regular pixel-wise or patch-wise MSEs, and the losses yielding from their neural network architecture. We aim to reproduce this strategy, in addition to taking into account the computer vision dimension of the task. Different type of distances between the input and the reconstruction will therefore be used to design the anomaly score, giving the method the best set of information on which to rely. Indeed, once the reconstruction has been performed and the most different patches have been identified on the test set, several multi-level metrics will be collected to characterize an anomaly present in the image. These metrics will be associated with each image, expressing a wide variety of information contained in the difference between the original and the reconstructed image.

Three different metric levels are considered. The first metric level describes the reconstruction quality image-wise. This is the case for the whole image pixel-wise MSE between the input and the reconstructed image (raw reconstruction loss), the pre-trained VGG-16 last layer MSE between the input and the reconstructed image features (raw perceptual loss), the MSE between the encoded latent representation and its quantized version (raw quantization loss), the patchGAN discriminator average loss for both the input and the reconstructed image (raw GAN losses), the ORB image matching difference (Karami et al., 2017), or the aggregated values of all the pixels resulting from the input-versus-reconstructed absolute difference. These aggregated values are the sum, maximum, minimum, mean, first quartile, second quartile (a.k.a. the median), and third quartile.

The second metric level describes the reconstruction quality pixel-wise. In this case, the goal is to retrieve information from the p highest pixel values resulting from the input-versus-reconstructed absolute difference. The same aggregated values are considered to qualify this set of most different pixels.

The third and last metric level describes the reconstruction quality patch-wise. The methodology involves computing matrix distances on the q patches selected via the p highest pixel values and the zoom-out-and-shift technique. We append each of these patch distance values into a sequence and apply the aggregated method to get a scalar that qualifies this set of patches. This is the case for several established distances between two matrices (Frechet, SSIM, Braycurtis, Canberra, Euclidian, Cosine, Wasserstein, Hamming, Minkowski, Jensen-Shannon divergence, etc.), as well as for the FID (triggering the selection). Figure 5 shows the architecture at the metric collection step.

For a dataset with the same type of images and a fixed position (like the PCBA one, where the same product model does not vary in rotation or translation), the same set of metrics, weighted by the normal variation inside the majority class, is computed. This technique brings business knowledge to the method, making it possible to reduce the distances where many normal variations are already observed in a normal class dataset. To do so, we isolate m unseen normal images, compute and normalize the pixel-wise average difference value between the input and the reconstructed images. We then obtain a mask (after a flip operation) reflecting the average difficulty that the reconstruction model encounters, due to the high normal variations. If a pixel varies a lot in the normal images, the mask pixel value will be close to zero. Otherwise, it will be close to one. These mask pixel values will be multiplied by the difference pixel values between a test input and reconstructed image, and will weight the difference computed (especially

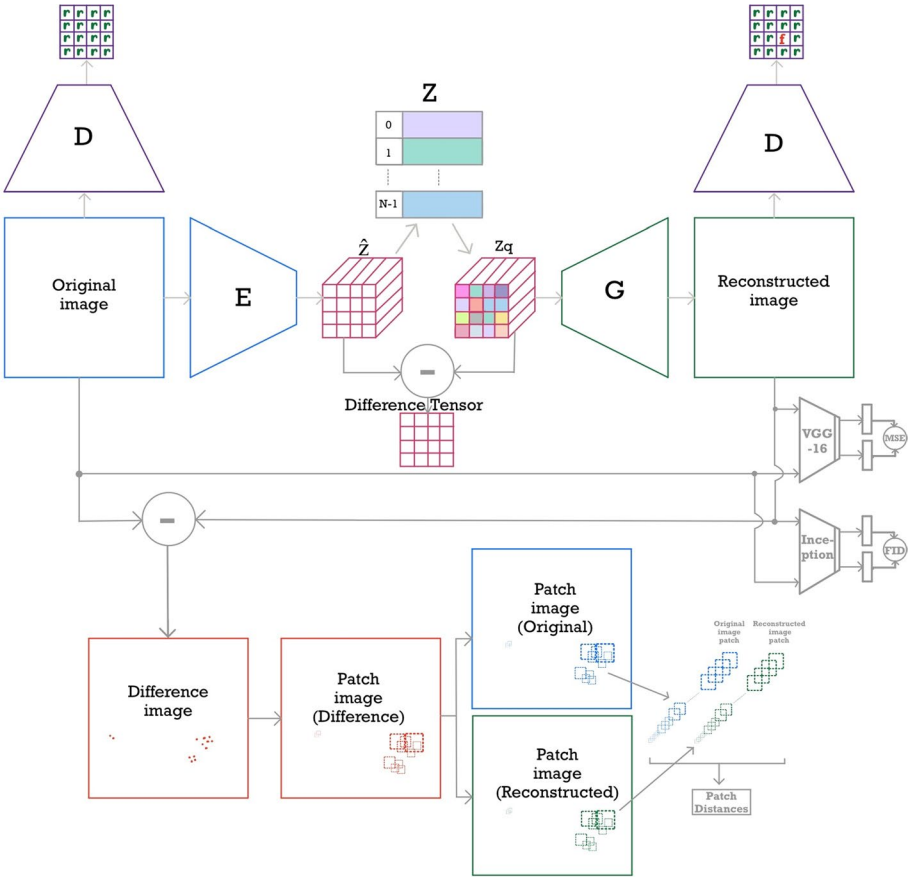

Fig. 5 (Color online) Overview of the *VQGanoDIP* architecture, at the inference step. A multi-level metric collection is extracted from the input and reconstructed images, containing insightful information to determine whether an anomaly is present or not

on high normal variation areas) with the help of a part of the business knowledge. Then the FID zoom-out-and-shift and the metrics collection are performed, based on these new p most different pixels. That doubles the metrics number (with and without the weighting mask), which will be used during the last step, detailed in the next subsection.

3.3 Third step: composite anomaly score creation

Associating an anomaly score to each test set image is the last step of the methodology. At this stage, we will use the few abnormal images we have at hand to create a classifier able to discriminate between the two classes. Indeed, instead of only using the reconstruction and eventually the latent loss as it is usually performed in anomaly detection techniques, we will feed a dataset built upon the metrics collected (instances in rows, metric values in columns) into a classifier, in order to let it build a new, composite, anomaly score that best discriminates the classes.

The anomaly score is designed in 4 phases. A data processing first phase is applied, removing features with a constant value. If several metrics are highly correlated (more than 95%), only one of them is kept. Values are also scaled into a [0–1] range.

Then, in a second phase, a randomized search for the hyperparameters optimization, efficient when the number of hyperparameters is large (Bergstra & Bengio, 2012), is applied to the entire dataset to achieve the best accuracy. If the dataset is imbalanced (this is the case for the MVTEC-AD datasets), we rely on the SMOTE algorithm (Chawla et al., 2002) to generate artificial positive samples (minority class) in the cross-validation stage, keeping pertinent the accuracy metric to optimize.

The third phase is a Leave-One-Out Cross-Validation (LOOCV) procedure, with the best-resulted hyperparameter set. This technique is computationally heavy, but even if we have a dataset composed of twice the number of the minority-class images (we take the number of normal images as the same we have for the abnormal ones), it is worth training as many models as there are instances, to get a reliable assessment. A stratified k-fold cross-validation is performed on the classifier, splitting all the instances (except one) in a training and validation set. A grid search procedure selects the best classifier type between several binary classification ones, namely a decision tree (DT), a random forest (RF), an extra trees classifier (ET), an adaptive boosting classifier (ADA), a light gradient boosting machine (LGBM), a gradient boosting classifier (GBC), an extreme gradient boosting classifier (XGBoost), a logistic regression (LR), a K nearest neighbor classifier (KNN), a gaussian naive bayes classifier (NB), a linear discriminant analysis (LDA) and a quadratic discriminant analysis (QDA).

In the final phase, we set the threshold as the lowest prediction probability that the classifiers associate to the abnormal test set. This way, we ensure that all the abnormal images in the test set are predicted as it has to be, and we can evaluate the classifier through the accuracy (or the geometric mean if the dataset is imbalanced), being degraded only with the normal images misclassification. Therefore, the prediction probability that the classifier gives to a test image to be abnormal is the composite anomaly score.

4 Evaluation

To summarize the above section, the proposed methodology is composed of three steps, which are a reconstruction model creation and anomalies localization, a metrics collection based on the input and reconstructed images, and a binary classifier training with the objective of building a composite anomaly score. After having detailed the methodology, this section is devoted to the experimental setup followed by the qualitative and quantitative model evaluation, finally discussed.

4.1 Experimental setup

For implementation purposes, several hyperparameters have to be tuned. In practice, we observed that the following decisions are the best to deal with the model performance and inference time. These following values are tailored for the PCBA dataset, but also tested on the MVTec-AD datasets (Bergmann et al., 2019).

In order to capture a large amount of insightful information on the PCBA dataset, we increased the default number of the codebook entries from 1024 to 2048 and its

Table 1 Size of the total available images, for the VQGAN reconstruction and the anomaly score classification model training sets of each dataset, with respect to the number of normal and abnormal data

Dataset	Total		VQGAN	Classifier	
	# Normal	# Abnormal	# Normal	# Normal	# Abnormal
PCBA	534	174	360	174	174
Cable	282	92	141	141	92
Carpet	308	89	154	154	89
Grid	285	57	143	142	57
Hazelnut	431	70	216	215	70
Leather	277	92	139	138	92
Screw	361	119	180	181	119
Transistor	273	40	136	137	40
Zipper	272	119	136	136	119

dimensionality from 256 to 512. This way, a large variety of textures, reflections, orientations and shapes can be captured by the codebook, and returned to the quantized latent representation.

For the PCBA dataset, 360 1024×1024 anomaly-free images are randomly selected to create the first step reconstruction model. The imbalanced nature of the dataset constrains us to work with only a few abnormal class images, compared to many normal class ones at our disposal. We have 174 abnormal images, so we randomly select 174 normal images (easily available, as it is the majority class) to get a balanced dataset of 348 images for the last-step composite anomaly score model creation.

For the MVTec-AD datasets, half of the normal images have been selected for the first step reconstruction model training. The other half of the normal images and all the abnormal images were selected for training and evaluation of the last step anomaly score model. Therefore, we end up with imbalanced datasets, this choice being motivated by the necessity to keep sufficient normal images for the classification step. Table 1 provides an overview of the number of images for each dataset.

On these test images, we apply the *GanoDIP* inference step as we developed in our previous work (Bougaham et al., 2021), keeping the $p = 100$ highest absolute difference pixels. The zoom-out-and-shift step is repeated $n = 4$ times, with an enlarged area of $\alpha = 4$ pixels each time. We therefore get 36 patches of 4×4, 8×8, 12×12 and 16×16 sizes. We finally keep the $q = 250$ worst FID patches, between the original and the reconstructed images, as the estimated abnormal candidates.

To evaluate areas of high normal variability, the weighting mask is based on $m = 30$ unseen random normal images and is only suitable for the PCBA dataset. Indeed, in the MVTec-AD datasets, objects and textures are rotated or translated, preventing the possibility of considering such a weighting mask.

The cross-validation to train the anomaly score classifier is a 5-fold stratified one. The randomized search for the optimal hyperparameters is executed for 500 iterations. The classifier selected is the one that gives the best geometric mean average on 10 independent runs, under the zero-false-negative constraint.

All experiments have been undertaken with an Nvidia RTX A6000 GPU, an Intel i7 CPU, using Python 3.8, Cuda 11.2 and Pytorch 1.10. For the training step, approximately 48 GB of GPU RAM is required to host the operations, for a single 1024×1024 image per

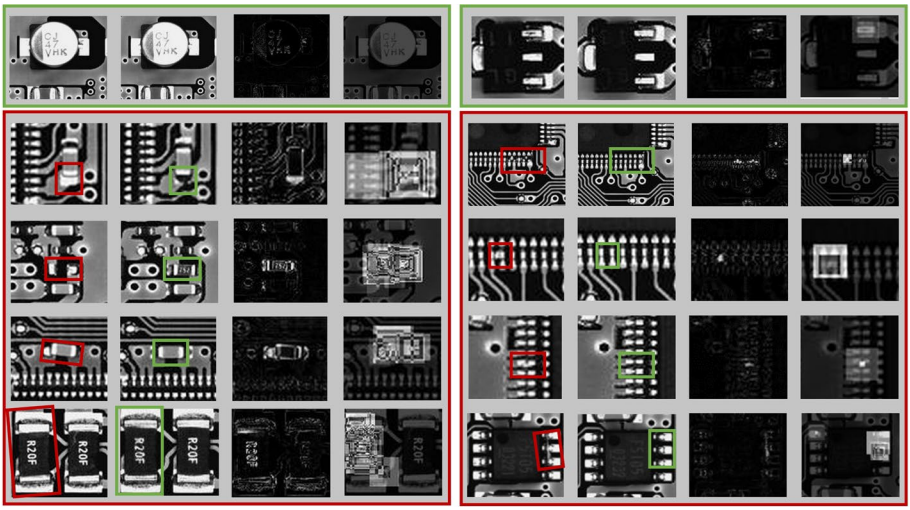

Fig. 6 (Color online) Two sets of original images (1st column) with the defect highlighted (red frame), reconstructed images (2nd column) with the recovered pixels highlighted (green frame), difference images (3rd column) and patch images (4th column) for 10 different zoomed areas ($\approx X10$) of the PCBA dataset (placed in rows). The first row shows anomaly-free areas unlike the four last rows, where a defect is observed. The left set shows component defects, and the right set shows solder defects

batch. For the inference step, about 5 GB of GPU RAM is needed and the inference time is about 50 s. This means that our GPU RAM capacity can handle a batch of up to 9 images, resulting in a per-image inference time of approximately 6 s. This time frame is suitable within the constraints of the business, making it viable for real-world production.

4.2 Qualitative assessment

The proposed methodology is qualitatively assessed, through the reconstruction quality (Sects. 4.2.1 for the private real-world PCBA dataset and 4.2.3 for the public MVTEC-AD datasets), and through a Visual Turing Test for the PCBA in Sect. 4.2.2.

4.2.1 PCBA reconstruction quality

A first way to evaluate the anomaly detection method proposed is to visually check how the original defects are recovered. Figure 6 shows several zoomed area examples where PCBA images contain anomalies that are correctly recovered, thanks to the *VQGAN* implementation. The reconstruction, difference, and abnormal patches estimation images are presented. The left set of images illustrates very small component shifts or absence, and the right set shows slight solder defects. We can see that the defects are correctly recovered (2nd column of each sets), proving the reconstruction efficiency for such complex data. The input data distribution is well followed for normal areas (first row of the figure), yielding low pixel differences (3rd column of each sets) after reconstruction. This statement means that the reconstruction model does not fall into the posterior collapse problem as it was the case in our previous work (Bougaham et al., 2021). The probable cause was a stronger generator

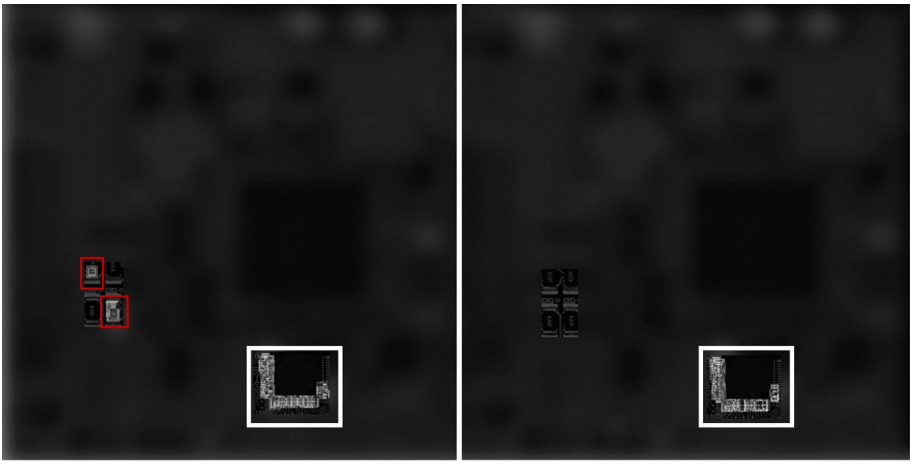

Fig. 7 (Color online) Abnormal patches estimation with (left image) and without (right image) the weighting technique. The abnormal patches estimation (false positives on the left unblurred red-framed area) disappears with this technique, focusing on true positives (right unblurred white framed area). Some parts of the images have been anonymized (material under intellectual property)

that always generated a quasi-identical image for any input image, whatever the latent vector variations. This limitation, which yielded a unique golden sample, is now solved.

Another step forward is the possibility of reconstructing very small anomalies, thanks to the 1024×1024 resolution. The *VQGAN* architecture makes it possible to consider an entire high-resolution input image at once, instead of patching it with a lower resolution, as it is the case for the *f-AnoGAN* (Schlegl et al., 2019) method, for instance. Therefore, difficulties of a challenging dataset like the PCBA one (small components in an information-rich global image) in the industrial context (entire image needed to evaluate the method with confidence before going into production) can be overcome.

Regarding the overall difference images, we can state that some specific zones are always noisy due to the normal variability of the dataset (marking on the components, reflections on large solder pads, details on the 2D serial number barcode that highly change, image by image). However, thanks to the weighting method, these false-positive differences are reduced, letting the *GanoDIP*-like technique choose the true positives. Figure 7 shows the relevance of this technique, which can only be applied with an adjusted position pre-processing task (here thanks to fiducial reference centering). We can appreciate how the focus on abnormal areas is improved when the weighting technique is applied, selecting the suitable p most different pixels (and thus the right q highest FID patches), by reducing this difference when a high normal variation has been previously observed. This improvement is important when distance metrics will be used for the composite anomaly score (signal noise ratio increased).

Despite these promising observations, some limitations remain. Figure 8 illustrates that for larger anomalies, the reconstruction is not as efficient, showing artifacts in the abnormal set of pixels. We can reasonably think that this comes from the latent representation dimensions, well fitted to deal with smaller details. Indeed, the choice for the codebook entries of 2048 and its dimensionality of 512 is particularly pertinent to catch the very small details in the image (like solder bridges or little component shifts) and the fine texture rendering (the PCBA silk or the solder pads reflects). This leads to difficulties for larger defect

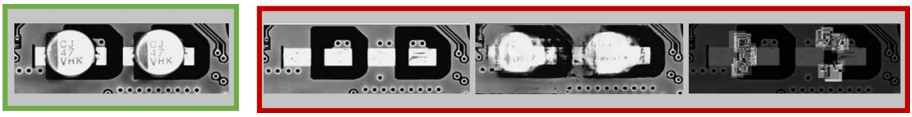

Fig. 8 (Color online) Zoomed area ($\approx X10$) of a normal product image (left-green-framed 1st image) and an abnormal product image (right-red-framed image). One can see in the abnormal image that the input (2nd image) presents a pretty large defect (absence components), reconstructed with artefacts (3rd image) but still well patched (4th image)

reconstructions. However, the artifacts reconstructed present all the same significant differences compared to the original abnormal image, and the *GanoDIP*-like patch isolation method can still focus on the right area.

The quality reconstruction is a key indicator, but it is not sufficient to estimate the methodology performance. Another indicator is the location correctness of the abnormal estimated patches. The difficulty lies in the fact that all the q patches will compete with each other to reveal the anomaly, and it is useful to notify that the first condition is to get at least one patch on the defect. For this concern, after the test set reconstruction, we observe that all the defects of the PCBA dataset are covered by a patch, which will be used as insightful information for the anomaly score creation.

4.2.2 PCBA visual turing test

Another way to assess the methodology qualitatively is to make a Visual Turing Test (VTT) on a normal set of PCBA images. Indeed, to implement the method in a real-world production line, domain experts need a high degree of confidence. The *VQGAN* model is hardly explainable (due to the multiple deep neural networks), but the VTT could reassure the experts on the reconstruction quality, being the first important block of the entire methodology. Therefore, this test has been conducted with 5 experts with different expertise levels, used to manipulate the PCBA images.

The protocol followed has been inspired by classical ones, reported in Salimans et al. (2016), Han et al. (2018) or Schlegl et al. (2019). 50 original normal images and 50 other reconstructed ones are randomly shuffled and shown to the experts. 16 seconds are given to the participants to determine if the image shown is real (original from the production line camera) or fake (reconstructed by the model). Between 2 images, 5 seconds of a blank screen is displayed to reset the visual memory and encode the judgment in a document. In the middle of the test (50th image), a 5 minutes break is taken to keep them focused until the end. No discussion between them is possible to avoid any eventual bias.

If the candidates cannot clearly state whether the images are artificial or not, then we can conclude that the reconstruction model generates realistic images, thanks to the architecture performance. This would be a good sign that this first reconstruction block is satisfying and brings confidence in the overall methodology. This procedure is a kind of human (expert) discriminator, like the one we have in the GAN part of the reconstruction model.

The results are presented in the Fig. 9. The correct classification average rate of 59.8%, with a 13.9% standard deviation, proves the difficulty of a domain expert distinguishing real images from fake ones. We can therefore conclude that the model generates high-fidelity images, a key argument for the users to adopt the algorithm.

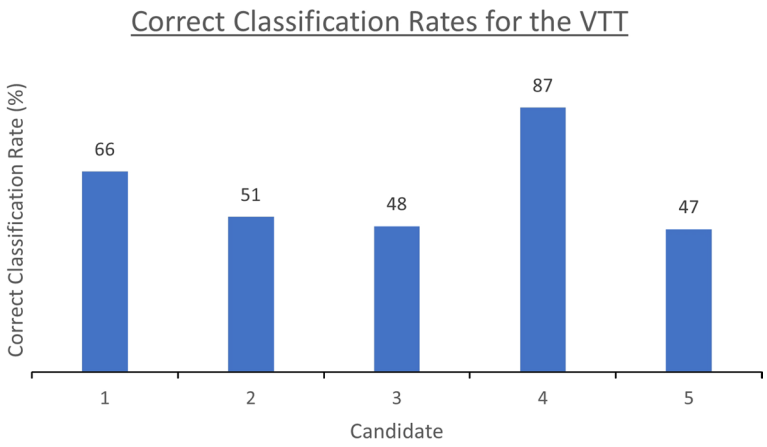

Fig. 9 Classification Rates for the 5 domain experts participating in the VTT

Fig. 10 Zoomed area ($\approx X20$) of the 2D serial number barcode where the original image (left image) gives a blurred reconstruction (middle image) with several pixel differences (right image)

An outlier stands out from the VTT results, with the highest score of 87%. The candidate with this score is a computer vision specialist responsible for designing anomaly detection algorithms for another production process. We can therefore understand why his biased attention differs from other participants. After a debrief session with him, it appears that some areas were insightful in judging the realness of the images. This is particularly the case for the 2D serial number barcode area, as shown in Fig. 10. Indeed, there are so many details and normal variations in this area that the model cannot reconstruct the dots composing the barcode matrix clearly. This yields to pixels that are smoothed, with a blurred effect. Hopefully, this limitation is not impactful because the weighting technique will reduce the difference pixel values. In addition, the defect opportunity in this area is very low.

4.2.3 MVTEC-AD reconstruction quality

The quality reconstruction can also be assessed on the popular MVTEC-AD datasets to figure out the genericity of the *VQGanoDIP* methodology. This work focuses on the 1024×1024 images resolution. Therefore lower resolution datasets were discarded.

Figure 11 shows the original, reconstruction, difference, and patch images for some examples of object and texture products, namely the screw, the hazelnut, the grid, and

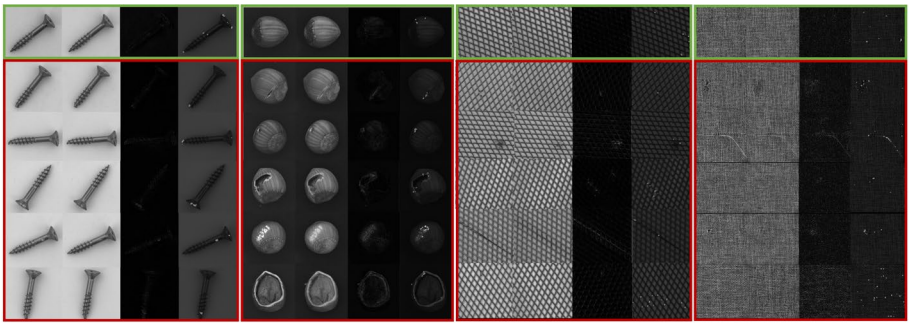

Fig. 11 (Color online) One normal image (1st row) and four abnormal images (4 last rows) of the screw (1st block), the hazelnut (2nd block), the grid (3rd block) and the carpet (4th block) datasets. For each block, the original (1st column), the reconstruction (2nd column), the difference (3rd column) and the patch (4th column) images are shown

the carpet datasets. We can confirm that the small defects are better recovered than the larger ones from this figure. For instance, a screw (1st block images) with a broken tail (2nd and 3rd rows) is more difficult to reconstruct than a scratch on the head (4th row) or the neck (5th and 6th rows). This is also the case for the hazelnut (2nd block images), where a rough crack (4th and 6th rows) is not well recovered, unlike a small cut (2nd row) or hole (3rd row). Despite these reconstruction difficulties, the abnormal estimated patches can still focus on the abnormal areas, even if the clustering effect is less observable.

Concerning the texture images, the same observation can be done (difficult reconstruction for rough defects but still estimated as abnormal), in addition to a larger split of the patches in the overall image. This is due to the high normal variations that offer the texture images (orientation, fibers, etc.), competing with the true-positive areas. We can therefore conclude that object images have better reconstruction performance than texture ones. This is due to the predominance object dimension of the PCBA images (compared to the texture dimension), which required hyperparameters selection adapted to this feature.

4.3 Quantitative assessment

A second way to assess the anomaly detection methodology is to measure established classification metrics, namely the accuracy, the precision, the geometric mean and the false-positive rate. In the case of imbalanced datasets, the geometric mean is a more relevant metric as it qualifies the performances on both classes, as well as the precision because it does not include true negatives (the majority class). There is no interest in measuring other metrics like the sensitivity (also named recall or true-positive rate), which will always be equal to 1 due to the absence of false negatives under the zero-false-negative constraint, or the AUCROC, as the only interesting threshold for our business case is the one able to detect all abnormal instances. As this specific anomaly score threshold cancels the false negatives, the confusion matrix is asymmetric, with the entire misclassified instances being false positives, giving the accuracy directly. The classification metrics are presented and discussed for the PCBA and the MVTEC-AD datasets.

Table 2 *VQGanoDIP* classification average metric values on 10 independent runs, for several classifiers on the PCBA dataset, considering the zero-false-negative constraint (ZFN columns) or not (STD)

Classifier	Accuracy(%)↑		Precision(%)↑		GMean(%)↑		FPR(%)↓		FNR(%)↓	
	STD	ZFN	STD	ZFN	STD	ZFN	STD	ZFN	STD	ZFN
ET	**94.65**	**87.93**	96.04	**80.58**	**94.64**	**87.09**	3.85	**24.14**	6.84	0
RF	93.71	78.97	95.42	70.77	93.68	75.74	4.43	42.07	8.16	0
XGBoost	93.28	78.42	92.13	70.37	93.25	74.92	8.16	43.16	**5.29**	0
LGBM	93.71	78.42	95.04	70.14	93.69	75.14	4.83	43.16	7.76	0
GBC	93.45	72.9	94.34	65.57	93.44	66.12	5.57	54.2	7.53	0
ADA	93.45	65.72	94.65	59.75	93.43	53.8	5.23	68.56	7.88	0
LR	93.3	58.39	94.15	54.73	93.3	39.14	5.75	83.22	7.64	0
DT	85.75	50.63	85.96	50.34	85.71	3.56	14.14	98.74	14.37	0
Q DA	90.52	50	**97.66**	50	90.2	0	**2.01**	100	16.96	0
KNN	91.64	50	93.7	50	91.57	0	6.15	100	10.57	0
LDA	94.28	50	95.95	50	94.26	0	3.91	100	7.53	0
NB	82.01	50	94.02	50	80.86	0	4.37	100	31.61	0

The ↑ sign means the highest the best, unlike the ↓ sign means the lowest the best. Values are sorted with the ZFN Accuracy column, and bold ones are the best of each column

4.3.1 PCBA classification metrics

Table 2 summarizes the average metric values on 10 independent runs, under the zero-false-negative constraint (ZFN columns), and, in a standard way, without this constraint (STD columns), on the PCBA dataset.

We can conclude that the Extra Tree Classifier offers the best classification metrics under the ZFN constraint, with an accuracy of 87.93% and 94.65% without this constraint (with a larger false-negative rate than a false-positive rate in this case). Notice that the the false-negative rate metric under the ZFN column is zero for all the classifiers, due to our quality exigence constraint. The linear and quadratic discriminant analysis (LDA, QDA), K nearest neighbor (KNN) and gaussian naive bayes (NB) classifiers raise 50% (the dataset imbalance rate value) of accuracy and precision, 0% of geometric mean, and 100% of false-positive rate, under the ZFN constraint. This situation happens because they encounter at least one challenging image for which the training failed to capture the discriminating features. Therefore, the lowest prediction probability for the positive class is zero, meaning that the classifier judges at least one positive instance with a 0% confidence to be positive. It yields an anomaly score threshold of zero, and all the instances will be classified as positives. We can conclude that these classifiers are not powerful enough to be used as our dataset anomaly score. The ensemble decision trees family seem much well fitted to the task.

Figure 12 shows the anomaly score distributions for the normal and the abnormal images of the PCBA dataset, built with the best classifier, raising a 87.93% accuracy. We can see the distributions are well separated (low anomaly score for the normal images and high anomaly score for the abnormal ones), even if the threshold (ensuring no missed detection) prevents an optimal split between them. It generates an overlap of the negative distribution, being the false positives (normal images scored above the threshold).

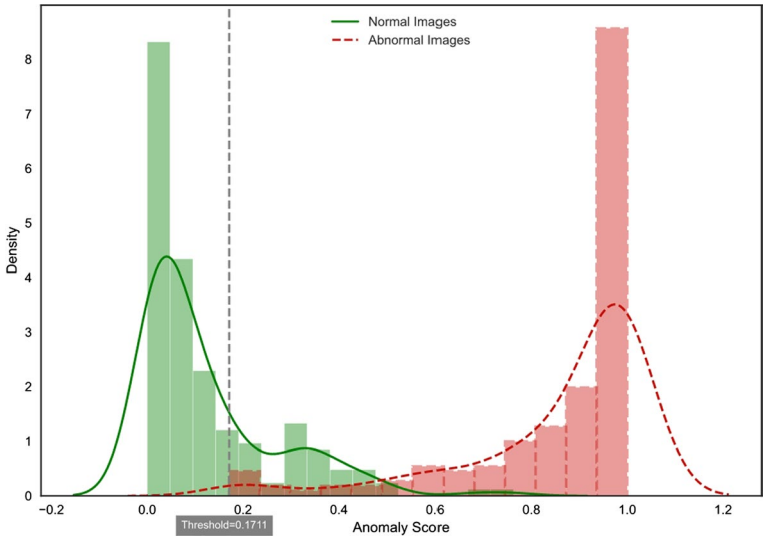

Fig. 12 (Color online) Anomaly score distributions of normal (solid-green line and bars) and abnormal (dashed-red line and bars) images for the PCBA dataset, with the threshold value (vertical dashed line in grey) that satisfies the ZFN constraint

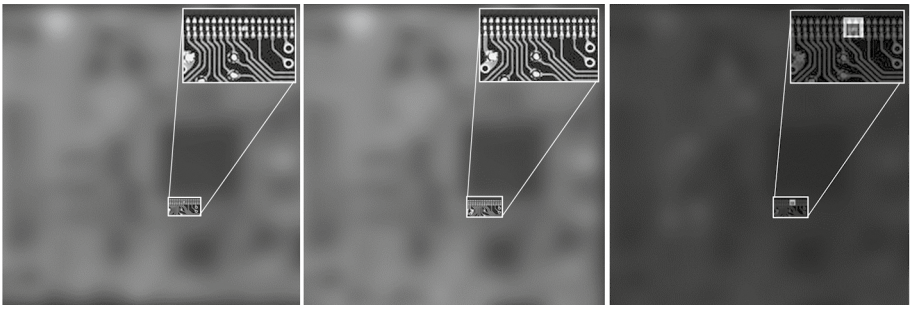

Fig. 13 Original (1st column), reconstructed (2nd column), and patch image (3rd column) with a zoomed view ($\approx X10$) of the PCBA most challenging image. Some parts of the images have been anonymized (material under intellectual property)

The zoomed abnormal area of the image conditioning the adjusted threshold (anomaly score of 0.1711) is presented in Fig. 13. This figure shows how difficult it is for the method to associate a high anomaly score with this very small defect, with respect to a low score for small normal variations. Indeed the small solder defect (a few mm^2 surface) in the input image is correctly recovered and an abnormal patch highlights the area thanks to the method. Nevertheless, this area competes with many other normal areas, giving difficulties for the classifier to build its decision function.

It is also interesting to determine the most influent collected metrics that impact the discrimination decision. Figure 14 shows the importance of the *GanoDIP*-like zoom-out-and-shift FID patches (for the 10 most influent metrics, 9 of them rely on these patches), as well

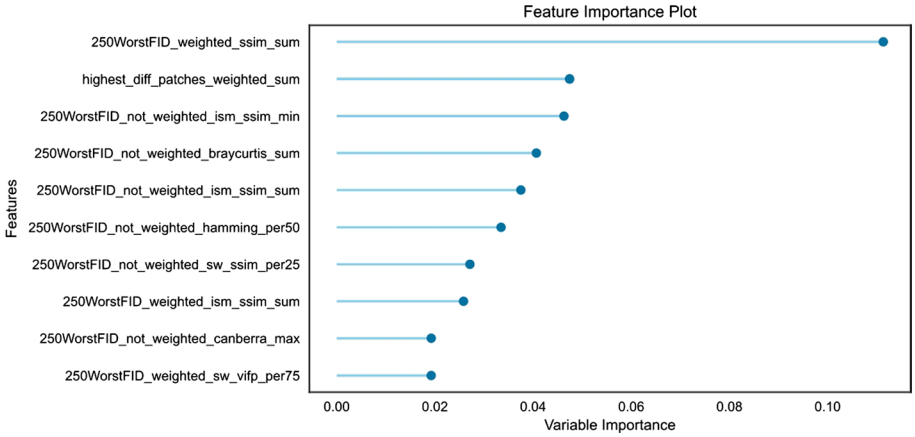

Fig. 14 10 most infuent collected metrics that impact the discrimination decision

as the weighting technique ($\frac{4}{10}$ most influent metrics) and the SSIM metric ($\frac{5}{10}$ most influent metrics). We can also notice that the most efficient aggregation technique is the sum of all the patch distances ($\frac{5}{10}$ most influent metrics). Finally, we can see that the most influent metric is the sum of the Structural SIMilarity (SSIM) values of all abnormal patches estimated, after applying the weighting mask to reduce the normal variations influence. It demonstrates the importance of considering the computer vision dimension in this task.

From a business point of view, each misclassification generated by the algorithm requires an operator visual inspection. Also, human misjudgement risks are proportional to the quantity to inspect. Therefore, an improvement on the accuracy directly impacts the time waste and the quality risk for the industrial partner. Here, thanks to the accuracy reached by the *VQGanoDIP* approach, the average inspection time required is decreased from 8 s to 1.9 s, and the operator misjudgment rate is divided by a factor of 4. This represents a significant improvement and is definitely promising to keep the partner competitive.

4.3.2 MVTEC-AD classification metrics

The same study has been performed on the public MVTec-AD datasets, especially on the images at the 1024×1024 resolution. The final summary of all the datasets, keeping the most accurate classifier under the ZFN constraint, is presented in Table 3.

This table shows that, under the ZFN constraint, datasets like Hazelnut, Leather or Grid show relatively low false-positive rates (6.88%, 23.77%, and 29.44% respectively), thanks to a correct anomalies reconstruction and a clear distinction between small normal and abnormal variations, inherent to the image complexity. Unlike these images, Fig. 15 shows how the reconstruction model has difficulties in following with fidelity the data distribution for the Carpet dataset, or recovering the large defects of the Cable dataset. In these cases, the patching technique cannot efficiently focus on the anomalies, and is completely fooled by the small normal variations.

These observations prove that the reconstruction quality, the data distribution fidelity, and the weighting mask are crucial elements for the anomaly detection task under the zero-false-negative constraint. If we do not consider the ZFN constraint, we can see that the method

Table 3 *VQGanoDIP* classification average metric values on 10 independent runs, for the most accurate classifiers (indicated into parenthesis) on the PCBA and the 1024×1024 MVTec-AD datasets, considering the zero-false-negative constraint (ZFN columns) or not (STD)

Dataset (Classifier)	Accuracy(%)↑		Precision(%)↑		GMean(%)↑		FPR(%)↓		FNR(%)↓	
	STD	ZFN	STD	ZFN	STD	ZFN	STD	ZFN	STD	ZFN
PCBA (ET)	94.65	87.93	96.04	80.58	94.64	87.09	3.85	24.14	6.84	0
Cable (ET)	78.28	59.49	76.69	49.79	75.09	55.94	13.05	66.95	35	0
Carpet (XGBoost)	86.55	56.71	82.26	45.95	85.24	56.03	10.45	68.31	18.65	0
Grid (LR)	96.03	79	98.08	58.64	93.38	83.85	0.7	29.44	12.11	0
Hazelnut (LGBM)	97.33	94.81	98.66	83.01	94.86	96.49	0.42	6.88	9.57	0
Leather (RF)	91.09	85.74	89.6	73.88	90.52	87.28	6.81	23.77	12.07	0
Screw (ADA)	93.5	74.7	92.53	61.43	93.05	76.02	4.86	41.93	8.99	0
Transistor (ET)	86.73	45.65	81.62	30.18	73.04	52.7	4.53	70.22	43.25	0
Zipper (XGBoost)	91.76	76.08	90.74	66.38	91.73	74.04	8.38	44.85	8.07	0

The ↑ sign means the highest the best, unlike the ↓ sign means the lowest the best

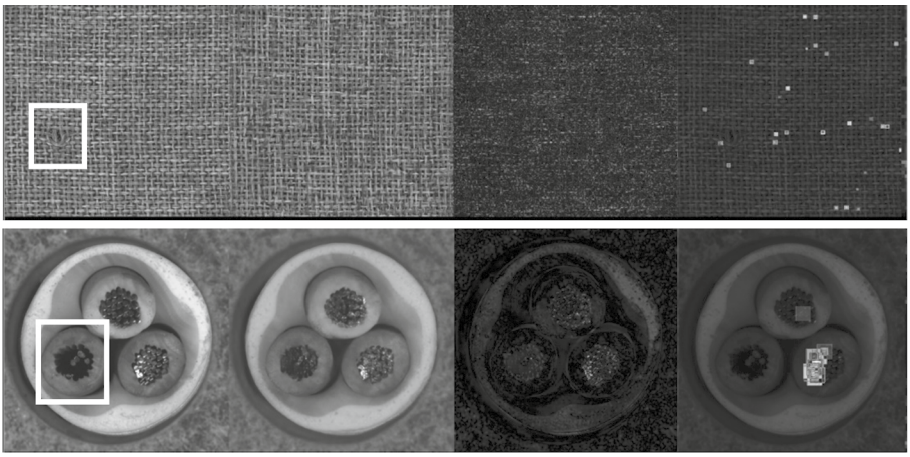

Fig. 15 Original (1st column) with the anomaly white-framed, reconstruction (2nd column), difference (3rd column), and patch (4th column) images for Carpet (1st row) and Cable (2nd row) image example. For the Carpet, we can see, on the difference image, all the small normal variations that the model could not well reconstruct. This yields, in the patch image, in many small patches everywhere but not in the anomaly area. For the Cable, we can see that the missing wires cannot be well recovered, which fools the patches focus

generates many more false negatives than false positives for all the datasets. This proves the difficulty of capturing the anomaly features in the entire overcrowded information contained in each image.

4.4 Summary

From a qualitative point of view, the reconstruction efficiency allows for the recovery of defects, and the model does not fall into the posterior collapse problem. Although small

defects are better recovered than larger ones, the abnormal estimated patches are still capable of focusing on abnormal areas in both cases. Object images demonstrate better reconstruction performance than texture ones, caused by the predominant object-type dimension of the PCBA images (compared to the texture dimension). The VQGAN architecture considers an entire 1024×1024 image at once, overcoming the difficulties posed by a challenging dataset like the PCBA one. However, some areas are always reconstructed with noise due to the high normal variability of the dataset. Additionally, artifacts may be observed in the abnormal set of pixels for larger anomalies as a result of the biased fitting to smaller components of the PCBA dataset. The differences in reconstruction between abnormal and normal data, as well as the weighting method, help to alleviate these issues and result in a satisfactory reconstruction quality.

From a quantitative point of view, the ensemble decision trees family, and specifically the Extra Tree Classifer, demonstrates remarkable performance on the PCBA dataset with an accuracy of 94.65% or 87.93% under the ZFN constraint. The MVTEC-AD datasets like Hazelnut, Leather or Grid also present satisfying results, thanks to their reconstruction quality. However, images with very small defects pose a significant challenge as the anomaly score becomes similar to normal images, causing difficulties for the classifier to make a decision. This challenge is nonetheless compensated by the anomaly score classifier performance. Finally, datasets such as Carpet and Cable face limitations in discrimination due to their struggle with reconstruction quality.

To ensure that the classifiers difference is statistically significant, we carried out a Wilcoxon rank-sum test on the geometric mean value under the ZFN constraint, between the best classifier and each of the 11 other classifiers. For all the datasets, the results show a p value always lower than the significance level α of 0.05, which demonstrates the statistical significance of the results.

5 Conclusion

An anomaly detection methodology suited for a real-world industrial use case is developed in this work. This is the continuation of a first work leading the *GanoDIP* method. The poor number of abnormal images (18) was the main limitation, thus yielding implementation difficulties due to the lack of defect variability. The current dataset contains around 10 times more abnormal images, with larger defect size, structure, and area variability. This amount is better, although still very small compared to the majority class. The proposal of this work is, therefore, to (i) be able to localize small defects, (ii) satisfy the zero-false-negative constraint, and (iii) reach the lowest false-positives rate possible. The *VQGanoDIP* methodology detailed in this paper reached the objectives thanks to a three-step methodology. It takes advantage of the vast amount of normal data, instead of a regular binary classifier. After a reduced anomalies estimation technique, the few abnormal data are indeed kept for further less-data-intensive processing.

The first step takes advantage of the recent advances in terms of image synthesis that make it possible to reconstruct an original image, following an input data distribution being anomaly-free. The *VQGAN* method, placed in an anomaly detection architecture, yields a strong representation of the normal class. It reconstructs very similar images to the original ones and, if any, replaces an abnormal set of pixels with an estimated normal one. The technique allows high-resolution reconstruction, fulfilling the business

case constraint, requiring to deal with small defects like solder bridges or electronic component shifts. Furthermore, only the majority class is required to train the reconstruction model in an unsupervised manner. Therefore, the imbalanced learning specificity is managed at this stage, and we save the few minority-class images that we have for the next step and the test set.

The second step of the methodology is a comparison of the 1024×1024 original and reconstructed images. A significant number of appropriate metrics are extracted from this comparison, including the different neural networks losses, as well as the computer vision distances on the worst difference patches, following the *GanoDIP* method strategy. To do so, a zoom-out-and-shift technique is performed on the worst patches to focus the metrics extraction on the highest Frechet Inception Distance areas. The objective is to make decisions on perceptual difference meanings instead of a regular absolute pixel difference. This step is applied to a balanced number of normal and abnormal class images, this number being conditioned by the abnormal set of images at hand.

The last step is to train a classifier able to act as the anomaly score to determine the image class. Its goal is to discriminate normal and abnormal images, thanks to the metrics collected. The ZFN constraint requires to set a low probability threshold on the classifier prediction, generating more false positives as a regular accuracy setup. The price to ensure the quality requirements is an accuracy decrease by 6.72%, reaching **87.93**% (instead of 94.65%).

For the business use case, the proposed methodology achieves a drop of the current inspection time from 8 to 1.9 s, and an estimated operator misjudgment rate **divided by 4**, which is a very satisfying achievement. It can be used as a baseline for many other use cases, where high-resolution images with small details and low variation between normal and abnormal areas are considered, especially under the ZFN constraint.

These promising results open new research questions for future works. Specifically, it would be interesting to find a better strategy to let the classifier learn the decision function integrating the zero-false-negative constraint directly, in an end-to-end manner. Another research direction could be to develop a reconstruction model that performs well on any type of dataset, whatever the normal variations characteristic or the anomaly sizes.

Acknowledgements The authors thank Jérôme Fink and Géraldin Nanfack for their insightful comments and discussions on this paper.

Author contributions AB conceptualized the ideas, designed the algorithm, carried out the experiments and wrote the manuscript. MEA, IL and BF supervised the work, provided critical improvements, helped writing, reviewed and approved the manuscript.

Funding Not applicable.

Availability of data and materials All public works and datasets have been cited in reference.

Code availability Not applicable.

Declarations

Conflict of interest The authors declare that they have no competing or conflict of interests.

Ethics approval The authors declare that this work is original, is not under consideration for publication elsewhere and has not been published previously. The authors approve the manuscript enclosed.

Consent to participate Not applicable

Consent for publication The authors consent to the publication of this work.

References

Abd Al Rahman, M., & Mousavi, A. (2020). A review and analysis of automatic optical inspection and quality monitoring methods in electronics industry. *IEEE Access, 8*, 183192–183271.

Abdallah, A., Maarof, M. A., & Zainal, A. (2016). Fraud detection system: A survey. *Journal of Network and Computer Applications, 68*, 90–113.

Akcay, S., Atapour-Abarghouei, A., & 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* (pp. 622–637). Springer: Berlin

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

Babic, M., Farahani, M. A., & Wuest, T. (2021). Image based quality inspection in smart manufacturing systems: A literature review. *Procedia CIRP, 103*, 262–267.

Bergmann, P., Fauser, M., Sattlegger, D., & Steger, C. (2019). Mvtec ad—A comprehensive real-world dataset for unsupervised anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 9592–9600).

Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. *Journal of Machine Learning Research, 13*(2), 281–305.

Bougaham, A., Bibal, A., Linden, I., & Frenay, B. (2021). Ganodip-gan anomaly detection through intermediate patches: A PCBA manufacturing case. In *Third international workshop on learning with imbalanced domains: Theory and applications* (pp. 104–117). PMLR.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). Smote: Synthetic minority oversampling technique. *Journal of Artificial Intelligence Research, 16*, 321–357.

Crispin, A., & Rankov, V. (2007). Automated inspection of PCB components using a genetic algorithm template-matching approach. *The International Journal of Advanced Manufacturing Technology, 35*, 293–300.

Down, M., Czubak, F., Gruska, G., Stahley, S., & Benham, D. (2010). Measurement system analysis. In *AIAG Reference Manual: Chrysler Group LLC*, Ford Motor Company, and General Motors Corporation, Southfield (pp. 103–123).

Eckstein, M. P. (2011). Visual search: A retrospective. *Journal of Vision, 11*(5), 14–14.

Erfani, S. M., Rajasegarar, S., Karunasekera, S., & Leckie, C. (2016). High-dimensional and large-scale anomaly detection using a linear one-class svm with deep learning. *Pattern Recognition, 58*, 121–134.

Esser, P., Rombach, R., & Ommer, B. (2021). Taming transformers for high-resolution image synthesis. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 12873–12883.

Filz, M.-A., Herrmann, C., & Thiede, S. (2020). Simulation-based assessment of quality inspection strategies on manufacturing systems. *Procedia CIRP, 93*, 777–782.

Han, C., Hayashi, H., Rundo, L., Araki, R., Shimoda, W., Muramatsu, S., Furukawa, Y., Mauri, G., & Nakayama, H. (2018). Gan-based synthetic brain MR image generation. In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)* (pp. 734–738). IEEE.

Hasoon, J. N., Fadel, A. H., Hameed, R. S., Mostafa, S. A., Khalaf, B. A., Mohammed, M. A., & Nedoma, J. (2021). Covid-19 anomaly detection and classification method based on supervised machine learning of chest x-ray images. *Results in Physics, 31*, 105045.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., & Hochreiter, S. (2017). GANs trained by a two time-scale update rule converge to a local nash equilibrium. In: Advances in neural information processing systems (Vol. 30).

Isola, P., Zhu, J.-Y., Zhou, T., & Efros, A.A. (2017). Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1125–1134).

Jia, H., Shi, J., & Chang, T.-S. (2004). An intelligent real-time vision system for surface defect detection. In *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004*. (vol. 3, pp. 239–242). IEEE.

Karami, E., Prasad, S., & Shehata, M. (2017). Image matching using sift, surf, brief and orb: Performance comparison for distorted images. arXiv e-prints, 1710.

Kiran, B. R., Thomas, D. M., & Parakkal, R. (2018). An overview of deep learning based methods for unsupervised and semi-supervised anomaly detection in videos. *Journal of Imaging, 4*(2), 36.

Krawczyk, B., Woźniak, M., & Schaefer, G. (2014). Cost-sensitive decision tree ensembles for effective imbalanced classification. *Applied Soft Computing, 14*, 554–562.

Li, H., & Li, Y. (2022). Anomaly detection methods based on GAN: A survey. *Applied Intelligence*. https://doi.org/10.1007/s10489-022-03905-6

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

Matteoli, S., Diani, M., & Corsini, G. (2010). A tutorial overview of anomaly detection in hyperspectral images. *IEEE Aerospace and Electronic Systems Magazine, 25*(7), 5–28.

Razavi, A., Van den Oord, A., & Vinyals, O. (2019). Generating diverse high-fidelity images with VQ-VAE-2. In *Advances in neural information processing systems* (Vol. 32)

Ren, Z., Fang, F., Yan, N., & Wu, Y. (2022). State of the art in defect detection based on machine vision. *International Journal of Precision Engineering and Manufacturing-Green Technology, 9*(2), 661–691.

Roth, K., Pemula, L., Zepeda, J., Schölkopf, B., Brox, T., & Gehler, P. (2021). Towards total recall in industrial anomaly detection. arXiv preprint arXiv:2106.08265

Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., & Chen, X. (2016). Improved techniques for training GANs. In: Advances in neural information processing systems (Vol. 29).

Sangalli, S., Erdil, E., Hötker, A., Donati, O., & Konukoglu, E. (2021). Constrained optimization to train neural networks on critical and under-represented classes. In *Advances in neural information processing systems* (Vol. 34).

Schlegl, T., Seeböck, P., Waldstein, S. M., Langs, G., & Schmidt-Erfurth, U. (2019). f-AnoGAN: Fast unsupervised anomaly detection with generative adversarial networks. *Medical Image Analysis, 54*, 30–44.

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

Sridhar, P., Arivan, S., Akshay, R., & Farhathullah, R. (2022). Anomaly detection using CNN with SVM. In *2022 8th international conference on smart structures and systems (ICSSS)* (pp. 1–4). IEEE.

Van Den Oord, A., Vinyals, O., & Kavukcuoglu, K. (2017). Neural discrete representation learning. In *Advances in neural information processing systems* (Vol. 30).

Vergara-Villegas, O. O., Cruz-Sánchez, V. G., de Jesús Ochoa-Domínguez, H., de Jesús Nandayapa-Alfaro, M., & Flores-Abad, Á. (2014). Automatic product quality inspection using computer vision systems. In *Lean manufacturing in the developing world: Methodology, case studies and trends from Latin America* (pp. 135–156).

Wang, W.-C., Chen, S.-L., Chen, L.-B., & Chang, W.-J. (2016). A machine vision based automatic optical inspection system for measuring drilling quality of printed circuit boards. *IEEE Access, 5*, 10817–10833.

Wang, J., Ma, Y., Zhang, L., Gao, R. X., & Wu, D. (2018). Deep learning for smart manufacturing: Methods and applications. *Journal of Manufacturing Systems, 48*, 144–156.

Xia, X., Pan, X., Li, N., He, X., Ma, L., Zhang, X., & Ding, N. (2022). Gan-based anomaly detection: A review. *Neurocomputing, 493*, 497–535.